STEM education in elementary grades:

Design of an effective framework for improving attitudes towards school science

Ph.D. Dissertation

Radu Bogdan Toma

UNIVERSIDAD DE BURGOS

UNIVERSIDAD
DE BURGOS

FACULTAD DE EDUCACIÓN

DEPARTAMENTO DE DIDÁCTICAS ESPECÍFICAS

Ph.D. Dissertation

STEM education in elementary grades: Design of an effective framework for improving attitudes towards school science

A thesis
submitted in fulfillment of the requirements
for the degree of

Doctor of Philosophy in Education

by

Radu Bogdan Toma

Supervised by:
Dr. Jesús Ángel Meneses Villagrá

Burgos, November 2019

The struggle itself towards the heights is enough
to fill a man's heart.
One must imagine Sisyphus happy.

El esfuerzo mismo para llegar a la cima basta
para llenar un corazón de hombre.
Hay que imaginarse a Sísifo feliz.

Albert Camus (1942, p.112)

Acknowledgments

I would like to show my greatest appreciation to my advisor Dr. Jesús Ángel Meneses Villagrá, who has been greatly tolerant and supportive of my scholarly pursuits. Your constructive feedback and encouragement have always been invaluable.

I am deeply grateful to Prof. Dr. Norman G. Lederman for the assistance given during my research stay at the Illinois Institute of Technology. I have greatly benefited from your expertise and generous support.

I want to also express my gratitude to the staff at the Centro Rural de Innovación Educativa de Burgos (CRIEB) for allowing me to conduct this doctoral thesis in this center. Special thanks to Daniel Arévalo and Abel Sancha for taking care of all the administrate matters to make the CRIEB the greatest test-bed that any researcher can dream of.

I would also like to show great appreciation to the University of Burgos for granting me a pre-doctoral scholarship that allowed me to undertake this dissertation on a full-time schedule.

I owe the deepest gratitude to my parents, for being an example of perseverance in overcoming life boundaries.

Finally, I am profoundly in debt to my partner, for being my companion throughout this endeavor. Your tireless patience and support taught me to always walk placidly amid this journey; and, your love remembered me that with all its starkness and broken dreams, it is still a beautiful world when you are by my side.

Agradecimientos

Quiero mostrar mi mayor agradecimiento a mi director de tesis, Dr. Jesús Ángel Meneses Villagrá, quien ha sido muy tolerante y me ha apoyado continuamente en todos mis esfuerzos académicos. Su constructiva retroalimentación y guía han sido siempre de un valor incalculable.

Estoy profundamente agradecido al Prof. Dr. Norman G. Lederman por la asistencia prestada durante mi estancia de investigación en el Illinois Institute of Technology. Su experiencia y generosos consejos me han resultado enormemente útiles.

Asimismo, quiero expresar mi agradecimiento a los docentes del Centro Rural de Innovación Educativa de Burgos (CRIEB) por permitirme realizar esta tesis doctoral en este centro. En especial, agradecer a Daniel Arévalo y Abel Sancha por solventar todas las cuestiones administrativas para hacer del CRIEB el mejor banco de pruebas con el que cualquier investigador pudiera soñar.

También quiero expresar mi agradecimiento a la Universidad de Burgos por haberme concedido un contrato predoctoral que me ha permitido realizar esta tesis a tiempo completo.

Del mismo modo, siento una inmensa gratitud hacia mis padres, por ser un ejemplo de perseverancia en la superación de las barreras de la vida.

Finalmente, estoy profundamente en deuda con mi pareja, por ser mi compañera en este camino. Tu incansable paciencia y apoyo me enseñaron a avanzar siempre plácido a lo largo de este viaje; y, tu amor, me recordó que, con toda su crudeza y sueños rotos, el mundo sigue siendo un hermoso lugar cuando estás a mi lado.

Abstract

This doctoral thesis specifically addressed the role of integrated STEM approaches in the development of elementary girls' and boys' attitudes towards school science. For such an endeavor, a mixed-method exploratory sequential design was adopted. During the first, qualitative strand, the development of a pedagogical approach, rooted in constructivist, socio-constructivist, and the tripartite-model of attitude, is advanced for the design and implementation of valid and plausible STEM teachings units. A pilot unit was developed and tested for 15 weeks with a total of 649 students enrolled from 3rd to 6th grades of elementary education. The analysis of three focus groups with six elementary school teachers that implemented the STEM unit, and the non-participant structured observation of 219 classroom sessions, revealed that the iterative process of implementation, evaluation, and re-design resulted in an authentic STEM unit that significantly addressed the interconnection between Science, Technology, Engineering, and Mathematics disciplines.

Next, during the second, quantitative strand, the effectiveness of this authentic STEM unit in improving students' attitudes towards school science was examined for 6 sequential weeks, using a one-group pretest-posttest quasi-experimental design. A total of 245 elementary school students participated in the 12-hour intervention unit. For data collection, existing measurement instruments were examined by means of two systematic reviews revealing a gap in the literature of valid and reliable attitude instruments. Therefore, two quantitative instruments, S-SSAS and SUCCESS, were specifically developed and psychometrically evaluated in this dissertation. Findings of the intervention suggested that the STEM unit seems to improve students' expectancies of success, enjoyableness of school science, self-efficacy in school science, and intentions to further enroll in school science subjects, without differential effects in terms of gender variable.

Taken together, this doctoral thesis advances a plausible conceptualization about what constitutes STEM education and how to best implement it in the Spanish educational system. Therefore, it carefully documents the iterative process of design, analysis, evaluation, and revision of a pedagogical framework that offers insights with regard to the characteristics of a valid, practical and authentic STEM education that possesses the potential to impact the attitudes towards school science of elementary school students.

Resumen

Esta tesis doctoral aborda específicamente el papel de los enfoques integrados STEM en el desarrollo de las actitudes de los niños y niñas de Educación Primaria hacia la ciencia escolar. Para tal fin, se adoptó un diseño de métodos mixtos de corte secuencial exploratorio. Durante la primera fase, de naturaleza cualitativa, se avanzó en el desarrollo de un enfoque pedagógico, enraizado en el constructivismo, el socio-constructivismo y el modelo tripartito de las actitudes, para el diseño e implementación de unidades de enseñanza STEM válidas y plausibles. Así, se desarrolló una unidad didáctica piloto que fue probada durante 15 semanas con un total de 649 estudiantes de 3º a 6º curso de Educación Primaria. El análisis de tres grupos focales realizados con seis maestros de Educación Primaria, así como los resultados de la observación estructurada no participante de 219 sesiones de clase, revelaron que del proceso iterativo de implementación, evaluación y rediseño se obtuvo una auténtica unidad didáctica STEM que aborda significativamente la interconexión entre las disciplinas de Ciencia, Tecnología, Ingeniería y Matemáticas.

A continuación, durante la segunda fase, de naturaleza cuantitativa, se examinó durante 6 semanas consecutivas la eficacia de esta unidad STEM para mejorar las actitudes de los estudiantes hacia la ciencia escolar, empleando un diseño cuasi-experimental pretest-post test de un solo grupo. Un total de 245 estudiantes de Educación Primaria participaron en la unidad de intervención de 12 horas. Para la recolección de datos, se examinaron los instrumentos de medición existentes mediante dos revisiones sistemáticas que revelaron una ausencia en la literatura de instrumentos de actitud válidos y confiables. Por lo tanto, dos instrumentos cuantitativos, S-SSAS y SUCCESS, fueron específicamente desarrollados y evaluados psicométricamente en esta tesis doctoral. Los resultados de la intervención sugirieron que la unidad STEM parece mejorar las expectativas de éxito, el disfrute de la ciencia escolar, la autoeficacia en la ciencia escolar y la intención de los estudiantes de 3º a 6º de Educación Primaria de matricularse en futuras asignaturas de ciencia, sin efectos diferenciales en términos de variables de género.

En conjunto, esta tesis doctoral propone una conceptualización plausible sobre lo que constituye la educación STEM y sobre la forma de implementarlo en el sistema educativo español. Así, se documenta detalladamente el proceso iterativo de diseño, análisis, evaluación y revisión de un marco pedagógico que ofrece detalles sobre las características de una educación STEM válida, práctica y auténtica que posee el potencial de mejorar las actitudes de los estudiantes de Educación Primaria hacia la ciencia escolar.

Table of contents

Chapter 1. Introduction

This chapter introduces the impetus for this doctoral dissertation. More specifically, section 1.1. describes the driving forces and the preliminary studies that prompted the development of this research. Then, section 1.2. discusses the educational and research problem that was investigated, whereas section 1.3. focuses on the specific aims and research questions.

1.1. Origins

In order to convey meaning and facilitate the understanding and interpretation of the different studies and sections compiled in this dissertation, it would be useful to begin by introducing how and why this research emerged; namely, its origins. Although there are many factors involved, the following five reasons were the ones that most encouraged this dissertation: (i) a decline in scientific related careers enrollments; (ii) students negative attitudes towards science; (iii) science education reforms; (iv) science education teacher's professional development; and (v) the STEM education movement. The significance of each of these factors for this dissertation is briefly described below.

1.1.1. Decline in enrollments

First, a shortage of graduates in science-related careers has been identified worldwide. Countries such as Australia or the United States of America have experienced in recent years a continuous decline in the number of students enrolled in these disciplines (Kennedy, Lyons, & Quinn, 2014; National Science Foundation [NSF], 2017). This lack of enrolments is also similar among many European Union countries (European Commission [EC], 2013), particularly in England (DeWitt & Archer, 2015), France (Rapoport & Thibout, 2018) and Netherlands (Buser, Niederle, & Oosterbeek, 2014). Spain is no stranger to this situation. Compared to 2004, in 2014 there were 23% fewer enrolments in scientific-related university degrees and 29% fewer enrolments in engineering and architecture (Ministerio de Educación, Cultura y Deporte [MECD], 2016). Furthermore, in Spain, gender segregation persists, with very few university graduates being women: only 25% and 2.5% in the case of Engineering & Architecture and Physics & Mathematics fields, respectively (Instituto Nacional de Estadística [INE], 2017). It is estimated that this situation could lead to a lack of scientifically literate citizenship and a dearth of a STEM workforce. These circumstances may compromise nations' ability to address complex problems facing the 21st-century, which in turn threatens global security (National Academy of Sciences [NAS], National Academy of Engineering [NAE], & Institute of Medicine [IOM], 2010).

1.1.2. Negative attitudes

Worries are dreadful. Coupled with a decrease in graduates in science-related careers, global trends of negative attitudes toward school science have been also identified. The international ROSE project (Sjøberg & Schreiner, 2010) concluded that in most developed countries students think of school science as less interesting than other subjects. Indeed, different attitudinal components such as lack of enjoyment (Palmer, Burke, & Aubusson, 2017), low levels of self-efficacy (Chachashvili-Bolotin, Milner-Bolotin, & Lissitsa, 2016; Sellami, El-Kassem, Al-Qassass, & Al-Rakeb, 2017) or perceiving science as irrelevant and not useful (Andersen & Ward, 2014) have been pointed out as detrimental for developing intentions to enroll in science-related studies. Recently, career choice and persistence has also been linked to achievement motivation factors such as expectancies of success (Guo, Parker, Marsh, & Morin, 2015).

In order to identify whether these patterns also occur in the context in which this dissertation was developed, the author of this dissertation conducted a set of preliminary studies investigating several variables that may influence student's intention to persevere in science-related studies, using elementary school students from the region of Castile and Leon, Spain. The results of these studies were used to create a baseline for the characteristics of the population under study in this dissertation, in relation to several critical constructs for scientific literacy.

1.1.2.1. Preliminary studies with elementary students

The first study explored the stereotypical image of scientists in 149 children of multicultural background, enrolled in 2nd to 6th grades of elementary education, using a modified protocol of Chamber's (1983) Draw a Scientist Test (Toma, Greca, & Orozco Gómez, 2018)[1]. The sample was comprised of Spanish students, Spanish students of gypsy ethnicity and second-generation Spanish students with roots from east-European countries.

In general, students had stereotyped conceptions of scientists, representing an image of the scientist as a male, especially in the case of boys, with Caucasian traits, who wear lab coats and who lonely conducts research in a laboratory with little to none interaction with other people. These stereotypes persisted in both girls and boys and was depicted in all grades investigated. In addition to stereotyped conceptions about the appearance or activities performed by scientists, the students in this sample also revealed beliefs about science as a very hard to achieve career, only suitable for very intelligent people. Therefore, it has been identified that these conceptions related to psychological components of scientists are engrained from elementary levels of the educational system, which could be holding back aspirations to pursue a career in science.

A second study investigated attitudes towards science and views of Nature of Science using the same sample as the Draw a Scientist study described above (Toma, Greca, & Orozco-Gómez, 2019)[2]. Data were collected using a modified version of the Test of Science Related Attitude instrument

---

[1] See Supplementary material 1

[2] See Supplementary material 2

developed and validated by Fraser (1981), and the Nature of Science instrument proposed by Hacıeminoğlu, Yılmaz-Tüzün, & Ertepınar (2012).

Regarding gender, boys had better attitudes toward Science than girls but more naïve views of the empirical aspect of Nature of Science. In relation to cultural background, second-generation Spanish students with east-European heritage reported significantly better attitudes toward Science than Spanish students and Spanish students of gypsy ethnicity, although no differences in Nature of Science views were found. Concerning grade level, 3rd graders had more positive attitudes toward Science than 5th and 6th graders and more informed views of the tentative Nature of Science. Therefore, this study concluded that there is a need to address the steady decline in positive attitudes toward Science and to improve students' views of Nature of Science from early elementary grades.

Therefore, since these two studies are consistent with international investigations suggesting that career aspirations in science start to develop at an early stage of the elementary education, with a considerable decrease in students' interest for science in upper grades and secondary education (Ali, Yager, Hacieminoglu, & Caliskan, 2013; Denessen, Vos, Hasselman, & Louws, 2015; DeWitt and Archer, 2015; Said, Summers, Abd-El-Khalick, & Wang, 2016), it is critical for Science Education to advance educational initiatives and policies to foster student's scientific vocations before they enroll in secondary education.

1.1.3. Science education reforms

Driven by the situation described above, science education reforms are being undertaken worldwide and inquiry-based pedagogies are being promoted as one of the main didactic strategies to reverse this situation. For example, in the United States, the National Research Council (NRC, 2000) postulated inquiry as a central strategy for science teaching and, more recently, the Next Generation Science Standards (NRC, 2013) promotes inquiry and scientific practices as a critical axis for science education in elementary school. In other countries, such as Chile, a new policy document suggests that inquiry should be used as a teaching strategy from the initial stages of elementary education (MINEDUC, 2012).

Likewise, the Spanish science education curriculum is slowly starting to promote inquiry-based science education through the three main curricula renovation advanced in the last two decades. Thus, in the 90s, the Ley Orgánica de Ordenación General del Sistema Educativo (LOGSE) [Organic Law on the General Organization of the Educational System] emphasized the need for scientific literacy through constructivism-based teaching and learning methods, and established a main subject called Conocimiento del Medio natural, social y cultural [Knowledge of the natural, social and cultural environment] that integrated content from both natural and social sciences disciplines (Ley Orgánica 1/1990, de 3 de octubre).

Years later, in 2006, the Ley Orgánica de Educación (LOE) [Organic Law of Education] was adopted in the Spanish educational system. While science education was still taught through the same subject as in the LOGSE law period, the LOE educational law promoted a competence-based curriculum and a global approach to learning. From a psychological point of view, this law proposes a didactic treatment that goes from global themes and knowledge to more analytical and specialized ones, thus presenting the reality through holistic lens instead of structured by areas of knowledge (Ley Orgánica 2/2016, de 3 de mayo).

Lastly, the Ley Orgánica para la Mejora de la Calidad Educativa (LOMCE) [Organic Law for the Improvement of the Quality of the Education] has recently been proposed and implemented in the Spanish education system. One of the main changes is that the Conocimiento del Medio natural, social y cultural school subject is now divided into two distinct subjects called Ciencias de la Naturaleza [Natural Sciences] and Ciencias Sociales [Social Sciences]. Through this change, greater attention is given to science education, which now has two weekly one-hour sessions specifically devoted to Natural Sciences disciplines, instead of the previous three hours that were shared with Social Science disciplines during the LOGSE and LOE laws. Another important change is the inclusion of a common standard for the six elementary grades (for students aged 6 to 12 years) called Initiation to scientific activity. This new standard explicitly advocates, for the first time in Spanish educational law, for inquiry approaches to scientific learning by planning and carrying out simple scientific investigations (Ley Orgánica 8/2013, de 9 de diciembre).

1.1.4. Professional development

The educational reforms that are being promoted will hardly have a significant educational impact if teachers do not have the required knowledge and skills to use innovative, inquiry-based teaching methodologies. There are numerous studies stressing that given the limited professional development in science of elementary school teachers, any science education reform that goes beyond the use of the textbook as the main teachings strategy will create many challenges because most teachers do not have the knowledge and skills necessary to use active-based pedagogy reflectively and appropriately in the classroom. In fact, research results reveals that elementary teachers hold fragmented, superficial and flawed scientific knowledge, which leads to traditional, lecture-based lessons for teaching science (Cañal 2000; Porlán et al., 2010). Similarly, previous research suggests that pre-service teachers struggle with proposing teaching units that are in line with inquiry-based approaches for science education (Seung, Park, & Jung, 2014).

In order to determine to what extent this aspect is present in the Spanish educational system, and more specifically within the context where this doctoral thesis has been developed (i.e., Burgos province), the author of this dissertation conducted a preliminary study on pre-service teaching ability to design inquiry-based science teaching units, described below. The results of this study were central for the design of the pedagogical framework for STEM-based science education proposed and tested in this dissertation.

1.1.4.1. Preliminary study with preservice teachers

The purpose of this preliminary study was to identify elementary school pre-service teachers' difficulties when developing inquiry-based teaching units for science education (Toma, Greca, & Meneses Villagrá, 2017). To this end, up to 157 teaching units designed by students in the final year of the University Degree in Primary Education at the University of Burgos have been analyzed. First, following Schwarz & Gwekwerere (2007), the teaching units have been categorized into a continuum from didactic units that promote traditional, teacher-centered lessons, to units that promote inquiry-based practices. Subsequently, the teaching units categorized as inquiry-based were evaluated using the Reformed Teaching Observation Protocol (RTOP), developed by Sawada & Piburn (2000), to detect the degree of coherence of the teaching units, and those aspects that are more difficult to integrate into an inquiry-based science unit.

The findings suggested that only 22% of the didactic units designed promoted an inquiry-based approach. Likewise, the application of the RTOP instrument revealed that three-quarters of the units designed to promote an inquiry-based approach were developed in a very guided way, which did not allow students to propose different research questions or design their own procedures for testing the formulated hypothesis. In other words, even though inquiry was promoted, students were limited to carrying out the hands-on activities already designed by the teachers. Therefore, these results highlight the need to improve elementary school teacher's professional development in science so that they can be able to implement those educational measures that are being demanded in the different educational reforms discussed in previous sections.

1.1.5. The STEM momentum

Finally, in this scenario characterized by continuous science education curricula reforms in order to ensure scientific literacy and increase the number of students interested in pursuing science-related careers, an educational movement, called STEM, emerged in science education policy discourse. This acronym stands for Science, Technology, Engineering, and Mathematics, and is gradually monopolizing attention in the science education panorama. Nevertheless, although the STEM movement has gained momentum in recent years, the research and educational agenda developed around this acronym is plagued with ambiguities and suffers from several critical limitations that compromise its educational value.

Hence, under the STEM acronym, diverse and sometimes contradictory educational initiatives are being promoted, including a greater emphasis on STEM coursework, improvement of STEM curricula through educational reforms, and a global call to abandon the individualized treatment of each disciplines in favor of the adoption of integrated approaches that emphasize connection across STEM school subjects and disciplines (English, 2016; Johnson, 2012; Kelley & Knowles, 2016). As a result, the dominant STEM discourse states that "(…) these subjects cannot and should not be taught in isolation, just as they do not exist in isolation in the real world or the workforce" (STEM Task Force Report, 2014, p. 7). It is further argued that integrative approaches may increase student's perception of the relevance of STEM subjects and foster favorable attitudes towards learning and engaging in STEM-related fields (NAE & NRC, 2014). This call for STEM integration through real-world problem-based learning in authentic environments has been echoed in recent international reforms, such as the Next Generation Science Standards (NRC, 2013) and the Common Core State Standards for Mathematics (National Governors Association Center for Best Practices [NGACPB], 2010).

In order to identify the viability of such an educational framework in the Spanish context, the author of this dissertation conducted a pilot study on the implementation of STEM-based science instruction.

1.2. Statement of the problem

Educational reforms aimed at improving elementary school students' scientific literacy, attitudes towards science, and intention to enroll in science-related careers are heavily promoted worldwide, as well as by the last Spanish educational reform. As a result, the STEM acronym is being adopted to refer to initiatives that are in tandem with these educational policies. Consequently, self-identified STEM schools and STEM-based curricula continue to grow daily, with many in and out-of-school STEM programs and initiatives mounting.

However, the STEM agenda is riddled with ambiguity as there is no clear conceptualization about what constitutes STEM and how to best implement it in the Spanish educational system. A common operational definition of STEM is missing (Breiner, Harkness, Johnson, & Koehler, 2012) and research about the implementation of integrative STEM is scarce and suffers of "(…) inconsistent use of language, failure to define terms, and lack of a theoretical framework for understanding integrated STEM education" (NAE & NRC, 2014, p. 138). Due to the plethora of STEM conceptualizations, lack of cohesive understanding, and dearth of research-based programs, STEM is at risk of being considered only as the neologism for science education. Indeed, some critical voices point to a STEM-ification of science education and consider STEM as an "(...) ideological positioning of science education rather than anything evidence-based" (Carter, 2017, p. 2) used "(…) to press for particular ideologically based improvements to public science reimagined as STEM" (p. 9).

In short, there are many aspects of the STEM movement that require in-depth reflection and investigation to determine whether this approach is appropriate for improving the scientific literacy of students, as well as for promoting positive attitudes towards science and intentions to pursue science-related studies. Otherwise, science education efforts under the STEM umbrella would run the risk of being nothing more than a deficit educational model that does not advance in the resolution of the problems facing the 21st-century science education.

1.3. Aim and research questions

In order to advance the ubiquitous journey of using STEM initiatives for preparing the current and next generation of citizens, vigorously promoted by the science education community and by stakeholders and policymakers, it is paramount to first articulate those theoretical and pedagogical underpinnings that such an educational paradigm should have, and then conduct rigorous educational investigations to identify the strengths and weaknesses of STEM-based science education. Therefore, the main aim of this dissertation is to contribute to the development of a coherent STEM framework that advances the teaching and learning processes that should take place when such approaches are being implemented in elementary science education, and to assess its impact on two educational outcome variables deemed important for scientific literacy, namely, expectancies of success and attitudes towards school science. Consequently, this dissertation intends to examine the following research questions:

Main research question	1. What are the characteristics of a valid, practical, and effective pedagogical framework that facilitates the implementation of STEM-based teaching units in elementary science education for promoting students' expectancies of success and attitudes towards school science?
Secondary research questions	1.1. What teaching strategies should be used to implement valid and plausible STEM-based teaching units for science education in elementary grades? 1.2. Does STEM-based teaching units promote students' expectancies of success and attitudes towards school science?
Tertiary research questions	1.2.1. For upper grades students, which teaching approach (structured vs. guided inquiry), if any, is more effective? 1.2.2. Is there any differential effect based on gender or grade level variable? Do these results hold when controlling for pre-intervention scores?

This is certainly a commendable goal, so it is not expected, nor claimed, that this doctoral thesis has fully achieved this task. However, it is hoped that progress has been made towards conceptualizing STEM as an educational paradigm, and in developing research-based didactic principles that guide the design of STEM-based teaching units that can be used to evaluate the appropriateness and effectiveness, if any, of STEM education.

1.4. Research design

This doctoral thesis adopted a mixed-methods multiphase design (Creswell & Clark, 2011). The multiphase design consists of an iteration of quantitative, qualitative, and mixed-methods studies that are sequentially conducted. Each new approach or study builds on what was previously learned and thus advance in addressing the central research question or project objective. This design is appropriate when there are a set of incremental research questions that address the same programmatic research objective, although cannot be fulfilled with a single study. Likewise, it is also useful when new research questions emerge as the project or study is being conducted. Therefore, the multiphase design provided an overarching methodological framework for this multiyear research dissertation.

Throughout this doctoral thesis, nine different, but closely related studies will be described. Although each one of these studies has their own research questions and objectives, they were specifically conducted so that they contributed to addressing the research questions introduced above. The relation between these studies is reported in **Figure 1**, which reads from the left side to right one.

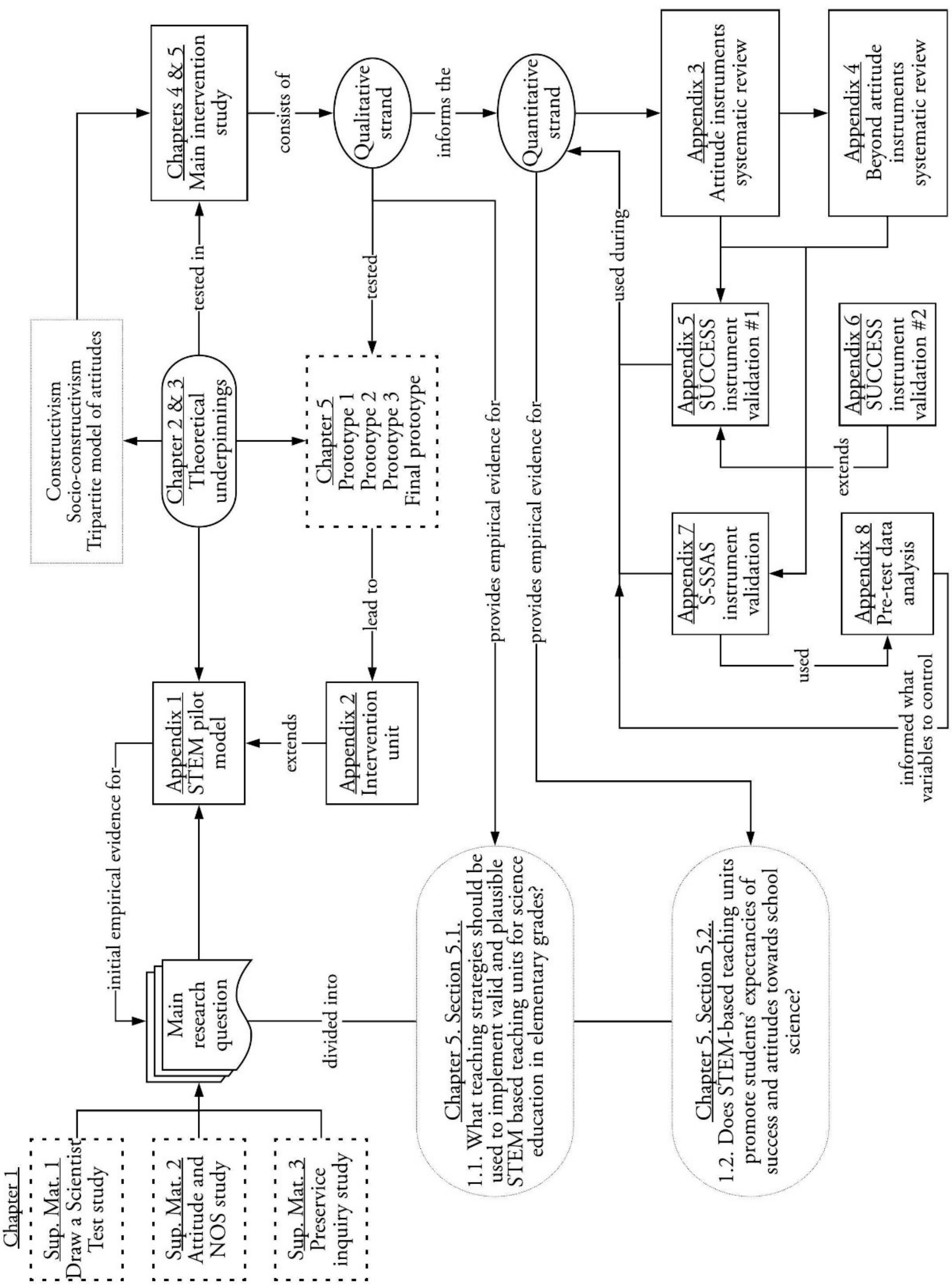


Figure 1. Pedagogical framework proposed for STEM-based science education

The first study introduces a pilot model of STEM education developed to test whether STEM education is viable in the Spanish educational system. This study uses a mixed-method embedded-design, which consists of collecting and analyzing both quantitative and qualitative data within a traditional quantitative design (Creswell & Clark, 2011). The quantitative data was collected using a Likert-type attitudinal scale and an achievement test, administered to elementary school students using a one-group pretest-posttest quasi-experimental design (Campbell & Stanley, 1963). The qualitative data was collected using structured interviews with in-service teachers. The quantitative analysis provided evidence about the effectiveness of STEM education in students' achievement and attitudes towards science. The qualitative analysis enhanced the overall design by adding information about teachers' perception of STEM-based education. This study is introduced in Chapter 3 and the full-text published manuscript is attached in Appendix 1.

The second study describes the intervention unit developed following the pedagogical approach for STEM education, that builds on the first study, and that is introduced in Chapter 3 of this dissertation. This study uses a qualitative approach by describing the didactic transposition of constructivist and socio-constructivist learning principles to the design of a teaching unit suitable for teaching science in an integrative way to students enrolled in 3rd to 6th elementary grades, in Spain. The theoretical underpinnings of this study are discussed in Chapter 2 and 3, and the full-text published manuscript is attached in Appendix 2.

The third and fourth studies report the results of two systematic reviews conducted in order to select the most appropriate measurement instruments for the main study. Both studies adopted a qualitative approach following the PRISMA statement for systematic and meta-analysis studies (Liberati et al., 2009). More specifically, the first systematic review analyzed the validity and reliability evidences of attitude towards science instruments. Next, building on this study, the second systematic review analyzed instrument development and validation practices in science education research and the psychometric properties of instruments measuring variables related to attitudes, such as motivation or self-efficacy. Both reviews are introduced in Chapter 4; likewise, the full-text accepted for publication manuscript of the first systematic review is attached in Appendix 3, and the full-text under review manuscript of the second systematic review is attached in Appendix 4.

The fifth study presents the development of the SUCCESS instrument, used in this dissertation to assess the efficacy of the pedagogical framework for STEM-based science education described in Chapter 3, and pedagogically transferred in the intervention unit attached in Appendix 1, on improving students' expectancies of success in school science. This study uses a mixed-methods exploratory sequential design, which consists of two interactive phases, being the first one qualitative and the second one quantitative (Creswell & Clark, 2011). The qualitative phase was used to develop the initial pool of items and by establishing its content validity using a penal of experts and cognitive interviews with the target population. Next, building on this exploratory first phase, the pool of items was administered to a large sample size. This second, quantitative phase, was used to examine the psychometric properties of the SUCCESS items in terms of validity and reliability evidences. This study is introduced in Chapter 4 and the full-text in review manuscript is attached in Appendix 5.

The sixth study further examines the psychometric properties of the SUCCESS instrument, using a survey-type quantitative design. This study is likewise introduced in Chapter 4 and the full-text published manuscript is attached in Appendix 6.

The seventh study describes the cross-cultural validation of the School-Science Attitude Survey (S-SSAS), originally developed by Kennedy, Quinn, & Taylor (2016). This instrument was used in this dissertation to examine the impact of the pedagogical framework for STEM-based science education on elementary school students' attitudes towards school science. This study adopted a mixed-methods embedded design. The main phase consisted of a survey-type quantitative design in which the psychometric properties of the Spanish version of the S-SSAS was examined, enhanced with a qualitative strand that examined the content validity and item interpretability of the students by conducting a panel of experts analysis and cognitive interviews with the target population. This study is introduced in Chapter 4 and the full-text published manuscript is attached in Appendix 7.

The eighth study consists of an analysis of the pre-test data collected during the implementation of the intervention unit. In this study, a survey-type of quantitative design was adopted by analyzing, through a person-centered approach, students attitudes towards school science. The data for this study was collected using the S-SSAS instrument. The analysis and results revealed the need to control for gender and school grade variables when analyzing the quantitative results from the main intervention study, described below. This study is introduced in Chapter 5, and the full-text published manuscript is attached in Appendix 8.

Finally, the ninth study consists of the main intervention performed in this doctoral thesis. More specifically, this study adopts a mixed-method exploratory design, in which the first, qualitative phase was devoted to the formative evaluation of the intervention unit using educational design research as a qualitative framework. Next, the second, quantitative phase consisted of the effectiveness evaluation through a one-group pretest-posttest quasi-experimental design (Campbell & Stanley, 1963). The method and design of this main study are in-depth described in Chapter 4, and the qualitative and quantitative findings are reported in Chapter 5. Lastly, these findings are discussed in Chapter 6, with special emphasis on interpreting the results in light of all previous studies.

Chapter 2. Theoretical framework

This chapter presents the theoretical framework underpinning this thesis dissertation. First, section 2.1. briefly reviews the origin and development of the STEM acronym, as well as recent studies conducted under the STEM umbrella. Next, section 2.2. discusses the theoretical underpinning and the existing literature related to the two educational constructs (i.e., expectancies of success and attitudes towards school science) examined in this dissertation to determine the effectiveness of the pedagogical framework for STEM-based science education proposed.

2.1. STEM and STEM education

The STEM acronym was initially originated in the 1990s by the National Science Foundation (NSF) as a "(…) strategic decision made by scientists, technologists, engineers, and mathematicians to combine forces and create a stronger political voice" (STEM Task Report, 2014, p. 9). The aim was to give more prominence to the disciplines of science, technology, engineering and mathematics (Sanders, 2009). The STEM acronym has gained strong political value and nowadays is embraced as a slogan for policymaker and advertisement of conferences, curricula and classroom resources (Bybee, 2013). Nonetheless, its meaning remains unclear and currently, there is a wide spectrum of conceptualizations, which compromise its educational value and hinders its real impact on educational curricula and teaching practices. Keefe (2010) concluded that the general public confuses the STEM acronym with stem cell research or plants, and Breiner et al. (2012), after examining STEM conceptions of faculty members, concluded that there is a lack of a common operational definition of STEM, with conceptualizations strongly resembling STEM as individual disciplines and with most members failing "(…) to demonstrate an understanding (…) even if they held academic appointments in STEM colleges and/or programs" (p. 9).

Consequently, from a policy perspective, STEM seems to be a slogan that calls for an improvement of science and mathematics education to fuel international competitiveness and economic security by fostering more graduates in these disciplines, retain talent, and recruit a talented STEM workforce from around the world (Breiner et al., 2012). From an educational standpoint, STEM is conceptualized through a broad continuum that moves from traditional disciplinary coursework (i.e., science, mathematics, technology, and engineering as separate disciplines) to recent conceptions that refers to STEM as the integration of science, technology, engineering, and mathematics curricula to closely resemble how STEM knowledge is developed and used in real life. In words of Dare et al. (2018), "STEM content should not be taught in isolation, but rather in a way that reflects how STEM knowledge is used outside of school; this knowledge is further contextualized or driven by some problem or issue" (p. 4).

This notion of STEM as curriculum integration places the emphasis on considering the four disciplines as a cohesive entity (Cunningham, 2014, 2018; English, 2016; Kelley & Knowles, 2016; Vasquez, Sneider, & Comer, 2013), which provide opportunity to teach and learn in real-world multidisciplinary contexts (Dare, Ellis, & Roehrig, 2018). Similarly, Bybee (2013) offered different conceptualizations of STEM education ranging from discipline-based perspectives to true integration of all four disciplines. For this author, STEM literacy encompass three aspects: (i) conceptual, that involves the understanding and knowledge of STEM-related disciplines, (ii) skills, which relate to the ability to apply STEM knowledge and procedures to real-world questions and problems, and (iii) attitudes, which refer to the motivation for engaging in STEM-related issues and the adoption of STEM-based ideas and reflective stances. Kelley & Knowles (2016) conceptualized STEM as "(…) the approach to teaching the STEM content of two or more STEM domains, bound by STEM practices within an authentic context for the purpose of connecting these subjects to enhance student learning" (p. 3). Moore et al. (2014) referred to STEM education as "(…) an effort to combine some or all of the four disciplines of science, technology, engineering, and mathematics into one class, unit or lesson that is based on connections between the subjects and real-world problems" (p. 38) and established that successful STEM education is characterized by (i) motivating context, (ii) mathematics and science content inclusion, (iii) student-centered, (iv) inclusion of engineering design, (v) teamwork and communication, and (vi) conceive failure as a learning opportunity through redesign. Recently, the letter A, referring to "arts", has been added to the acronym, turning it from STEM to STEAM (STEM + Arts), which refers to the incorporation of creative thinking in problem-solving and the development of creativity (Kim & Chae, 2016; Kim, Lee, Yang, Lee, Jang, Kim, 2019).

Although closely linked with recent definitions of STEM, the notion of curricula integration is not new and dates to the Aikin's Eight-Year Study about the reconstruction of the secondary school curriculum through core-type programs characterized by a student-centered, cross-disciplinary approach who addressed the need for making connection across subjects (Aikin, 1942; Pinar, 2010). Combined with the lack of consensus on what constitutes STEM, misunderstandings increase as perspectives on STEM as curricula integration are also gaining momentum. Researchers and policy developers are even more confounded about the existing perspectives on integration and the forms it should take in the classrooms as approaches to integration varies also from disciplinary-based integration to transdisciplinary-problem based approaches (Gresnigt, Taconis, van Keulen, Gravemeijer, & Baartman, 2014; Hurley, 2001; Vasquez et al., 2013).

In relation to defining integration, the National Academy of Engineering and the National Research Council (2014) conceptualized it as "(…) working in the context of complex phenomena or situations on tasks that require students to use knowledge and skills from multiple disciplines" (p. 52). Beane (1997) defined integration as the curriculum planning around problems based on the real world without concerning about standardized subjects' areas or content. The STEM Task Force Report (2014) stated that convenient integration of the four disciplines is not enough and addressed the need for "(…) cohesive and active teaching and learning approaches" (p. 9) in linking STEM disciplines.

Regarding approaches to integration, Haggis & Adey (1979) referred to coordinated, combined and amalgamated form of integration depending on the extent to which the disciplines are taught in combination or not. Fogarty (1991) described and categorized ten ways to integrate curriculum within single disciplines (i.e., fragmented, connected and nested), across several disciplines (i.e., sequenced, shared, webbed, threaded and integrated), and within and across learners (i.e., immersed and networked). Hurley (2001) established science and mathematics integration along a continuum of five levels: sequenced (planned and taught sequentially), parallel (simultaneously planned and taught through parallel concepts), partial (planned and taught separate and together), enhanced (one of the subjects is the dominant while the other is apparent), and total form of integration (planned and taught together equally).

More recently, after reviewing existing literature on curriculum integration, Gresnigt et al. (2014) proposed a taxonomy for integration based on the extent to which educational elements such as goals, lessons, explanations or grades are being shared between curricula subjects. These authors listed a six-level hierarchical taxonomy with integration approaches being fragmented, connected, nested, multidisciplinary, interdisciplinary, and transdisciplinary, and concluded that the greater the form of integration, the greater the possibilities of fostering students' positive attitudes towards the curriculum, developing 21st-century skills and awakening the enthusiasm of teachers. On the other hand, Gresnigt and colleagues warned that greater curricula integration also requires more commitment, professional development, teacher support, and necessary resources.

In short, the literature refers to STEM education both from a point of view centered on each discipline and from an integrating perspective that calls for the interdisciplinary teaching of these disciplines. As Bybee (2013) stated, "Many request a definition, and few agree with one when it is presented" (p. X).

Given the lack of consensus in defining STEM education, before advancing it is necessary to clarify how STEM is conceptualized in this dissertation. As mentioned above, in the literature STEM is conceived as integrating at least two disciplines. However, this approach adds nothing new to the Science & Mathematics integration approach promoted in the 1970s and 1980s, whose educational value is, at best, equivocal (Czerniak, Weber, Sandmann, & Ahern, 1991). Therefore, in this dissertation, it is argued that STEM initiatives must integrate the four disciplines in a synergistic way so that the result of the integration is more than the mere sum of its parts. Likewise, contrary to existing studies conceptualizing STEM as a teaching methodology (i.e., Martín-Páez, Aguilera, Perales-Palacios, & Vílchez-González, 2019), in this dissertation STEM is conceived as an educational paradigm that seeks to establish connections between the four STEM disciplines. Thus, the definition of STEM that is advanced in this thesis is:

❖ *STEM is an educational paradigm aimed at intentionally reducing the boundaries between Science, Technology, Engineering, and Mathematics by fostering holistic teaching and learning episodes based on authentic transdisciplinary real-world problems wherein the similarities, differences, and relationship between the concepts, procedures, and skills of each one of these disciplines are fostered and explicitly established.*

2.1.1. STEM education empirical literature

Several empirical studies have examined the impact of a STEM approach on different educational constructs, such as achievement, engagement, and interest. Thus, Kim, Kim, Yuan, Hill, Doshi, & Thai (2015) reported that STEM lessons using robotics improved elementary education pre-service teachers' interest and enjoyment. Barak & Assal (2018) concluded that using robotics within the STEM framework provides attractive learning environments, and Chen & Chang (2018) found that students participating in an integrated robotics STEM course outperformed student in traditional courses in terms of knowledge acquisition and development of interest and orientation towards STEM-related careers. Likewise, Psycharis & Kotzampasaki (2019) study revealed that using computational tools (i.e., microcontroller boards) during integrative STEM lessons improves student's computational thinking skills.

Some studies adopted an engineering-based approach to STEM, concluding that STEM learning through engineering design improve middle-secondary students' interest towards STEM subjects and careers (Shahali, Halim, Rasul, Osman, & Zulkifeli, 2017). Likewise, Lou, Chou, Shih, & Chung (2017) concluded that STEM projects with a focus on design might be helpful in improving student's creativity. Other studies integrated STEM education with project-based learning, concluding that this approach can enhance secondary school students learning of STEM knowledge and facilitate the integration of the STEM disciplines (Tsai, Chung, & Lou, 2018), and improve attitudes towards the subject of engineering in freshmen students (Tseng, Chang, Lou, & Chen, 2013). Finally, at K-5 educational level, Lamb, Akmal & Petrie (2015) suggested that STEM integrated learning may improve self-efficacy, science interest, spatial visualization and mental rotation (i.e., transformation process related to the ability to manipulate and rotate mentally the presented stimuli).

While these results provide initial empirical evidence supporting the effectiveness and relevance of STEM approaches, these findings should be interpreted with caution due to the multiplicity of STEM approaches used in the reviewed studies. In this regard, Martín-Páez et al. (2019) have identified that the acronym STEM appears in empirical studies addressing from one discipline in isolation to interdisciplinary approaches that attempt to integrate all four disciplines that make up the acronym.

2.2. Attitudes

Defining attitudes is not an easy task as attitudes are not directly observable or measurable. Despite the existence of multiple different conceptualizations, there is a general agreement on conceiving attitudes as a multifaceted rather than unidimensional construct, which encompasses cognitive, affective and behavioral aspects (Khine, 2015; Simpson, Koballa, Oliver, & Crawley, 1994). Shaw & Wright (1968) considered that beliefs about a particular object (i.e., cognitive component) leads to affective reactions towards that particular object (i.e., affective component) and that this may affect the behavior towards that object (i.e., behavioral component). Therefore, for these authors, there is a three-stage process that goes from the cognitive dimension (beliefs), then to the affective one (affective reactions), and then to behaviors (predisposition to act in one way or another). Similarly, Oppenheim (1992) associated the cognitive aspect to beliefs about a certain object and the affective component, much more stable and profound than the cognitive one, to the values related to that object. Therefore, attitudes are conceived as individual's favorable or unfavorable evaluations formed as individuals develop beliefs about the object under study (Ajzen, 1991; Fishbein & Ajzen, 1975). In words of Eagly & Chaiken (1995), attitudes can be defined as "(…) a psychological tendency that is expressed by evaluating a particular entity with some degree of favor or disfavor" (p. 414). Thus, attitudes encompass those evaluative aspects of a cognitive, affective and behavioral nature referring to beliefs, thoughts, feelings, and emotions towards a given object (e.g., science) or a behavior related with the object (e.g., studying science).

2.2.1. Attitudes towards science

Along with the development of literate scientific citizenship, the development of positive attitudes towards science is likewise a critical objective included within the scientific literacy agenda. The importance of attitudes towards science is reflected in the literature reporting a strong relationship between enjoyment of science lessons with expressed intentions to further engage in science learning experiences (Ainley & Ainley, 2011a; 2011b), and in the positive relationship between interest in science and performance and achievement (Bybee & McCrae, 2011; Newell, Tharp, Vogt, Moreno, & Zientek, 2015). Given that students' attitudes toward science may affect their science performance, their desire to study science in the future, and the choice of science as a career (Koballa, 1988), attitudes have perennially been a very significant line of research in science education, especially in recent years given the steady decline of individuals interested in studying science in secondary and tertiary education (Archer et al., 2010; Kennedy et al., 2014; Potvin & Hasni, 2014).

Although in recent decades studies focused on diagnosing students' attitudes toward science have proliferated considerably, the concept under study remains unclear. Indeed, Osborne, Simon, & Collins (2003) stated that attitudes toward science are "(...) somewhat nebulous, often poorly articulated and not well understood" (p. 1049), a problem that has been warned since more than three decades (Munby, 1983). In this regard, Klopfer (1971) offered insight about the concept under study by categorizing the attitude toward science construct as a set of affective behaviors toward science as an enterprise, scientists, scientific inquiry, scientific careers and towards science-related activities in general. Similarly, Schibeci (1983) concluded that attitude toward science deals with several attitude objects, such as science as a discipline, the study of science, science lessons, among many others. Gardner (1975) provided further clarity by addressing the differences between "scientific attitudes", conceived as those elements inherent to scientific thinking and research, and "attitudes toward science", conceptualized as the sociological, psychological and affective conceptions about science. In short, attitudes towards science comprise various aspects, such as science in general, the subject of school science, scientific careers, science teachers or the teaching of science, with each one of these aspects being characterized by many more detailed facets (Kind, Jones, & Barmby, 2007; Toma, in press; Toma, Lederman, & Meneses Villagrá, under review[3]).

Consequently, past studies focusing on attitudes toward science have incorporated many constructs such as anxiety toward science, self-esteem in science, motivation toward science, enjoyment of science (Tytler & Osborne, 2012), students affective feelings and cognitive judgments of science (Zhang & Campbell, 2011), unfavorable outlook of science (Abd-El-Khalick, Summers, Said, Wang, & Culbertson, 2015), and perception of scientists and value of science to society (Hillman, Zeeman, Tilburg, & List, 2016), among many others.

2.2.2. Attitudes towards school science

It is important to emphasize that in this dissertation the attitudinal object under investigation is the domain of school science instead of science in general. First, of the STEM disciplines, the science subject is the closest to student's awareness at this educational stage, so it seems to be more realistic and reliable to measure attitudes towards a subject they already know instead of, for example, engineering, which is a discipline of which they presumably had no knowledge or past experiences at all.

Secondly, because in the Spanish educational system, individuals' interest in studying a STEM-related university degree should first enroll in science subjects (i.e., biology, geology, physics, and chemistry) at the secondary school stage, which would later lead to a scientific and technological baccalaureate specialized for a career in STEM (LOMCE, 2013). Therefore, it seems appropriate to assume that students ending the elementary education stage with negative attitudes towards school science (not science in general) will avoid enrolling in scientific-type subjects later on in the educational system, thus falling out of the STEM pathway from very early ages.

Finally, because behaviors towards engaging with the target object (i.e., studying school science), rather than toward the object itself (i.e., science as an enterprise) are reported to be better predictors of individual behaviors (Ajzen & Fishbein, 2005; Fishbein & Ajzen, 2010). Therefore, students may have a positive attitude towards science as an enterprise, but negative attitudes towards learning school science, which may affect their intention to pursue science-related studies despite their favorable attitudes towards science in general. The decision on focusing on studying school science instead of science, in general, is further reinforced by the above argument on STEM pathways; thus, certainly a student will be more likely to enroll in science subjects during high school and baccalaureate if he or she has positive attitudes towards studying science at earlier stages of the educational system.

2.2.3. Expectancies of success in school science

Although different attitudinal components have been pointed out as detrimental for developing intentions to enroll in further science, research on attitudes conducted in science education is affected by a lack of conceptual clarity of the target construct and by methodological issues derived from

[3] See Appendix 3 and 4

poorly developed and unreliable measurement instrument (Blalock et al., 2008; Gardner, 1975; Munby, 1997), even in instruments developed in recent years (see Appendix 3 and 4). One prominent theory that can be used to advance research in understanding students (dis)engagement with science-related studies is Jaqueline's Eccles and colleagues Expectation-Value theory (EVT), which aims to explain the relationship between motivations and academic achievement by linking performance, persistence and choice of a given task or activity with individual's beliefs about expectancies of success and the value placed on that task or activity (Eccles et al., 1983; Eccles, Wigfield, Harold, & Blumenfeld, 1993; Wigfield & Cambria, 2010). Recently, the EVT model has been used to explain the selection of STEM subjects (Andersen & Ward, 2014), so it is postulated as an ideal theoretical framework for examining disinterest in STEM-related degrees.

The first key construct of the theory, which is the one of interest in this dissertation, is "expectancies of success", defined by Eccles et al. (1983) as individual's beliefs about the success he or she will have in a given task or activity, whether in the immediate or long-term future. Conceptually, there are two dimensions that underlie individuals' expectancies of success: (i) beliefs in one's own ability, which refers to the individual's perception of his or her current competence to perform a given activity in the present, and (ii) expectations, which refers to the individual's perception of successfully performing that given activity in the future. However, studies have shown that beliefs and expectations are highly correlated and cannot be differentiated empirically (Wigfield & Eccles, 2000). Therefore, expectancies of success include internal beliefs about both the ability to perform a task in the present and in the future.

Using a hypothetical example related to this dissertation, a student enrolled in elementary education may feel highly competent in learning science, which may lead to the development of positive beliefs about his or her ability to succeed in future science courses or subjects. This high expectancy of success may foster the selection of science subjects during secondary education or the engagement with extracurricular science activities; eventually, this individual may pursue a science-related degree.

2.2.4. Attitude toward science empirical literature

It is difficult to establish patterns about the attitudes of primary school students due to the disparity of constructs that have been studied as a reflection of attitudes towards science. Thus, while some studies have focused on studying students' attitudes towards the "social implication of science", "enjoyment of science", or "interest in scientific issues" (Ali et al., 2013; Denessen, Vos, Hasselman, & Louws, 2015), others have addressed more general aspects, such as pressure from teachers and parents to be successful in school science subjects (George, 2006). However, there seems to be a trend of more unfavorable attitudes among girls than boys (Caleon & Subramaniam, 2008; Denessen et al., 2015; DeWitt & Archer, 2015; George, 2006; Hacıeminoğlu, 2016), as well as a general decline in favorable attitudes towards science in higher school grades (Akpinar, Yildiz, Tatar, & Ergin, 2009; Ali et al., 2013; Denessen et al. 2015; DeWitt & Archer, 2015; George, 2006; Said et al., 2016).

In the Spanish context, most studies have focused on the Secondary stage (e.g., Arandia, Zuza, & Guisasola, 2016; Pérez-Franco, de Pro Bueno, & Pérez, 2018; Vázquez & Manassero, 2009), reporting results that are more positive when compared to the international literature. However, gender differences, favoring boys, have also been identified in the Spanish contexts. Thus, among studies that have included Primary Education students in their sample, girls seem to have a less positive attitude towards science than boys (de Pro Bueno & Pérez, 2014; Marbá-Tallada & Márzquez, 2010; Pérez & de Pro Bueno, 2018; Vázquez & Manassero, 2008). Likewise, although less pronounced than in other countries, as grade level increase, Spanish students positive attitudes tend to decrease (de Pro Bueno & Pérez, 2014; Marbá-Tallada & Márquez, 2010, Vázquez & Manassero, 2008). However, it should be noted that these differences are not as accentuated or intense as some international results; hence, in the Spanish context, there are studies that have found no association between attitudes towards science and gender or grade level variables (e.g., Fernández-Cézar & Pinto-Solano, 2017).

Chapter 3. Pedagogical proposal

This chapter describes the pedagogical framework for STEM-based science education proposed in this doctoral thesis. First, section 3.1. discusses a study on a pilot STEM model for elementary education. Next, section 3.2. presents the psychological underpinnings, based on constructivists and socio-constructivists learning theories, of the pedagogical framework for STEM-based science education developed in this dissertation by extending and modifying the pilot model introduced in section 3.1. Then, section 3.3. introduces the main characteristics of the pedagogical framework, discussing how each letter (and therefore, discipline) of the STEM acronym is conceptualized at an educational praxis level in the light of the constructivists and socio-constructivists theories described. Afterward, section 3.4. addresses the didactic transposition of the proposed pedagogical framework to the design of teaching units in elementary education. Finally, section 3.5. summarizes the intervention unit and provides specific examples of how the pedagogical framework proposed was implemented in this dissertation.

3.1. Background: A pilot model of STEM education

A pilot model of STEM education was designed and tested in real-world classroom to examine the impact if any, this type of integrative teaching approaches would have on students' attitudes towards Science and achievement, and to inspect in-service teachers' opinions about STEM approached to science education (Toma & Greca, 2018). In this section, an extended abstract of this study is presented; the full published article is attached in Appendix 1.

A STEM prototype consisting of four phases that seeks to encompass each STEM discipline was developed. This STEM model draws on Martin-Hansen's (2002) coupled inquiry that combines a guided and open inquiry investigation within one project. Thus, in the first phase, entitled inquiry invitation, the teacher proposed an engineering-based, real-world problem that served as a context to teach science-related content matter. During the second phase, students engaged in a guided inquiry in which they conducted different hands-on activities using scientific practices and technology, and interpreted data using mathematics. The third phase consisted of an open inquiry during which students discussed the results obtained in the guided inquiry and proposed and developed new research questions necessary to solve the initial problem, with less guidance from the teacher. The fourth and final phase, named inquiry resolution, required the design or implementation of a technological solution.

Following this conceptualization of STEM education, a unit on simple machines for fourth graders was designed and tested. During the inquiry invitation phase, the teacher introduced the following engineering-based problem: How did the Egyptians transport the stone blocks for building the pyramids without using modern machines? Next, in the guided inquiry, the teacher indicated students how to construct three simple machines (i.e., pulleys, inclined planes, and levers) using LEGO™ building blocks. Students performed several hands-on activities to determine the force needed to move a small LEGO block using these simple machines. Students discussed the results obtained and the physics laws that govern each machine were deduced with teacher facilitation. In the third phase, students engaged in an open inquiry and investigated new variables that might have an impact on the force that is necessary in order to move the pyramid block using simple machines, such as the ruggedness of the inclined plane surface, the length of the inclined plane, or if a simple or compound pulley is used. Finally, during the last phase, students modeled the route of a stone block from its origin to the hypothetical pyramid, pointing out the way in which simple machines would have been used.

This unit was tested in two conventional classrooms from one elementary school located in an urban area of the city of Burgos, in Spain. Two classes from another (yet similar in characteristics) school formed the control group that studied the sample simple machine unit through traditional, teacher-centered pedagogies. The Test of Science Related Attitudes (TOSRA) scale developed by Fraser (1981) was adapted and used to assess the influence of the pilot STEM model on students' attitudes toward science. In addition, an achievement test with 25 questions was developed following teachers' criteria and standards of evaluation. Finally, teachers from the treatment group were interviewed to gather information about their thoughts on teaching science through STEM.

Findings revealed that after the implementation of the pilot model, students included in the treatment group reported more positive attitudes towards science than control group students did. More specifically, they reported higher scores on attitudes towards scientists, scientific inquiry, adoption of scientific attitudes, enjoyment of science lessons, and leisure interest in science constructs. Regarding gender, there were no significant differences between and within each treatment or control group. Similarly, the achievement test suggested that this prototype might foster science learning, with all students improving their content knowledge about simple machines. As for the interviews, both teachers argued that, although they were surprised by students' results and the classroom atmosphere during the STEM project, they would not use it as the main approach for science teaching due to lack of knowledge and because the design of a STEM unit was too time demanding.

The results of this study reinforce the need to advance in the development of a suitable framework for implementing STEM education in the Spanish educational system. In addition to the logistical limitations pointed out by the teachers who have participated in its implementation, this pilot study on STEM also revealed limitations in terms of conceptualizing STEM education. Hence, it should be acknowledged that in this pilot study, STEM was represented mainly as Science and Mathematics. The referent of this model was the school science content and the scientific procedures (i.e., inquiry), slightly complemented with mathematical procedures used for the analysis and interpretation of the data. Engineering was only anecdotally present in the form of an initial problem that requires a resolution. As for technology, it was only present in the form of the material used during the inquiry-based activities. In other words, this model did not contemplate the explicit use of engineering design process or technology as guiding frameworks for helping students develop a solution to the problem.

Therefore, this model resembles what Bybee (2013) called "STEM as a reference for Science and Math perspective" (p. 75), where Science is the dominant discipline, Mathematics is the secondary focus, and Technology and Engineering are almost invisible and do not have an important function in the teaching unit. Therefore, the results of this pilot study revealed the need for further studies aimed at improving STEM approaches in terms of both conceptualization (i.e., STEM as being not only S&M) and practicality (i.e., STEM as a teaching method that in-service teachers consider viable and practical for teaching science education at elementary school).

3.2. Psychological underpinnings

To guide the process of designing STEM-based teaching, it is first necessary to know the process through which students acquire and develop meanings of new information, what cognitive processes take place, and how conceptual meanings are ultimately assimilated. Given the complexity of human cognition, this dissertation draws on multiple learning theories and builds on those principles that seem to be of value for developing a pedagogical framework for STEM-based science education.

Consequently, the underlying elements of constructivist and social-constructivist learning theories are analyzed, and most relevant aspects of the theories of Piaget (1964, 1974), Vygotsky (1979, 1981), Ausubel (1963, 1968, 1982) and Bruner (1961, 1966) are drawn for constituting the psychological underpinnings of the pedagogical framework for STEM-based science education proposed in this doctoral thesis.

3.2.1. Jean Piaget's theory of cognitive development

The theory of cognitive development proposed by Piaget (1964, 1974), also known as Genetic Epistemology, postulates that cognitive development is the result of a construction process through which the learner actively constructs and corrects his set of schemes as a result of biological maturation and the stimulation of environmental experiences. The individual understands and reacts to the stimuli presented in the world through different knowledge block, named schemas.

Piaget proposes a pattern of maturation that goes through different stages of increasing complexity, from an initial sensory-motor stage, in which the subject understands the world through senses and learn to respond through motor activity to the various stimuli presented to his senses, to a final fourth stage of formal operations in which the ability to think about abstract concepts is developed. Each stage of cognitive development represents a higher qualitative level of knowing or thinking. More complex and advanced logical forms of reasoning are gradually developed in a process that involves assimilation and accommodation. Piaget conceives the mind as a cognitive structure that tends to function in equilibrium; in the interaction with the world, the learner integrates new structures into existing coherent systems (i.e., assimilation) and the mind tends to restructure itself and reach new states of equilibrium (i.e., accommodation) through which the subject adjusts to the demands of the environment.

3.2.2. Lev Vygotsky's sociocultural theory of cognitive development

The sociocultural theory postulated by Vygotsky (1979, 1981) sought to explain the development of human beings in terms of social interaction. Vygotsky's theory posits that psychological functions are socially embedded, and that human mental cognition is a mediated process in which systems of symbols (e.g., language) are essential for the development. From his sociocultural perspective, learning and development are socially driven and takes place as learners interact with other peers and objects in collaborative environments. Consequently, learning is conceived as a mediated process, first at a social level and then at the individual level as a result of the interaction between the learner and other experienced members of the society.

This social and cultural interaction is mediated by the tools and systems of symbols that have been culturally constructed by a particular social group. These tools are subsequently transmitted through social interaction to other subjects who internalize them. In other words, the development of higher mental processes is subject to the transformation of social symbols, first between people (i.e., interpersonal) and later within the subject itself (i.e., intrapersonal) through a complicated and prolonged internalization process that necessarily requires the interaction between individuals. In this transformation, the individual internalizes the social constructs and shared a set of systems and symbols, giving rise to cognitive development. As the learner internalizes a greater number of set of systems, the cognitive development and the psychological operation that the individual can perform increase so as the range of activities and situations to which the individual can apply these psychological operations.

Some of the most prominent aspects introduced by Vygotsky is the notion of the zone of proximal development (ZPD), defined as the distance between the level of actual cognitive development of an individual and his or her level of potential development under guidance or in collaboration with more capable peers. The ZPD defines functions that are in the process of maturation with external help, and it represents the region in which cognitive development occurs (i.e., learning potential). As ZPD is highly related to an individual's cognitive development, it is dynamic and constantly changing as a function of the new cognitive development that is achieved.

3.2.3. David Ausubel's learning theory

The meaningful learning theory proposed by Ausubel (1963, 1968, 1982) is a learner-centered learning theory that focuses on and highlights what happens in the classroom when students are learning. Unlike Piaget's or Vygotsky's developmentally focused theories, Piaget aims at understanding and explaining the nature of learning, what conditions can facilitate or hinder it, and how to manipulate instructional elements and factors to enhance meaningful learning. Therefore, this theory is focused on the process of teaching and learning and place emphasis on the previous concepts and ideas that the learner already possesses.

As a result, Ausubel differentiated between meaningful and mechanical or rote learning. On the one hand, learning is significant when an individual is able to relate new information to the existing cognitive structure in a non-arbitrary and non-literal way. This interaction between the existing and the new knowledge leads to the transformation of the ideas already established in the cognitive structure. Therefore, the previous knowledge is modified based on the acquisition of new meanings, which give rise to new and more powerful and enriched ideas. Given that it is the learner who actively organizes and re-structures his own existing knowledge based on the new information, meaningful learning is conceived as the mechanism through which individuals gives meaning to new content, relating it to the ideas and concepts already available in owns cognition. On the other hand, when learning is not significant, it tends to be mechanical and resemble rote-based learning. In these cases, the new information does not interact with the existing cognitive structure, so the learning is produced in an arbitrary and mechanical way, without meaning for the learner.

Ausubel highlights three key aspects to enable meaningful learning. The first one is an emotional aspect and refers to the necessity for the student to have a favorable attitude towards learning and to be willing to learn, hence, this is closely related to attitudes towards school science investigated in this dissertation. The second condition is related to the logical significance of the new knowledge or information that must be presented in a logical way so it can be potentially related to the existing cognitive structure of the learner. Finally, the third condition that would facilitate the acquisition of meaningful learning is the presence of a stable cognitive structure with which the new information can be related and anchored.

3.2.4. Jerome Bruner's cognitive theory

Bruner (1960, 1961, 1966) introduced the concept of "Discovery learning", that conceived students as active learners who developed their own knowledge by organizing and categorizing information using a coding system. Contrary to Piaget, he postulated that children can be taught effectively regardless its stage of development and that development is a continuous process instead of a series of separates stages. Therefore, for Bruner, cognitive development can be fostered without having to wait for stage development readiness. As Vygotsky, he emphasized the child environment as a key element for the development, stating that other people should help the child through the process of scaffolding, like Vygotsky's ZPD. Thus, Bruner lessons must be designed to foster this discovery of the relationship between existing information. Doing so, education should aim for a "Spiral curriculum" (Bruner, 1988), where complex ideas are first connected to existing knowledge at a simplified level, and subsequently revisited in later school grades at a more complex and ever-increasing level of abstraction.

3.3. Characteristics of the framework

Several implications for the proposed framework arise from the theories introduced in the previous section (**Tables 1** and **2**). First, constructivist and social-constructivists theories consider learning as the construction of meanings. According to these perspectives, learning stems from interpretations based on individuals own experiences. In other words, instead of being simply acquired or transmitted, learners construct personal interpretations of the world through personal experiences. Therefore, individual's knowledge is always tentative and subject to change, and therefore, the pedagogical framework for STEM-based science education should focus on helping students organize and relate meaningfully the new information to their existing knowledge.

Table 1. The implication of Piaget's and Vygotsky's theories for the pedagogical framework proposed

Key principles	Implication for the STEM-based pedagogical framework
Piaget	
1. Biological stages of maturation	1.1. STEM key concepts and procedures should be carefully selected and adapted to learner's biological stage of cognitive maturation
2. Assimilation and accommodation of new knowledge	2.1. The STEM unit should start with a real-world problem aimed at creating disequilibrium in the existing cognitive structure 2.2. The STEM unit should allow students to develop knowledge meaningfully, and to apply this knowledge to new situations to facilitate accommodation. 2.3. Active teaching methodologies should be used to foster assimilation of the new knowledge through situations that encourage learning through discovery
Vygotsky	
1. Psychological development as a socially mediated process between advantage (i.e., MKO) and fewer advantage individuals	1.1. Cooperative working groups should be established with learners of different cognitive development 1.2. The teacher should act as a resource person (i.e., the metaphor of "teacher as a guide")
2. Socially and culturally system of symbols that influence the social interaction	2.1. Each phase of the STEM unit should have clearly established goals and milestones that scaffold students learning
3. Cognitive development as the internalization and transformation of socially and culturally established a system of symbols	3.1. The STEM unit should be based on authentic problems that creates opportunities for the system of symbols (i.e., STEM concepts and skills) to be internalized by the learners 3.2. Students should participate in rich and engaging activities that foster conceptual understanding that are transferable to other situations and contexts
4. Zone of proximal development (ZPD)	4.1. The teacher should be the mediator for the cognitive development by rooting designing the STEM unit according to student's prior knowledge and alternative conceptions 4.2. Practices should be carefully intertwined with scaffolding strategies and explicit instruction (when required) aimed at developing students' ZPD

Second, socio-constructivists authors argue that learning takes place by the interaction between the learner and environmental factors that can enhance or undermine the learning of the knowledge matter of study. Therefore, it is essential to establish realistic learning episodes where the learning task and concepts are delivered through situations that are relevant to learner's previous experiences. Consequently, in the proposed framework, learners should be engaged in real-world situations where the new knowledge is contextualized and used, and the instructional methods should assist learners in actively exploring and building their own understanding. There is a shift in the role of teachers

Table 2. The implication of Ausubel's and Bruner's theories for the pedagogical framework proposed

Key principles	Implication for the STEM-based pedagogical framework
Ausubel	
1. Meaningful learning is the process of linking new knowledge to the existing cognitive structure	1.1. New concepts and skills should be related to the existing cognitive structure of the learner through active teaching strategies and scaffolding
2. The learner has an active role in this process	2.1. Active teaching methodologies, that actively engage student in the learning process by exploring, discussing, and critiquing the new information, should be used 2.2. Students should be allowed to develop their own knowledge through questioning, exploring, elaboration and co-construction of a solution to the authentic problem underpinning the teaching unit.
3. Emotional disposition of the learner	3.1. Teaching strategies that are highly motivating for students should be used
4. Logical significance of the new knowledge	4.1. The new information and the materials should be adapted to learners' existing cognitive structures. The key concepts of each discipline should be strategically selected so that it can be delivered in an integrative manner
Bruner	
1. Discovery learning	1.1. The STEM unit should promote students active and collaborative participation 1.2. Teachers should connect the new concepts with existing knowledge in an authentic learning experience
2. Spiral curriculum	2.1. During the unit, STEM concepts and skills should be iteratively applied to different situations of increasing abstraction and complexity.

from teaching facts that are passively memorized by students, to designing experiences that foster learning through authentic and relevant contexts where students actively apply the new ideas and knowledge developed to diverse situations.

Third, since Piaget's theory conceives cognitive development based upon biological stages of maturation, the idea that teaching should be adapted to the cognitive development of the learner is fundamental. This notion relates to the need to introduce certain concepts at a particular maturation age and not before the learner reaches an appropriate level of cognitive development or maturation. Likewise, since developmentally appropriate concepts should be introduced to foster new cognitive assimilation and adaptation processes, the instructional approaches must provoke cognitive conflicts (i.e., disequilibrium), understood by Piaget as situations that encourage the construction of new processes of assimilation and accommodation of the new information. Accordingly, and also consistent with Bruner's theory, learning must be student-centered and should be acquired through processes based on discovery learning in which the teacher creates meaningful contexts and guides and facilitates this discovery through scaffolding.

Fourth, Vygotsky's sociocultural theory has several implications related to the social environment in which the learning takes place. On the one hand, given that higher mental processes are originated because of social interaction, students should learn in contexts where collaboration and discussion between peers are fostered and framed within activities that have a clear purpose and goals. Social interaction should not only be embedded in activities with concise goals but also be extended overtime for the internalization of shared systems and thus cognitive development to be achieved. Since this is a long and slow process, students should be able to test and reflect on their new cognitive development and internalized systems in real environments and situations by increasing progressively the level of abstraction and complexity of the concepts introduced, as Bruner also stated. Therefore, extensive time must be allowed for the process of internalization of shared social systems to take place.

On the other hand, Vygotsky's theory has important implications related to external support. As the principle of ZPD suggests, the instruction must be focused on the zone of potential development where individuals interact with a more capable person, also named most knowledgeable other (MKO), which provides the assistance needed to help the learner in the internalization process. This support, likewise promoted by Bruner as scaffolding, has direct educational implications for the creation of working groups, suggesting that they should be composed of learners with different levels of cognitive development so that less advanced learners may operate within their ZPD through the assistance provided by the most knowledgeable other.

Finally, Ausubel's theory has also fundamental educational implications for the pedagogical framework proposed in this dissertation. Consistent with the principle of meaningful learning, teacher instruction should focus on two main aspects. On the one hand, the existing cognitive structure of the learner (i.e., previous ideas and alternative conceptions) should be considered. In order to promote meaningful learning, teachers must examine what students already know about the concept under study or concepts related to the new information so that this existing knowledge can serve as support for the new information. If the existing structure is not ready for meaningfully learn the new material, it would be necessary to improve first the existing cognitive structure to anchor the new information in it. Therefore, the new content should be adapted to the existing cognitive structure of the learners, presenting introductory materials and using advanced organizers that serve as a cognitive link between what the student already knows and what he or she must learn, as advances by Bruner's theory as scaffolding.

On the other hand, attention should be devoted to how new information is delivered. Since meaningful learning is fostered by an active role of the learner, the content should be taught using active pedagogies that foster emotional disposition in students for learning the new content. Therefore, the pedagogical framework for STEM-based science education should allow students to develop understandings through actively engaging in questioning and exploration of the phenomenon under study. This aspect is closely related to the cognitive and affective dimensions of the tripartite model of attitudes (Khine, 2015; Simpson et al., 1994). When engaged in student-centered learning episodes, individuals experience a different way of learning science; therefore, this meaningful learning of the concepts under study could improve student's cognitive (e.g., "relevant" instead of "useless") and affective (e.g., "fun" instead of "boring") of the school science subject. In this regard, active teaching strategies are known for being highly motivating for students, especially for young learners in elementary education (see Demirel & Dağyar, 2016, and Schroeder, Scott, Tolson, Huang, & Lee, 2007 for two meta-analysis studies).

3.4. Didactic transposition

As introduced in Chapter 2, in this dissertation STEM is defined as an educational paradigm that intentionally integrates the concepts, procedures, and skills of science, technology, engineering, and mathematics disciplines through transdisciplinary real-world problems (**Figure 2**).

Consistent with constructivist and social-constructivist learning theories, and with the tripartite model of attitudes, the STEM teaching unit is articulated around an authentic real-world problematic phenomenon. This real-world problem stems from the science and mathematics curricula and is presented such that it is relevant to the students and appropriate for their biological stage of maturation. Likewise, this initial problem gives rise to different learning episodes that require two specific phases: knowledge development (phase 1) and knowledge application (phase 2). More specifically, on the one hand, the "knowledge development" phase is devoted to developing scientific knowledge and understandings of the phenomenon under study through inquiry investigations that require mathematical and technology knowledge. On the other hand, the "knowledge application" phase aims at the application of the acquired knowledge to the resolution of the real-world problem proposed, through the design of different solutions using the engineering design process, mathematics, and technology, thus accommodating the new knowledge developed. Both phases are delivered using student-centered teaching strategies that aim at developing scientific and engineering skills appropriate for student's maturational stage. Likewise, students' engagement in scientific and engineering investigations might trigger students' attitudes towards school science learning by improving their feelings and emotions (affective dimension of the tripartite model of attitude), and thoughts and beliefs (cognitive dimension) about school science. Eventually, students would develop positive attitudes towards school science subjects, which might promote their intention to enroll in further science subjects (behavioral dimension). These phases are discussed in depth in the following sections.

3.4.1. Knowledge development phase

The knowledge development phase consists of engaging students in inquiry investigation(s) to develop the scientific knowledge that would help understand the phenomenon under study and thus addressing the S (Science) of the STEM acronym.

The term inquiry has been prominently used in science education in relation to (i) scientists – the scientific inquiry as what scientists do to study natural world and develop scientific knowledge; (ii) students –the inquiry learning as a process by which students engage in the investigation of natural phenomenon to learn science concepts, develop scientific or inquiry skills, and acquire understandings of nature of science and scientific inquiry; and (iii) teachers –the inquiry teaching as the pedagogy to teach science (Crawford, 2014; Minner, Levy, & Century, 2010). Therefore, the proposed pedagogical framework for STEM science teaching conceives inquiry both as a mean that

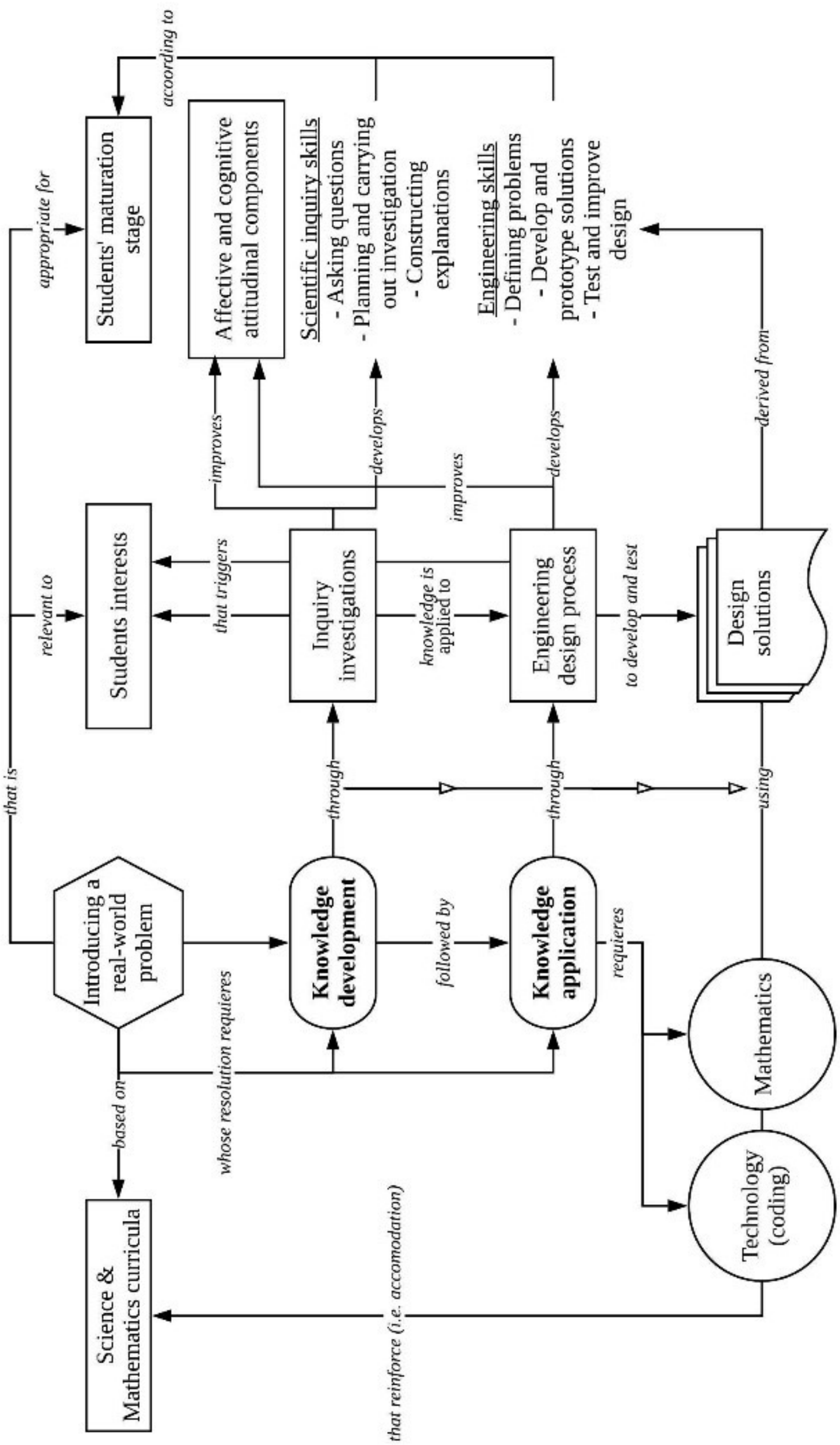


Figure 2. Pedagogical framework proposed for STEM-based science education

relates to inquiry as an instructional approach used for the teaching and learning of scientific content, and also as an end, where inquiry is an educational outcome to be studied and understood in the context of science education as a part of the science content itself (Abd-El-Khalick et al., 2004). Therefore, when engaged in inquiry-based science instruction students are expected to learn scientific concepts through the use of scientific-like practices in authentic contexts, improve understanding of Nature of Science (Lederman & Lederman, 2014; Lederman, 1992; Schwartz, Lederman, & Crawford, 2004) and Scientific Inquiry (Abd-El-Khalick et al., 2004; Flick & Lederman, 2004; Lederman, 2009), and develop scientific-related skills such as identifying problems, formulating research question and hypothesis, designing and conducting of experiments, analyzing data, and drawing conclusions (Keselman, 2003; Pedaste & Sarapuu, 2006)[4]. More specifically to this dissertation, students do an inquiry to mimic scientific work and develop scientific skills (i.e., inquiry as an end) and they engage in inquiry investigations to also learn the scientific core concepts (i.e., inquiry as a mean). These views have been promoted in international (AAAS, 1993; NGSS Lead States, 2013) and national (LOMCE, 2013) educational reforms, is nowadays considered as central for the achievement of scientific literacy.

Consistent with the ZPD and scaffolding principles promoted by social-constructivist theories, inquiry investigations are divided into smaller units that are logically connected to help the teacher structure the unit and guide students through different phases. These smaller units form a cycle, known as the inquiry cycle (Pedaste et al., 2015). In the science literature, the inquiry cycle adopts many different forms and frameworks, such as the learning cycle originally proposed by Karlplus & Thier (1967) and subsequently adapted and popularized by Bybee et al. (2006) as the 5E instructional model, or Wecker, Kohnle, & Fischer's (2007) three phases cycle. In a review study, Pedaste et al. (2015) identified up to 109 different terms referring to inquiry phases, with core features that can be grouped into five general inquiry steps, mainly orientation, conceptualization, investigation, conclusion, and discussion.

Likewise, based on how much information and guidance is provided to students, inquiry investigation can be categorized in a four-level continuum of inquiry, from confirmatory, where students confirm a phenomenon following specific instruction with the research question and procedure already provided by the teacher and where the potential results are known in advance, to an open-type inquiry where students follow self-directed investigations by formulating research questions, design procedures and generate explanation based on the evidence collected (Herron, 1971). Similarly, Martin-Hansen (2002) concluded that inquiry investigations can be structured (cookbook type lessons), open (fully student-centered that closely mirror real scientific practices), guided (where students are assisted by the professor through the inquiry cycle), or coupled, which combines guided and open-type inquiries into a 5 step sequential cycle: (i) invitation to inquiry, (ii) teacher-oriented guided inquiry, (iii) student-oriented open inquiry, (iv) inquiry resolution, and (v)

[4] It should be noted that the development of Nature of Science knowledge (Lederman, 2007), as well as understandings about Scientific Inquiry (Lederman et al., 2014) is beyond the scope of this doctoral dissertation and, therefore, was not addressed in the intervention unit.

assessment.

In short, consistent with constructivist, socio-constructivist, and the tripartite model of attitudes, inquiry investigations are used to foster an active role of the students and allow them to engage in authentic problem-solving investigations to learn scientific concepts and scientific skills. More specifically, in this dissertation, the coupled inquiry model is adopted (Martin-Hansen, 2002), consisting of conducting several investigations about the phenomena under study following the Pedaste et al.'s (2015) 5-phase inquiry cycle, and progressively decreasing teacher guidance to allow students to participate more autonomously[5].

3.4.2. Knowledge application phase

During the second phase of the pedagogical framework for STEM-based science education, students apply the scientific knowledge acquired during the knowledge development phase to the design, construction, and testing of solutions that may solve the initial real-world problem proposed at the start of the unit, thus addressing the E (Engineering) of the STEM acronym. In this dissertation, the E represents the use of engineering practices that foster the learning of scientific curricular content through the application of the acquired knowledge to the design and construction of a solution.

The inclusion of engineering practices into science education curricula allows students to discover the interconnection between science and engineering (NGSS Lead State, 2013). In science educational policy documents, engineering is conceptualized as a STEM content integrator (NRC, 2012), and the focus is on student's engagement in what is considered to be eight essential practices for understanding science and engineering (**Table 3**). While there is an evident overlap between science and engineering practices, the NRC (2012) address the differences as forms that are mutually complemented and strengthened, and therefore, should be made explicit during instruction.

From a pedagogical point of view, engineering practices are addressed using the engineering design process (EDP) that structures the inclusion of engineering in elementary classroom through five steps: identify the problem, reflect on the goals, establish criteria for the design, build a prototype, and test and improve the prototype until arriving to the best solution (Cunningham, 2014, 2018). The EDP is an iterative cycle consisting of different phases named ask (identification of the problems and constraints), imagine (ideas brainstorming and selection of the best one), plan (designing and selecting needed materials), create (creation and testing of a prototype), and improve (analysis of the strength and limitations of the prototype and ways of improvement). Therefore, unlike the inquiry cycle that is used for advancing theoretical knowledge and understandings of the phenomenon under study, the engineering design process engages students in the design, construction, and testing of different prototypes of an artifact that may solve the problem formulated

---

[5] The real-world problem introduced in the intervention unit designed for this dissertation required students to engage in two inquiry investigations to develop the pertinent knowledge for proposing a solution to the initial problem during the knowledge application phase. However, it should be noted that other problems used for a STEM unit may require more than two inquiry investigations.

at the start of the unit.

3.4.3. Technology and mathematics

Considering elementary school student's biological stage of maturation and because of the Spanish curricula that do not explicitly stress technology as a separate subject or discipline, in the pedagogical framework proposed in this dissertation technology is conceptualized through an instrumental-engineering lens (c.f., Feenberg, 2017). This implies that technology is portrayed as both making and using artifacts and instruments that can enhance science discoveries and engineering applications, a conceptualization that is in line with curricula reforms (e.g., NGSS Lead States, 2013). Therefore, technology is represented by the devices used during the scientific inquiry cycle for the design, building, and testing of prototypes using the engineering design process, and for the use of mathematical calculations and representations developed during both the knowledge-development and knowledge-application phases.

Likewise, consistent with the Spanish curricula for elementary education, and the scientific and engineering practices mentioned in **Table 3**, this dissertation advocates for the inclusion of computational thinking and computer programming of devices in the knowledge development and application phases. Computational thinking is defined as the ability to engage in problem-solving analysis and resolution by drawing on computer science concepts (Wing, 2006, 2008). In other words, CT is the ability to transfer and apply computational programming processes to new contexts that are not necessarily related to programming (Berland & Wilensky, 2015) in order to "(…) solve

Table 3. The distinction between science and engineering practices

Science	Engineering
Ask questions about phenomena	Define problems that need solutions
Use models to develop explanations about natural phenomena	Use models to test possible solutions to a new problem
Scientific investigation seeks to test, revise or develop new theories and explanations	Engineering investigation seek to identify parameters that determine the effectiveness, efficiency, and durability of their designs
Produce data to find patterns through the derivation of meaning using tabulation, graphical interpretation, and statistical analysis	Produce data to compare which solution meets specific design criteria that best solves the problem within specific constraints
Use mathematics and computational thinking to represent variables and study their relationships	Use mathematics and computational thinking to test and assess potential solutions
The goal is to explain world features through the construction of theories that can be applied to a specific situation of phenomenon	The goal is to develop design solutions to a specific problem based upon scientific knowledge and within particular criteria and needs
Reasoning and arguments are integral for identifying the best possible explanation to natural phenomenon among a range of lines of reasoning	Reasoning and arguments are integral for identifying the best possible solution to a problem among a range of competing ideas

Extracted and adapted from NRC (2012)

problems effectively and efficiently" (Shute, Sun, & Asbell-Clarke, 2017, p. 171), which requires to master up to six main skills known as decomposition -splitting a problem into common elements that are more easily addressed-, abstraction -retrieval of fundamental aspects of a system-, algorithm -design logical and ordered instructions-, debugging -identification and fixing of errors-, iteration -refinement of solutions-, and generalization -transfer of skills to other situations- (Shute et al., 2017).

CT is gradually being considered as the 21st-century skill and has been found to be beneficial in the development of problem-solving skills (Chalmers, 2018) and in the fostering of science, technology and mathematics understandings (Barak & Zadok, 2009; Nugent, Barker, Grandgenett, & Adamchuk, 2010).

In the proposed pedagogical framework for STEM-based science education, technology and computational thinking are fostered through the introduction of coding-related activities aimed at programming devices that replace conventional laboratory-type technology/material (e.g., programming a thermometer instead of using a conventional one). This conceptualization is consistent with the NGSS Lead States (2013) that included the use of mathematics and computational thinking as one of the eight key science and engineering practices that are considered as essential for all students to learn. Therefore, "Students are expected to use laboratory tools connected to computers for observing, measuring recording and processing data" (p. 58) and also to develop computational thinking skills such as searching and organizing data, create algorithms, and use simulation of natural phenomenon.

As for the M (Mathematics) of the STEM acronym, this dissertation includes it through mathematics curricula content and through mathematical skills. In the proposed pedagogical framework, mathematics is learned in a real-world context where they became meaningful and relevant. Thus, mathematics is essential throughout the STEM unit for the collection and interpretation of data during the scientific inquiry, for the use of the engineering design process in which mathematical reasoning is essential during the prototype design phase, and for the use of technology, in this framework being represented by computational programming, where mathematics is used to inform the user about the programming of different apparatus that can enhance scientific discoveries and its engineering applications. This conceptualization is in line with the National Research Council (2012), that considers mathematics as a tool for the understanding of science and engineering practices: "(…) Mathematics often brings these two fields together by enabling engineers to apply the mathematical form of scientific theories and by enabling scientists to use powerful information technologies designed by engineers" (p. 65). Therefore, "Students are expected to use mathematics to represent physical variables and their relationships and to make quantitative predictions." (NGSS Lead States, 2013, p. 58).

3.4.4. Approach to integration

The report of the STEM Task Force Report (2014) highlighted the need for cohesive and active approaches to teaching and learning, linked to STEM disciplines. Therefore, in the proposed pedagogical framework, curricular integration is conceived as the learning of STEM disciplines through contexts in which complex phenomena and situations require the use and application of the knowledge and skills of these four disciplines (NAE & NRC, 2014). Specifically, this dissertation defines integration as an operational amalgamation of conceptual, procedural, and attitudinal curricular contents of the different school STEM subjects and disciplines. Thus, it is advocated for a complementary overlap between these contents through interdisciplinary approaches that are effectively articulated through multiple learning episodes tackling grade-appropriate, real-world problems in which skills and concepts of each STEM discipline are tightly linked and explicitly emphasized as they interconnect and become interdependent.

Hence, a reduction of the boundaries and limits between each discipline is advocated, which would improve the connection between the concepts, procedures, and skills of each discipline. The taxonomy proposed by Gresnigt et al. (2014) is adopted, which proposed a six-level hierarchy to integration from fragmented to connected, nested, multidisciplinary, interdisciplinary, and transdisciplinary approaches, depending on the degree of integration sought between the different disciplines. Therefore, consistent with the nature of curriculum integration, there is a certain loss of the discipline perspective, and major real-issues such as climate change or water pollution became the central emphasis of the educational experience. However, the use of explicit-reflexive teaching strategies is likewise advocated to reflect on the characteristics of each STEM discipline and on the similarities and differences between each discipline.

While this conceptualization highlights the characteristics that a perfectly integrated-STEM approach should have, the framework proposed in this dissertation is intended to be dynamic in order to adapt to the limitation of the educational system. Thus, it is not advocated that all science units should be delivered following an integrative approach. Rather, the need for lecture-based learning episodes is recognized, and therefore integration is promoted only for the teaching and learning of curricular contents that can be approached through real problems, appropriate to student's cognitive development, and in which both the uniqueness and the interconnection of each STEM discipline can be made explicit and understood by students.

In summary, the pedagogical framework for STEM-based science education advanced in this dissertation proposes scientific and engineering practices as the backbone of a STEM unit. Using a two-phase procedure, students engage first in scientific practices to develop scientific knowledge and understandings of the phenomenon under study (Phase 1: knowledge development). Then, students get involved in engineering practices to apply this knowledge to the design and development of a technological solution for the phenomenon under study (Phase 2: knowledge application), thus explicitly tackling the interconnection between the STEM disciplines and discovering the relevance of each one. Finally, technology is explicitly addressed through the inclusion of computer programming, which, in turn, would support mathematics integration as fundamental for the coding process and for data collection, analysis, and interpretation during both learning phases.

3.5. The intervention unit

This section presents a summary of the lesson plan designed and implemented in this dissertation (Toma, 2018). The full lesson plan has been published as a book chapter and is therefore attached in Appendix 2.

At the beginning of this unit, students are confronted with a problematic situation related to the colonization of a planet in which the existence of water has been discovered, that addresses curricular units from the Science and Mathematics subjects (**Table 4**). Adapted from the science fiction film "The Martian" (Scott et al., 2015), this problematic situation introduces the story of an astronaut who finds himself stranded and alone on the Mars planet after the crew of this manned mission were forced to abort the space expedition due to a strong dust storm. In order to survive until the rescue day, the astronaut must find how to determine if the water found on this planet is suitable for human

Table 4. Curricular content addressed in the intervention unit

Science and Mathematics curricular contents
Science - Initiation to scientific activity. Experimental approach to some questions related to Science. - Project planning and reporting of the results - Usefulness of some products and materials innovations for the advancement of society - Matter: properties, states, and changes. Study and classification of some materials by their properties: hardness, solubility, state of aggregation, texture, color, shape, plasticity, and conductivity - Prediction of alterations in the movement and shape of bodies due to the effect of state changes - Types of mixtures. Separation of components of a mixture by filtration, evaporation, and magnetization. Physical procedures for separating mixtures - Physical-chemical parameters of water quality - Science knowledge that improves health (purification and potabilization of water, etc.)
Technology - Use of information and communication technologies (ICT) to search for and select information, simulate processes and present conclusions - Guided search of information on the internet. Time control and responsible use of ICT - Benefits and risks of technology and products - Appropriate use of ICT as a learning resource - Integration of ICT in the learning process to obtain information, make calculations, solve problems and present results - Computational thinking: algorithm, abstraction, decomposition, debugging, iteration, and generalization*
Engineering - Design and construction of simple structures that fulfill a function or condition to solve a problem from modulated parts.
Mathematics - Mathematical strategies and procedures: Graphs and tables - Evaluation of the results obtained - Measurements using conventional instruments and units of measurement in everyday contexts - Critical analysis of the information presented - Realization of simple statistical studies putting into practice the phases: Obtaining and recording data, presentation in tables, transformation into graphs and evaluation of the results

* Based on Shute, Sun & Asbell-Clarke (2017).

consumption. Therefore, this problematic situation requires the development of two inquiry investigations to determine the physical-chemical parameters of the quality of the water found on this planet (first inquiry) and the mixture separation techniques appropriate to purify the water (second inquiry), as well as the application of the engineering design process for the design and construction of an effective water filter to convert the water found in this planet suitable for human consumption.

During the knowledge development phase of the unit, students engage in a coupled inquiry (Martin-Hansen, 2002) to develop the knowledge that is necessary to understand the problematic situation. This phase consists of two inquiry investigations. In the first one, students analyze the temperature, pH, turbidity, suspended solids, and levels of chlorine and nitrate in different water samples with the help and guidance of the teacher, thus introducing the Science discipline from the STEM acronym. Likewise, during this phase, students use mathematics to interpret the results obtained and to computationally program or code few microcontroller boards using visual programming language to convert it into different data measurement sensors, thus introducing the mathematics and technology component of the STEM approach. Using different external measurement sensors, this microcontroller board can be programmed to become a measurement tool of different parameters. For example, in this unit, instead of using a traditional existing thermometer to measure the temperature of the water, students code the microcontroller board to convert it into a thermometer. This inquiry investigation concludes by reflecting on the physicochemical parameters that determine the quality of different water samples and reflect about the importance of this knowledge for everyday life situations, for the environmental problems of potable water shortages in many underdeveloped countries, and on the usefulness of science and scientific knowledge for society. The teacher also explicitly emphasizes the interrelationship between science, technology, and mathematics for the development of new knowledge.

During the second inquiry investigation, students learn about different methods of mixture separation techniques to improve the quality of the contaminated water. During this second investigation, the teacher's level of guidance decreases considerably, and students get involved in a more active way. Thus, students propose and test different methods of water purification using household and conventional laboratory material (i.e., metal grids, coffee filters, Bunsen burner, etc.). In this way, students carry out research related to the techniques of separation of mixtures. As during the first investigation, students program the microcontroller board to convert it into a motion sensor, accelerometer, and a stopwatch to determine the efficacy and the time needed to filter the solid components out of the water using each separation technique. Finally, this second stage may also end with explicit-reflections on the integration of science, technology, and mathematics for the development of new knowledge, and on the usefulness of this knowledge for real life[6].

[6] It should be noted that, although beyond the scope of this dissertation, aspects related to the understanding of scientific inquiry could also be addressed (Lederman et al., 2014). For example, students may discover that there is no single method or sequence of steps established to develop new knowledge, that procedures are guided by the question asked, and that explanations are developed from the collection of new data and from what was already known about the phenomenon studied.

After the knowledge development phase of the intervention unit, students are expected to (i) understand the parameters that determines the water quality and be able to follow the correct procedure to measures these parameters in different water samples, (ii) know the different separation techniques that can be used to separate solid components from the water, and the correct procedure to do so, and (iii) code the microcontroller board to convert it into the different measurement tools needed for these tasks.

During the second phase of the intervention unit, named knowledge application, students use the engineering design process to apply the scientific knowledge acquired during the knowledge development phase to the design and construction of a solution that helps solving the initial problem, thus introducing the engineering discipline from the STEM acronym. More specifically, students design and test different technological processes and tools (i.e., water filters) and analyze their effectiveness in terms of economic and material cost, filtering time, and efficiency in improving water quality. Following several iterative design and test phases, students improve their design until reaching the optimal solution that helps to solve the initial problem. In this phase, it is expected that students would apply the scientific, mathematics and computational programming knowledge and skills developed during the knowledge development phase. At the end of the second phase of the STEM unit, the teacher and students explicitly reflect on the connection between science, engineering, mathematics, and technology for solving problems of social relevance. For example, students reflect that engineering design is made according to some specification, constraints, and goals, that there is an intrinsic relationship between engineering and science (i.e., the utility of scientific knowledge to engineering), and that there are also differences between science and technology (e.g., while engineering aims at developing technological solutions, science scope is the development of knowledge).

Chapter 4. Methods

This chapter provides information on the methods of the main study conducted in the doctoral thesis. Section 4.1. introduces the methodological approach adopted, with reference to the qualitative and quantitative strand of the study. Next, sections 4.2. and 4.3. describes the context in which this study has been conducted and the sample, respectively. Subsequently, section 4.4. presents data collection methods, with subsection 4.4.1. focusing on the qualitative strand, and the 4.4.2. subsection on the results of two systematic reviews on attitude quantitative measurement instruments and on the development and validation of the instruments used in this dissertation. Finally, section 4.5. explains how data was analyzed.

4.1. Mixed-methods exploratory sequential design

For the main study of this dissertation, a mixed-method exploratory sequential design was adopted. This design consist of two sequential phases, being the first one qualitative and the second one quantitative (Creswell & Clark, 2011. More specifically, the exploratory design being with prioritizing qualitative data collection and analysis in the first phase or strand. Then, building on the qualitative data, the researcher conducts a quantitative phase that builds on, extends, or apply the features of the first qualitative strand. According to Creswell & Clark (2018), "This feature may be the generation of new variables, the design of an instrument, the development of activities for an intervention (…)" (p. 67). Consist with the characteristics of this mixed methods design, the main study of this dissertation starts with collection and analysis of qualitative data that is used to translate the qualitative findings into an approach (i.e., intervention unit) that can be tested quantitatively to determine its effectiveness for improving elementary school students' expectancies of success and attitudes towards school science. More specifically, the study begins with a series of iterative implementation of the intervention unit that are qualitatively evaluated in terms of validity (i.e., the extent to which the unit and its implementation can be considered STEM) and practicality (i.e., whether in-service elementary teachers are able to implement it according to the pedagogical framework of reference). Each new iteration builds on a previous prototype of the intervention unit until a reasonable approach and product are achieved. Then, the quantitative phase starts, consisting of a quantitative intervention to examine the impact of the last unit prototype in the dependent variables under study. **Figure 3** includes a diagram with the main features of each strand, and the next section describes in-depth the characteristics of the qualitative and quantitative strand, separately.

Strand	Procedure	Evaluation	Product
Qualitative data collection	- Focus groups - Non-participant structured classroom observation	RQ 1.1. What teaching strategies should be used to implement valid and plausible STEM-based teaching units for science education in elementary grades?	- Transcripts
Qualitative data analysis	Direct and conventional content analysis		- Challenges themes - Teaching strategies
Quantitative data collection	- SUCCESS instrument - S-SSAS instrument	RQ 1.2. Does STEM-based teaching units promote students' expectancies of success and attitudes towards school science?	- Numeric data
Quantitative data analysis	Univariate and multivariate statistics		- Expectancies of success and attitudes
Overall interpretation	Discussion in the light of existing literature and findings	Primary research question	Empirical evidence on plausibility and effectiveness of STEM

Figure 3. Diagram for the exploratory sequential design adopted in the main intervention study

4.1.1. Qualitative strand

4.1.1.1. Educational design research

For the development of the qualitative strand, educational design research methodology was adopted. This methodology can be defined as an innovative research approach in educational practice that involves the systematic design, development, and evaluation of educational interventions aimed at addressing complex problems (Plomp & Nieveen, 2007). Educational design research is an alternative research approach to traditional research methods that aims at designing and developing practical and effective educational interventions (e.g., pedagogical approaches, programs, materials) based on collaborative and reflexive labor between researchers and practitioners (i.e., teachers).

This approach emerged out of the need to address the credibility gap resulted from traditional research approaches (e.g., survey or correlational type of studies) that hardly provide any solution to the problems and issues of everyday practice beyond the mere identification and description of the problem itself (Design-Based Research Collective, 2003; Reeves, 2006; Van den Akker, Fravemeijer, McKenney, & Nieveen, 2006). Put differently, as opposed to traditional research paradigms that examine the educational problem in isolation and within artificial environments (e.g., “laboratory” type of contexts), thus producing an incomplete and partial understanding of the problem being studied (Barab & Squire, 2004), educational design research is developed in the naturalistic contexts where the educational problem has been identified in order to develop (i) empirically rooted innovative interventions and (ii) generalizable principles, also known as design principles (i.e., "how-to" guidelines), that can be generalized to broader and more diverse contexts (Reeves, 2006). Therefore, educational design research is carried out iteratively in close collaboration with researchers and practitioners to increase the likelihood that the intervention will be practical, effective, and scalable to similar contexts, therefore being more suitable for educational contexts were major and complex problems where specific guidelines for solutions are not yet available. This is clearly the case with STEM education.

Although this paradigm resembles action research, the main difference between both paradigms lies in the fact that the latter is appropriate for professionals who seek to improve their educational practice based on research, rather than generating design principles (Denscombe, 2007). Following Plomp & Nieveen (2007) conceptualization, educational design research is:

> “(…) the systematic study of designing, developing and evaluating educational interventions (such as programs, teaching-learning strategies and materials, products and systems) as solutions for complex problems in educational practice, which also aims at advancing our knowledge about the characteristics of these interventions and the processes of designing and developing them.” (p. 13).

This definition highlights the cyclical nature of educational design research, consisting of successive prototypes where analysis, design, evaluation and revision and re-design of the educational intervention are iterated until reaching the desired outcome or design-principle that advance solutions to the problem. Therefore, interventions or design principles not resulting in the desired outcomes are refined and further tested in the naturalistic context until the design principles appear to be effective. Each re-designed intervention is known as prototypes (Plomp & Nieveen, 2007), and from each new prototype, new design principles contributing to closing the gap between the educational practice and the desired outcome are advanced.

4.1.1.2. Design research phases

Educational design research has three main phases, named: (i) preliminary research, (ii) prototyping, and (iii) assessment (Plomp & Nieveen, 2007). First, during preliminary research, needs, context analysis, and literature review reviews are performed to develop a conceptual framework underpinning the first prototype. During this phase, the problem in context is identified and specifically formulated through practitioners and expert consulting, analysis of similar problems or interventions, focused literature review and/or preliminary studies. The prototyping phase involves the design, development, implementation, reflection, and re-design of a sequence of prototypes aimed at improving the educational intervention. Therefore, during this phase, several tentative interventions and design principles are advanced and refined through iterative cycles until a desired prototype is achieved. Finally, once the intervention is effective and the design principles are established, the assessment phase is performed. The main aim of this phase is to evaluate whether the intervention is practical (i.e., if practitioners can implement the educational innovation and the design principles).

4.1.1.3. Quality criteria

The four quality tenets of educational design research proposed by McKenney, Nieveen, & Van den Akker (2006) are met in this dissertation as follows:

❖ Rigor. This tenet refers to the rigorous standards for scientific research (Shavelson & Towne, 2002), including, but not limited to, (i) pose of significant research questions, (ii) research linked to relevant theory, (iii) adequate use of research methods and instruments that allow investigation of the research question, and (iv) transparent disclosure of research to allow professional scrutiny and critique. Consequently, (i) the significance of the research questions addressed in this dissertation are justified by the growing literature demanding the development of conceptual and pedagogical frameworks to STEM Education (Martín-Páez et al., 2019); (ii) the pedagogical framework for STEM-based education is rooted in constructivism and social-constructivism theories (see Chapter 3 and Appendix 2), whose relevance to understanding the teaching and learning process has been established for decades; (iii) both quantitative and qualitative research methods have been used in an attempt to respond the research question proposed, and the measurement instruments used where comprehensively developed and psychometrically validated (see Appendices 5, 6, and 7); and (iv) in-depth details and original published studies are collected in this dissertation to allow transparent reporting of the studies performed.

❖ Relevance. The second quality tenet raises questions about whether the research questions are relevant for educational and policy practice, and whether the research group have the working knowledge needed for the direct investigation of the research questions in natural settings. As stated before, the relevance of this dissertation is underlined by national and international literature that demands an understanding of the teaching and learning processes that must take place when STEM approaches are adopted. In this regard, this dissertation proposed both a pedagogical framework for STEM education, which could inform science teaching practices at the elementary educational level and also stimulate research in this line of investigation. As for the needed background to conduct this research, the author of this dissertation has specific background in elementary education, which is the target educational stage of the different studies described in this dissertation, and the research group, or more specifically the advisor of this dissertation, has an extensive experience of more than two decades in the use of constructivist-based didactic strategies, such as inquiry teaching and learning approach, this being a central teaching strategies using in the proposed framework.

❖ Collaboration. The third quality tenet refers to the need for collaboration between the research team and practitioners. In this regard, the design, development, implementation, and revision of each prototype of the proposed STEM framework was jointly conducted in direct collaboration with six elementary school teachers that participated in the design, implementation, and refining of the intervention unit, and with indirect collaboration of more than 36 visiting teachers from the schools that participated during the formative evaluation of the unit.

❖ Validity. Finally, in order to develop effective design principles for STEM education, the pedagogical framework proposed (and, therefore, the teaching unit designed) should meet the following validity criteria for high-quality interventions (Nieveen, 1999): (i) content validity: the framework is based on state-of-the-art knowledge; (ii) construct validity: the STEM disciplines are linked to each other in a meaningful way and the knowledge development and knowledge application phases of the unit are logically designed; (iii) Practicality: the framework aims at being realistically usable in natural settings by the end-user or practitioners (i.e., elementary school teachers); and (iv) effectiveness: the use of the pedagogical framework proposed should result in the desired outcomes, i.e., development of elementary school students' expectancies of success and attitudes towards school science (this aspect is addressed through the quantitative strand).

4.1.2. Quantitative strand

For the development of the quantitative strand, a one-group pretest-posttest design was adopted (Campbell & Stanley, 1963). This type of research design is often used to determine the effect of an intervention on a particular aspect of a given sample. Its main features are (i) use of a single group of participants, and (ii) a sequential arrangements of data collection (Allen, 2017). The first characteristics involves that all participants from the target sample are part of a single condition (i.e., the intervention group) receiving the same treatment. The second characteristics denote that the dependent variable is assessed before and after the intervention is implemented, following a O1 X O2 design, in which O represents the data collection and X the implementation of the intervention. Within this design, the effect of an intervention is determined by exploring the difference between the first evaluation (i.e., O1: pretest) of the dependent variable and the second one (i.e., O2: posttest).

4.2. Context and temporalization

For this dissertation, a collaboration agreement has been established with the Centro Rural de Innovación Educativa de Burgos (CRIEB, Burgos Rural Center for Educational Innovation), approved by the Conserjería de Educación de Burgos (Department of Education of Burgos). The CRIEB is a state-funded educational center, located on the peripheries of the city of Burgos, intended to provide complementary educational instruction for students from rural areas of the Burgos province. This means that the target population of this dissertation includes all the elementary schools from the Burgos province with up to 12 units/classes. Therefore, over an academic year, up to one thousand students enrolled in 3^{rd} to 6^{th} grades of elementary education attend the CRIEB school center for 5 days (from Monday to Friday) on an educational boarding school regime.

The collaboration agreement consisted of using the CRIEB school center during the 2017-2018 academic year for testing the pedagogical framework for STEM-based science education developed in this dissertation. To this end, a total of 21 full weeks of intervention were established. Thus, 894 elementary school students coming from up to 36 state-funded schools located in the province of Burgos (**Figure 4**) attended the CRIEB school center and participated in this Ph.D. thesis. Of them, 274 were enrolled in 3^{rd} or 4^{th} grades, and the remaining 620 students were 5^{th} or 6^{th} graders. In September 2017, schools were randomly assigned to one of the 21 weeks so that each intervention week consisted of no more than 50 students.

Each intervention week, a new group of students attended the CRIEB school center from Monday morning to Friday noon. During the week, students participated in lessons developed following the pedagogical framework proposed in this dissertation and in other relaxation activities (e.g., going to the municipal swimming pool, to the cinema, etc.). The intervention lessons included 12 teaching hours, implemented by the six elementary school teachers who worked full time at the CRIEB school center during the 2017-18 academic year. This amount of time corresponds to the number of classes that are usually dedicated to each unit of school science in conventional schools. Students were grouped according to the school grade they were enrolled in, thus forming a group for 3rd and 4th graders, and another group for 5th and 6th graders.

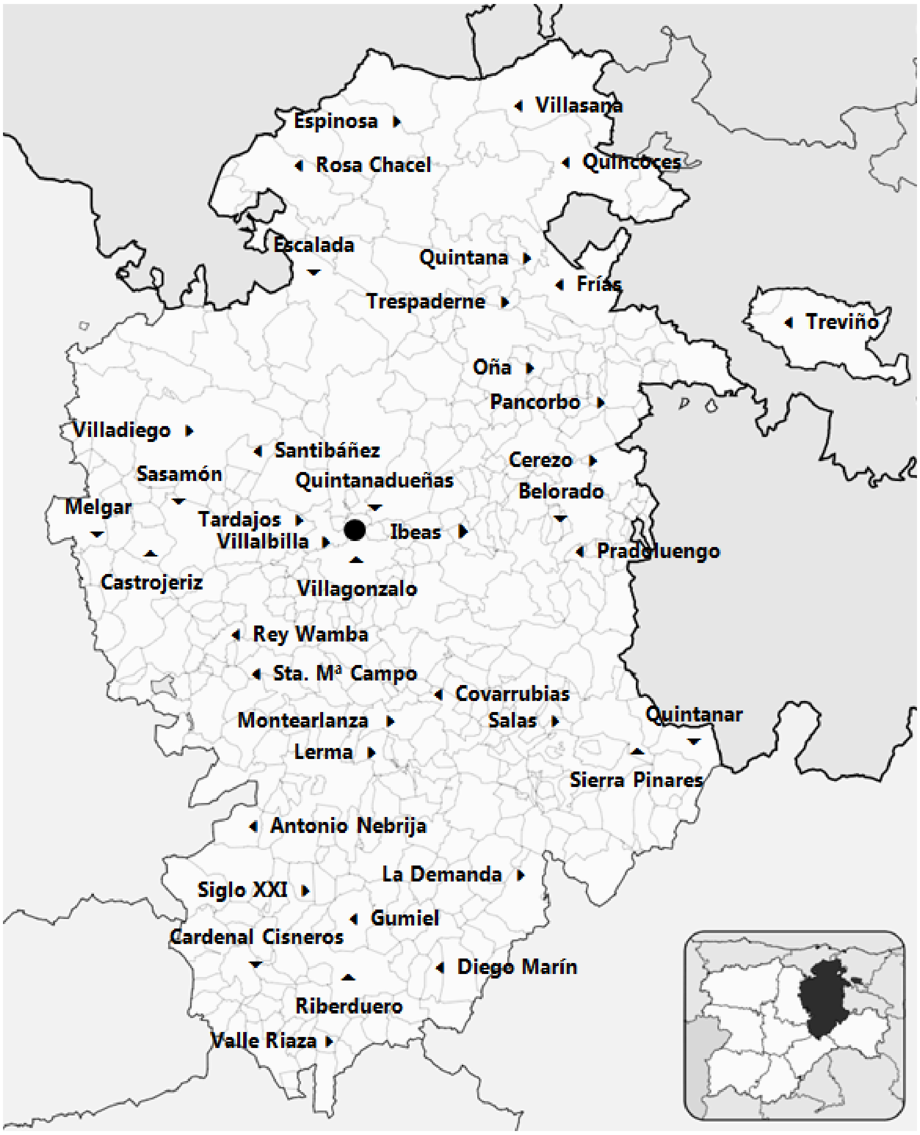


Figure 4. Cities and towns from which the educational school centers that participated in this dissertation were drawn. Note: ● represents the city of Burgos.

Of the 21 intervention weeks, the first 15 weeks were devoted to improving the intervention unit following the educational design research (Plomp & Nieveen, 2007) approach described in previous sections (i.e., qualitative strand). During these 15 weeks, adjustments were made until suitable design principles were developed, thus improving the validity and practicality of the STEM-based intervention unit. In this way, the CRIEB school center became the ideal testing site that matches the characteristics of the educational design research approach. First of all, the circumstance of having 21 weeks in which new students were attending the intervention site allowed several refining of the intervention unit framework during the qualitative strand (see Chapter 5.1.), in order to overcome the difficulties encountered after each trial. Secondly, unlike conventional schools, the CRIEB school does not have a pre-established schedule. This allowed establishing an educational schedule that simulated that of a conventional school, with the particularity that all the Science lessons usually taught in approximately 5-6 weeks have been condensed into one week. Third, the availability of six elementary school teachers has allowed the pedagogical framework to be developed in a naturalistic school-type context, thus reducing the bias introduced by the researcher and endowing to the investigation a more realistic and practical approach. Finally, the participation of students from such diverse schools and contexts allows for greater generalization of the results, at least as far as rural schools are concerned.

Once the design principles were developed and the practicality and validity of the intervention unit were established, the effectiveness of the pedagogical framework for STEM-based science education was assessed (i.e., quantitative strand). Thus, the last 6 weeks of the 2017-18 academic school year were devoted to examining whether the intervention unit improves elementary school students' expectancies of success and attitudes towards school science (**Figure 5**).

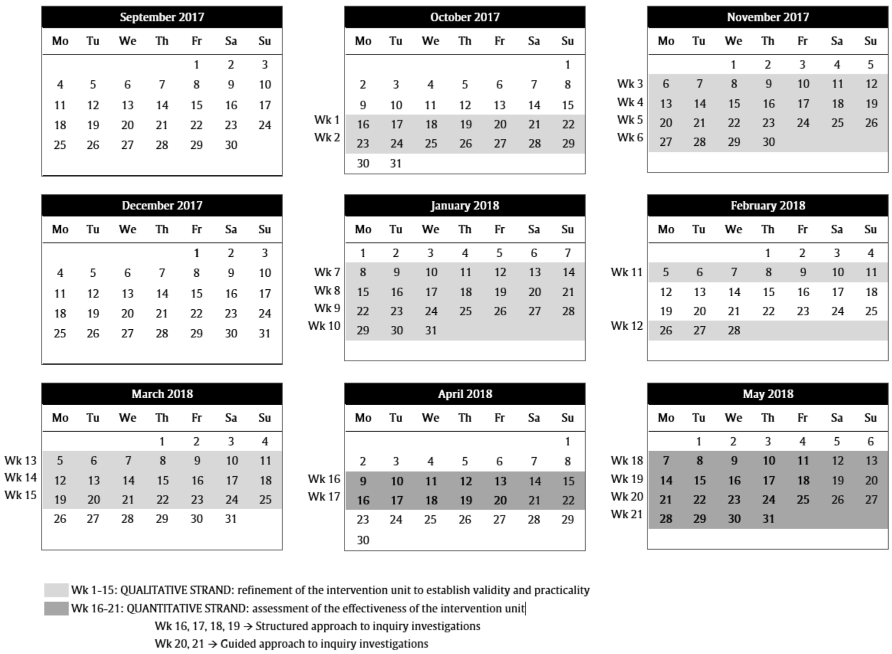


Figure 5. Distribution of the intervention weeks at the CRIEB school center

4.3. Sample

4.3.1. Qualitative strand

A convenience sampling has been established by means of the teachers who have worked in the CRIEB school center during the implementation of this doctoral thesis. The sample consisted of six elementary school teachers (2 women), in charge of teaching the different sessions of the didactic unit described in Chapter 3. Although all six teachers had extensive teaching experience in elementary education (approx. 10 years each), none of them were familiar neither with the STEM movement nor with the inquiry and engineering design process teaching strategies.

Therefore, the month before the start of the intervention (i.e., September 2017), the teachers participated in a short, nine total hours, professional development workshop. The workshop consisted of a first part dedicated to a brief theoretical introduction to STEM education as well as to the teaching strategies used within the pedagogical framework, and a second part focused on explaining the lesson plan. The teachers then received the materials and classroom notebooks that students would use during the teaching unit.

It should be noted that the lack of a more comprehensive and long-term training professional development program was a deliberated decision made by the author of this dissertation, in order to increase the generalization of the results to other educational contexts. Given that Spanish teachers tend to avoid active methodologies for science teaching, such as inquiry-based learning (Cañal, 2007; Romero-Ariza, Quesada, Abril, Sorensen, & Oliver, 2019), the aim was to develop a pedagogical framework that would allow teachers not experienced in inquiry-based or engineering design teaching strategies, to be able to implement integrated STEM for science education.

4.3.2. Quantitative strand

The sample for the quantitative strand was established by means of convenience sampling techniques, based on the students who have attended the CRIEB school center during the last six weeks of the 2017-18 academic school year (from April 9th to May 31st, 2018).

A total of 245 elementary school students participated in the evaluation of the effectiveness of the pedagogical framework proposed in this dissertation. Of them, 67 were enrolled in 3rd (n = 20) and 4th (n = 47) elementary grades, and 178 were 5th (n = 82) and 6th (n = 96) graders. Half of the students were girls (n = 123) and the mean age of the students was 10.15 (SD = 1.22) for the girls and 10.19 (SD = 1.10) for the boys.

For data analysis, students were grouped in lower (3rd and 4th graders) and upper (5th and 6th graders) grade groups, consistent with the classroom distribution followed each week. In addition, upper grades students were randomly assigned to either the structured or the guided approach to inquiry investigations. After an initial inspection of the collected data, five and six cases were deleted from the expectancy of success and attitudes databases, respectively, for univariate outliers. Additionally, using Mahalanobis distance, three more cases were deleted from the S-SSAS database for multivariate outliers (**Tables 5** and **6**).

Table 5. Total sample size and valid cases for lower grades students after removing outliers

Gender	All sample	Expectancy of success	Attitudes
Girls	36	36	35
Boys	31	29	31
Total	67	65	66

Table 6. Total sample size and valid cases for upper grades students after removing outliers

	All sample		Expectancy of success		Attitudes	
Gender	Structured	Guided	Structured	Guided	Structured	Guided
Girls	48	39	47	38	45	36
Boys	52	39	51	39	50	39
Total	100	78	98	77	95	75

4.4. Data collection

4.4.1. Qualitative strand

Qualitative methods were used to investigate the difficulties experienced by the teachers during the implementation of the intervention unit. Data were collected from two main sources. The first one was focus groups (Morgan, 1996) with the teachers reflecting on the implementation challenges they experienced. Three focus groups were conducted during the 15-week period of the prototyping phase. In addition, the second source was non-participant, structured observations and field notes undertaken by the author of this dissertation (Bogdan & Biklen, 2007; Mulhall, 2003). The author of this dissertation attended all lessons held in the educational center CRIEB during the 21 weeks that lasted the intervention (from October 2017 to May 2018). More specifically, during the 15 weeks corresponding to the prototyping phase, non-participant observations were performed in 209 one-hour sessions.

4.4.1.1. Focus groups

A focus group is a discussion group in which several individuals are interviewed at the same time in order to obtain information and generate reactions on the topic under study (Beck, Bryman, & Futing, 2004). In the case of this doctoral thesis, the objective was to understand the limitations that teachers have had in implementing the didactic unit based on the STEM model advanced in this dissertation, and to jointly propose solutions to overcome these barriers and, therefore, to make this type of pedagogical approach more viable. Three focus groups of approximately one hour each have been carried out. To this end, teachers were instructed to write down the difficulties they have experienced in each session they have implemented so that they could discuss them in-depth during the focus groups. During each focus group session, each teacher-related his or her experience, the difficulties encountered, and the proposed improvements. At the end of the session, a consensus was reached on the measures to be taken in order to improve the implementation of the didactic unit.

Table 7. Codes used for the non-participant structured observation

Aspects	Codes
Teachers propositional knowledge	The teacher had a solid grasp of the lesson
	The lesson is presented in a logical and clear fashion
	Connections with other disciplines were explored and valued
Students procedural knowledge	Students made predictions or hypotheses
	Students devised means for testing their predictions or hypotheses
	Students engage in thought-provoking discussions
Classroom culture	The teacher asked questions that triggered students thinking
	Active participation was encouraged
	The teachers acted as a resource person, instead of delivering lecture-type of instructions

4.4.1.2. Non-participant structured observation

The author of this dissertation performed direct, structured, non-participant observation of each session. The aim was to gain a better understanding of the educational environment created using the pedagogical framework for STEM-education advanced in this dissertation. To this end, the researcher took field notes on different aspects of the implementation process. More specifically, the rubric developed by the author and colleagues (Toma, Greca & Meneses Villagrá, 2017), based on Pitburn et al. (2000) Reformed Teaching Observation Protocol, was used to identify: (i) teachers propositional knowledge (what they know and how they organize and present the material in a learner-oriented manner), (ii) students procedures knowledge (to what extent teachers allowed students to engage mentally and experimentally during the session, and (iii) the classroom culture (to what extent there is communicative interactions between students). The specific codes used during the non-participant observation are described in **Table 7**.

4.4.2. Quantitative strand

Data was collected using paper and pencil instruments upon students' arrival at the CRIEB school center (i.e., pretest on Mondays) and the last day of each intervention week (i.e., posttest on Fridays). To avoid socially desirable responses, the author of this dissertation collected all the data in the absence of classroom teachers from the CRIEB and visiting schools, and students were informed that participation was voluntary, anonymous and that responses would not affect their school grades.

Two exhaustive reviews have been carried out to select the most appropriate instruments for the quantitative strand. To this end, the first systematic review examined attitude towards science instruments published after Blalock et al.'s, (2008) study (Toma, in press). Given the major conceptualization issues identified in the first review, a second systematic review, hereafter referred to as "beyond attitudes review" (Toma, Lederman, & Meneses Villagrá, under review), was conducted to analyze the psychometric properties of instruments measuring other valuable psychological constructs that were conceived as attitudes (e.g., self-efficacy, motivation, emotions, engagement, etc.) in the instruments analyzed in the first attitude study. In the next section, an extended abstract of both review studies is presented. The full-text accepted article for the first review is attached in Appendix 3, and the full-text of the second review, which is currently under peer review, is attached in Appendix 4.

4.4.2.1. Systematic reviews

Both studies were conducted following a systematic procedure consistent with the PRISMA statement (Liberati et al., 2009). Potentially relevant articles were identified in the Web of Science (WOS) Core Collection and Science Direct databases, in the case of the attitude review, and in the WOS and Scopus databases in the case of beyond attitude review. Likewise, a snowball technique, which consists of examining the reference list of selected articles for relevant studies not identified through the databases, was used in both review studies. A total of 14 studies met the inclusion criteria of the attitude review and 61 studies for the beyond attitude review.

The findings of these systematic reviews revealed that Blalock et al.'s (2008) and Munby's (1983) criticisms are still valid for current instruments used in science education research. The results demonstrate that although a rather large body of instruments was developed and published in the science education research in the last 14 years, the conceptual and methodological quality of most of them are below modern standards for educational and psychological testing. Most instruments lacked theoretical foundations and were almost based on existing instruments with conceptually poor definition of the attitude construct. In addition, many of the attitude assessments were limited in terms of validity and reliability of psychometric evidence. Likewise, several limitations related to data reporting and misuse of some psychometric tests were also identified. These results are generalizable to the Spanish literature, as reported in the instrument validation study reported in the next section (Toma & Meneses Villagrá, under review).

These results calls into question the confidence that can be placed in the results derived from studies using instruments whose validity and reliability is at stake and calls for the adoption of rigorous psychometric analysis procedures to develop valid and reliable measurement instruments that would help to support (or refute) the assumptions and consensus reached in this line of research.

The implications of these findings for the main intervention study were twofold. First, it raised the need to develop valid and reliable measurement instruments to be used in the assessment of the effectiveness of the proposed pedagogical framework for STEM-based science education. Second, given the lack of confidence that can be placed in existing attitudes towards science studies using psychometrically poor instruments, especially in the Spanish context, a descriptive study of the target population of this dissertation (i.e., students from the Burgos province) should be carried out to determine their attitudes using more robust instruments, which will be described in Chapter 5.

4.4.2.2. SUCCESS instrument

Given the lack of consensus in the science education literature about what is measured when attitudes towards science are examined[7], the development and validation of a first instrument were performed. In this section, an extended abstract of two validation studies about the first quantitative

[7] See Appendices 3 and 4

instrument used in this dissertation, called SUCCESS, is presented (Toma & Meneses Villagrá, under review; Toma, 2019). The first, full-text article, which is currently under peer review, is attached in Appendix 5. The second full-text published validation article is attached in Appendix 6.

The SUCCESS is a theoretically driven (based on EVT theory), short (6 items), unidimensional (measures expectancies of success in school science), and Likert-type (five response options) instrument rooted in social psychology theories of achievement motivation. More specifically, the SUCCESS was developed based on Eccles et al.'s (1983) EVT theory, described in Chapter 2 of this dissertation. The use of this theoretical framework was based on past and recent research indicating that student's self-efficacy and expectancies of success in science and mathematics are low (Sellami et al., 2017; Meece, Wigfield, & Eccles, 1990; Wigfield, 2004), factors that may negatively affect the development of their intentions to enroll in future science-related studies. Given that students' self-efficacy influences the type of activities in which students judge themselves to be competent (Bandura, Barbaranelli, Caprara, & Pastorelli, 2001), and given that students with a high self-concept are more likely to persist in science and engineering careers (Mau, 2003), the development of expectancies of success in school science from the elementary stages of the education system may be beneficial for counteracting the loss of interest in science at later stages. Despite existing literature calling for the evaluation of expectancies of success from an early age, the instruments revised in the systematic reviews introduced in the above section do not tackle this aspect. Therefore, there is a need to address this gap in the literature and to develop a valid and reliable instrument to assess whether the pedagogical framework designed in this dissertation is effective in fostering students' expectancies of success in school science.

The development and validation of the SUCCESS instrument followed robust guiding principles based on Classical Test Theory (CTT, DeVellis, 2017; Nunnally & Bernstein, 1994) and the Standards for Educational and Psychological Testing (AERA, APA, & NCME, 2014). In the first study (Toma & Meneses Villagrá, under review[8]), the expectancy of success construct was defined and specified by drawing on theories of achievement motivation. Next, based on the theoretical underpinnings adopted, a deductive approach based on literature review was used for the development and adaptation from existing literature an initial list of potentially relevant items. In the second phase, one panel of experts composed of university professors and elementary education teachers examined the content validity of the proposed items, leading to a refined pool of items. Next, the readability of the items was examined through cognitive interviews with the target population, which led to the final pool of items. Finally, in the third and last phase, the retained items were administered to a large-scale sample, consisting of Spanish elementary school students, and the psychometric properties in terms of item quality, construct validity based on Exploratory Factor Analysis (EFA), convergent validity and internal consistency reliability were analyzed.

In the second study (Toma, 2019[9]), the psychometric properties of the SUCCESS instrument were further examined. More specifically, the construct validity was examined using Confirmatory

---

[8] See appendix 5

[9] See appendix 6

Factor Analysis (CFA); criterion validity was examined using the Pearson correlation coefficient between the SUCCESS and the attitudinal S-SSAS instrument reported in the next section; finally, the reliability of the SUCCESS instrument was examined through the Cronbach α coefficient.

Taken together, the results of both studies suggest that SUCCESS is a valid and reliable instrument with great content, construct and criterion validity, and acceptable internal consistency reliability. Based on these findings, it can be concluded that reveal that the proposed instrument is conceptually consistent with the "Expectancies of success" construct first introduced by Eccles et al. (1983) and methodologically robust in terms of modern validity and reliability psychometric evidences. Therefore, it can be used with confidence in this dissertation.

4.4.2.3. S-SSAS instrument

In addition to developing and validating the new instrument called SUCCESS, it was decided to conduct a cross-cultural adaptation and psychometric validation of an existing attitude toward school science instrument (Toma & Meneses Villagrá, 2019a). In this section, an extended abstract of the second quantitative instrument validated and used in this dissertation is presented. The full-text published validation article is attached in Appendix 7.

From the list of reviewed instruments in the systematic reviews reported in previous sections, the 10-items School-Science Attitude Survey (SSAS, Kennedy et al., 2016) was chosen as an adequate candidate to be adapted and validated for Spanish speaking students. This instrument was selected because (i) it was originally developed and validated following a robust procedure, (ii) it measures the main attitude constructs used in science education research, and (iii) is short and easy to administer, therefore suitable for the CRIEB context; due to its brevity, it allowed to administer others measurement instruments, such as the SUCCESS, something inconceivable if a more extensive instrument had been chosen due to student administration fatigue.

The School-Science Attitude Survey is an instrument designed to examine student's attitudinal profile to the area of school science through ten items that address the six common attitudinal constructs used in the literature of attitudes toward science. The SSAS was translated into Spanish following a cross-cultural translation procedure. Content and face validity were confirmed before large-scale administration using a panel of experts and cognitive interviews with the target population.

Psychometric evaluation of the Spanish School Science Attitude Survey (S-SSAS) indicated an adequate level of internal consistency reliability and that response was well distributed along with the response categories, showing great sensitivity and no evidence of extreme response tendency. Intraclass correlation coefficient supported its temporal stability reliability, with a 10-days span between the first and the second administration. Pearson correlation coefficient reported acceptable predictive validity with a strong correlation between expected attitudinal constructs based on results reported in specialized literature of attitudes. Likewise, the S-SSAS showed great concurrent validity, thus being highly correlated with two attitudes measures of conceptual convergence already validated in the literature. In addition, discriminative validity between S-SSAS constructs was confirmed by obtaining similar results to those of its original version in terms of gender and rural or urban school variables. Taken together, it can be concluded that the S-SSAS represents a first effort to provide a valid and reliable measure to examine attitudes toward school science of Spanish-speaking elementary students, especially when there are time constraints for data collection. Therefore, it can also be used in this dissertation with confidence.

4.5. Data analysis

4.5.1. Qualitative strand

Focus groups were audio-recorded and transcribed to allow for in-depth analysis of the challenges reported and improvement measures described, and non-participant structured observation was recorded in a field note. Data were analyzed using both direct and conventional content analysis (Hsieh & Shannon, 2005). First, direct, theory-driven content analysis was performed. Common difficulties reported in the literature related to using active-based (e.g., inquiry) strategies served as initial coding categories to identify the extent to which CRIEB teachers have experienced the same challenges. Next, a conventional coding analysis was used to identify further challenges mentioned by the CRIEB teachers or observed by the author of this dissertation. Thus, transcripts and field notes were read repeatedly; key words and sentences capturing difficulties experienced, and solutions proposed, were highlighted. Using an iterative process, sentences describing similar ideas were grouped and codes that are more general were created; finally, similar codes were congregated to create meaningful clusters (Patton, 2002).

4.5.2. Quantitative strand

Data was analyzed using univariate and multivariate statistics, which pertains to the family of general linear model (GLM) and is used to detect group differences when there are several dependent variables that are conceptually related (Field, 2009; Leech, Barret, & Morgan, 2005), like in the case of attitudes towards science construct. Although multivariate analysis is an extension of the analysis of variance (ANOVA), most researchers just wrongly conduct a series of separate ANOVAs for each dependent variable (i.e., each attitude construct). In the Spanish context, most studies analyzed attitudinal instrument results without considering its factorial structure. Therefore, authors (e.g., Vázquez & Manassero, 2008; Marbá-Tallada & Márquez, 2010; Pérez-Franco, de Pro Bueno & Pérez-Manzano 2018) analyzed each item independently, making up to a total of 25-50 individual comparisons without using any α value correction. Unfortunately, this leads to a common problem in the literature of attitudes towards science, related to type I error inflation, derived mainly from erroneous statistical practices such as multiple t-tests or ANOVAs without adjusting the level of significance (α value).

For example, following the common criterion of $\alpha = 0.05$, in a study that examines the hypothesis that there are differences between boys and girls in a single dependent variable, the probability of obtaining a false positive (i.e., Type I error) is only 5% $[1 - (1 - \alpha)]$; therefore, the researcher may conclude, with a 95% confidence (1 - 0.05), that the difference observed between two groups is not due to chance. However, since attitudes towards science are a multidimensional construct (Gardner, 1975), measured by numerous different dependent variables representing the affective, cognitive and

Figure 6. Simulation of false-positive inflation in multiple tests without adjustment of the α level, based on the formula $1 - (1 - \alpha)^N$, where N is the number of t-tests performed on the same sample.

behavioral dimensions of attitudes (Eagly & Chaiken, 1995; Simpson et al., 1994; Tytler & Osborne, 2012), studies in this line of research often use multiple t-tests or ANOVAs, one for each attitude dimension, without adjusting the common criteria of $\alpha = 0.05$. In these cases, by not adjusting the level of α to the number of tests to be performed, the probability of false-positive increases exponentially with the number of tests performed (**Figure 6**). Thus, with two repeated t-tests or ANOVAs without α adjustment, there will be a 9.75% probability of type I error (i.e., obtaining a significant result when, in reality, it is not significant). With ten tests with uncorrected α, the probability of a false positive increases to 40%, and so on. Therefore, the likelihood that the results derived from the Spanish studies mentioned above are only a reflection of type I errors rather than significant results is very high. To avoid this problem, a more conservative α value should be used, by establishing some type of correction (i.e., Bonferroni correction), or ideally, by using multivariate analysis when data meet the several assumptions of multivariate statistics. In the following subsections, these assumptions are discussed for both SUCCESS and S-SSAS instruments.

4.5.2.1. Expectancies of success in school science

For lower grades students (3rd and 4th graders), data was analyzed using a 2 (gender: girls/boys) x 2 (time: pretest/posttest) mixed between-within subjects' analysis of variance (ANOVA). Preliminary checks were conducted to examine assumptions for mixed-between-within subjects ANOVA: Levene's test showed that the variances for expectancy of success scores were equal, both for pretest $F\,(1, 63) = .937$, $p = .336$ and posttest scores, $F\,(1, 63) = .1.420$, $p = .238$, thus the assumption of homogeneity of variance was not violated. Likewise, the Box's test was not significant, $M = 6.892$, $p = .084$, indicating that the assumption of homogeneity of covariance across the groups was not violated. Consequently, Wilks Lambda results will be reported (Tabachnick & Fidell, 2007).

Likewise, lower grades data was further analyzed using one-way between-group analysis of covariance (ANCOVA) to determine whether the gender results hold when controlling for pretest scores. In order words, ANCOVA was used to determine if girls and boys responded differently to the intervention unit when controlling for their pretest scores. The independent variable was gender (girls/boys). The dependent variable was the posttest scores on the SUCCESS instrument (expectancy of success in school science), administered following completion of the intervention unit. Pretest scores on the SUCCESS instrument, administered prior to the intervention, were used as a covariate to control for individual differences. Preliminary checks were conducted to examine assumptions for ANCOVA: The distribution of pretest and posttest scores for each gender revealed linearity between the dependent variable (posttest score) and the covariate (pretest score), thus, the assumption of linear relationship was not violated. Levene's test showed that the variances for expectancy of success scores were equal, F (1, 63) = .229, p = .634, thus the assumption of homogeneity of variance was not violated. Finally, there was no interaction between the independent variable (gender) and the covariate (pretest scores), F (3, 61) = .878, p = .076, thus the assumption of homogeneity of regression slopes was not violated, and therefore ANCOVA results can be interpreted with confidence (Tabachnick & Fidell, 2007).

For upper grades students (5^{th} and 6^{th} graders), data was analyzed through a 2 (teaching approach: structured/guided inquiry) x 2 (gender: girls/boys) x 2 (time: pretest/posttest) mixed between-within subjects' analysis of variance (ANOVA). Preliminary checks were conducted to examine assumptions for mixed-between-within subjects ANOVA: Levene's test showed that the variances for expectancy of success scores were equal, both for pretest F (3, 171) = 1.818, p = .146 and posttest scores, F (3, 171) = .560, p = .642, thus the assumption of homogeneity of variance was not violated. Likewise, the Box's test was not significant, M = 14.390, p = .624, indicating that the assumption of homogeneity of covariance across the groups was not violated. Consequently, Wilks Lambda results are reported (Tabachnick & Fidell, 2007).

Moreover, upper grades data was further analyzed using a 2 x 2 between-groups analysis of covariance (ANCOVA) to determine whether the gender results hold when controlling for pretest scores. In other words, ANCOVA was used to examine if girls and boys responded differently to the structured or guided approach to inquiry. The independent variables were the teaching approach (structured/guided inquiry) and gender (girls/boys). The dependent variable was the posttest scores on the SUCCESS instrument, administered following completion of the intervention unit. Pretest scores on the SUCCESS instrument, administered prior to the intervention, were used as a covariate to control for individual differences. Preliminary checks were conducted to examine assumptions for ANCOVA: The distribution of pretest and posttest scores for each approach revealed linearity between the dependent variable (posttest score) and the covariate (pretest score), thus, the assumption of linear relationship was not violated. Levene's test showed that the variances for expectancy of success scores were equal, F (3, 171) = 1.122, p = .342, thus the assumption of homogeneity of variance was not violated. Finally, homogeneity of regression slopes was the only assumption violated, F (3, 171) = 6.153, p = .014, which may increase to likelihood of false-negative results (i.e., type II error); therefore, upper-grade results when controlling for pretest scores should be interpreted within this limitation (Tabachnick & Fidell, 2007).

4.5.2.2. Attitudes towards school science

Lower grades data were analyzed using a 2 (gender: girls/boys) x 2 (time: pretest/posttest) multivariate analysis of variance (MANOVA) with repeated measures. Preliminary checks were conducted to examine assumptions for MANOVA: Levene's test showed that the homogeneity of variance was met for all pretest and posttest (p's = .101 to .830). Likewise, the Box's test was not significant, M = 110.999, p = .187, indicating that the assumption of homogeneity of covariance across the groups was not violated; consequently, Wilks Lambda results will be reported. Matrix scatterplot inspection for each attitude factor revealed that the assumption of linear relationship was not violated. Finally, Pearson correlation revealed that the maximum strength of the correlation among the attitude variables was r = .562, between intention and enjoyableness factors; since these values are lower than .8, the assumption of multicollinearity and singularity was met, and therefore MANOVA results can be interpreted with confidence (Tabachnick & Fidell, 2007).

Likewise, lower grades attitudinal data were further analyzed using a one-way between-groups multivariate analysis of covariance (MANCOVA) to control for pretest scores. The independent variable was gender (girls/boys). The dependent variable was the posttest scores on the S-SSAS instrument (attitude towards school science dimensions), administered following completion of the intervention unit. Pretest scores on the S-SSAS instrument, administered prior to the intervention, were used as a covariate to control for individual differences. Preliminary checks were conducted on the attitudinal data to examine assumptions for MANCOVA: The distribution of pretest and posttest scores for each approach revealed linearity between the dependent variable (posttest score) and the covariate (pretest score), thus, the assumption of a linear relationship was not violated. Levene's test showed that the variances for each attitudinal factor were equal, thus the assumption of homogeneity of variance was violated (p's = .356 to .620). Consequently, Wilks' Lambda results will be reported. The Box's test was not significant, M = 30.628, p = .153, indicating that the assumption of homogeneity of covariance across the groups was not violated. Finally, homogeneity of regression slopes was violated for factors measuring intention, F (2, 63) = 7.844, p = .001, and enjoyableness, F (2, 63) = 7.031, p = .002, which may increase to likelihood of false-negative results (i.e., type II error); therefore, MANCOVA results should be interpreted within this limitation (Tabachnick & Fidell, 2007).

Upper grades attitudinal data were analyzed using a 2 (teaching approach: structured/guided inquiry) x 2 (gender: girls/boys) x 2 (time: pretest/posttest) multivariate analysis of variance (MANOVA) with repeated measures. Preliminary checks were conducted to examine assumptions for MANOVA: Levene's test showed that the homogeneity of variance was met for all but the posttest scores for the intention, F (3, 166) = 5.181, p = .002 and perceived difficulty, F (3, 166) = 2.824, p = .040 attitudes factors. Likewise, the Box's test was significant, M = 326.184, p = .015, indicating that the assumption of homogeneity of covariance across the groups was violated; consequently, Pillai's Trace results will be reported. Matrix scatterplot inspection for each attitude factor and pedagogical approach revealed that the assumption of linear relationship was not violated. Finally, the Pearson correlation revealed that the maximum strength of the correlation among the attitude variables was r = .616, between usefulness and relevance factors; since these values are lower than .8, the assumption of multicollinearity and singularity was met (Tabachnick & Fidell, 2007).

Finally, a 2 x 2 between-groups multivariate analysis of covariance (MANCOVA) was conducted on the upper-grade attitudinal data to determine if the structured and the guided approach to inquiry had different impact on girls and boys, after controlling for their pretest scores. The independent variables were the pedagogical approach (structured vs. guided) and gender (girls and boys). The dependent variable was the posttest scores on the S-SSAS instrument, administered following completion of the intervention. Pretest scores on the S-SSAS instrument, administered prior to the intervention, were used as a covariate to control for individual differences. Preliminary checks were conducted on the attitudinal data to examine assumptions for MANCOVA. The distribution of pretest and posttest scores for each approach revealed linearity between the dependent variable (i.e., posttest score) and the covariate (i.e., pretest score), thus the assumption of linear relationship was not violated. Likewise, Levene's test showed that the variances for perceived difficulty, F (3, 166) = 3.144, p = .027, and self-efficacy, F (3, 166) = 2.698, p = .048, were not equal, thus the assumption of homogeneity of variance was violated. Consequently, Pillai's Trace results will be reported. The Box's test was not significant, M = 86.059, p = .069, indicating that the assumption of homogeneity of covariance across the groups was not violated. Finally, homogeneity of regression slopes was also violated, which may increase to likelihood of false-negative results (i.e., type II error); therefore, MANCOVA results may not accurately reflect the differential gender effect of the intervention STEM-based unit, if any.

Chapter 5. Findings

This chapter presents the findings of the main intervention study of this dissertation. More specifically, section 5.1. reports on the iterative implementation, evaluation, and re-design of the intervention unit developed following the pedagogical framework for STEM-based science education introduced in Chapter 3. Therefore, this section describes the changes that have been made to improve the practicality and viability of such an educational approach in the Spanish elementary school science education. Next, section 5.2. reports on the effectiveness of the intervention unit (in its latest version after making all relevant improvements) on fostering students' expectancies of success and attitudes towards school science. These results are reported by controlling for school grade (lower grades: 3rd and 4th graders; upper grades: 5th and 6th graders), gender (boys and girls), pretest scores, and for upper graders, by comparing between delivering the knowledge development phase of the intervention unit using structured vs. guided inquiry investigations.

5.1. Qualitative strand

5.1.1. First prototype

The first focus group was held after two implementation weeks (from October 16th to October 27th, 2017); thus, each teacher had the opportunity to implement twice each session that they were in charge of. **Table 8** reports the main codes derived from the first focus groups and from the non-participant observation of 30 total sessions, and the educational measures adopted in prototype 2 to overcome these challenges. It should be noted that during the first three prototypes, the STEM unit described in Chapter 3 and attached in Appendix, lasted 15 total sessions instead of 12 sessions.

Table 8. Challenges identified in the first prototype and educational measures

Codes representing challenges	Educational measures
- Classroom management - Messiness (King et al., 2001) - Technical problems related to computational coding	- Classroom management strategies and restructuring of the classroom space. - Use a laptop instead of iPads from coding the microcontroller board and using a micro USB cables instead of Bluetooth

The codes with a reference were derived using direct content analysis, and the ones without any reference using conventional content analysis.

Three main codes emerged in the first prototype, mainly related with logistical and organizational difficulties. Thus, the first great difficulty mentioned by all teachers was that of class management. Far from the conventional image of students behaving well in quiet and orderly classrooms, all teachers from the CRIEB school center have faced a classroom climate to which they were not accustomed. Hence, during the session in which the inquiry investigations and the engineering design process was conducted, the active participation and the interactions between students considerable increased the noise level, that reached much higher levels than in a traditional, teacher-centered session. This made teachers feel that they were not having good classroom management skills and that therefore; such an educational paradigm would not be suitable for science education. In short, teachers argued that, unlike when they deliver lecture-style lessons that promoted stillness and limited interaction among students, the intervention unit triggered too much noise and aggravated the misbehavior of students.

From the field notes taken during classroom observations, it was clear that the classroom often reached very disordered levels, with teachers having great difficulty in being able to focus students' attention on the indications they were giving, especially during the sessions in which the inquiry investigations were conducted. These aspects added to the messiness generated by the hands-on investigations, and the fact that the level of student engagement and noise certainly reached limits that would not be allowed in a conventional school, seemed to have overwhelmed the teachers. Therefore, it would seem that classroom management problems were closely related to the teachers' lack of techniques to manage the control of the classroom when implementing active teaching strategies, which made the implementation of the intervention unit very disorganized.

On the other hand, some technical problems were reported in relation to the computational coding of the microcontroller board. More specifically, the microcontroller boards that were programmed as measurement sensors required the use of iPads tablets and a visual-coding application that required internet access through Wi-Fi connection. The first problem reported by the teachers was related to the poor Wi-Fi connection, which often made it difficult for some workgroups to code the microcontroller board. Lack of Wi-Fi connection, iPads that we're stuck with freeze screen errors or that did not received access to the Wi-Fi were very frequent problems experienced during the programming sessions. The second problem was related to the Bluetooth connection; as the coding program, once created in the iPads, needed to be transferred to the physical microcontroller board via Bluetooth connection, students experienced multiple connectivity issues. Since there were multiple iPads and microcontroller boards that were used simultaneously in a classroom, it was often impossible to connect each tablet to the desired microcontroller board to transfer the program.

To overcome these challenges, several changes have been made to the first prototype of the intervention unit. First, the distribution of the class has been changed completely. Thus, working groups have been created based on cooperative learning principles (Johnson & Johnson, 1999). These working groups were generally made up of four or five students and were maintained in all lessons. Also, a shelf has been created in which all the material was organized, according to the sessions to be implemented, with only one student from each working group being authorized to take the material to the working tables and to return it back to the shelf at the end of each lesson, thus reducing class disorder and noise. In this way, it was intended to avoid the hustle and bustle

Table 9. Challenges identified in the second prototype and educational measures

Codes representing challenges	Educational measures
- Lack of pedagogical knowledge (Toma & Greca, 2018) - Lack of logical structure of the didactic sequences (Aragüés, Quílez, & de la Gándara, 2014)	- Structuration of each lesson according to the inquiry-cycle and engineering design process phases.
- Difficulties of group work and inconsistent contribution and engagement between the member of each working group (Anderson, 2002)	- Introduction of specific group roles

The codes with a reference were derived using direct content analysis

generated during the sessions in which the inquiry investigations and the engineering design process was carried out.

Second, some strategies shown in the literature to be useful in facilitating the implementation of inquiry-based science teaching (i.e., Keeley, 2008; Qablan & DeBaz, 2015) were implemented in each session. Techniques such as "Raise hand for silence" and a modified version of the "traffic light cups" (i.e., one student in each group was in charge of communicating to the teacher their difficulties or questions about the lesson being implemented) were used to grab students' attention, especially during practical work sessions, and also to make the interaction between teacher and students more organized and effective. By introducing these classroom management techniques, it was intended to help the teacher attend, in a controlled and well-ordered manner, all the problems that working groups face during the knowledge development and knowledge application phases of the unit.

Finally, regarding the microcontroller boards, it was decided to use USB micro-cables instead of the Bluetooth to transfer the program to the microcontroller board, and to use computer laptops with cable internet when Wi-Fi internet connection failed. It should be noted that these logistical difficulties related to the coding-phase of the unit persisted throughout the intervention, albeit in a less severe way as in the first prototype. Since these were aspects that teachers could not control (e.g., technical assistance, faster Wi-Fi internet connection, up-to-date computer software), it was impossible to reach a definite solution.

5.1.2. Second prototype

The second prototype, which included the changes mentioned above, was implemented for 4 weeks (from November 6th to December 1st, 2017). After this time, and a total of classroom observations of 60 lessons, the second focus group was held. Field notes revealed, and teachers reported, that the changes made to the first prototype allowed them to better deal with aspects related to classroom management. However, the analysis of the classroom observation and the focus group transcription revealed several important challenges that needed attention (**Table 9**).

The most important challenge was related to teachers' ability to teach the unit using inquiry and the engineering design process strategies. Thus, it has been observed that teachers lacked solid didactic knowledge to structure the different phases of the inquiry investigations and for discussing those aspects most important for each stage. For example, the progression from formulating an investigable research question and formulating hypothesis was not explicitly made to the students.

Likewise, the importance of designing a procedure that can address the formulated hypotheses, and that is ultimately related to the research question being investigated, was neither highlighted. Therefore, the didactic unit was delivered in an unclear, disorganized, and illogical order.

In addition, the two investigations about water quality parameters and the mixture separation techniques, developed during the knowledge development phase, were disjointed and the connections between them were inconsistent. The lack of explicit connection between the knowledge development and the knowledge application phase was also a constant problem that threatened the coherence of the intervention unit, and thus, the validity of the STEM-based framework. Thus, although teachers knew some fundamentals about what to do in each lesson, they were not explicit about it with the students, and therefore, students did not understand the relationship between science and engineering, nor the importance of mathematics and technology. Too very often, students did not even understand the reason behind some activities. Thus, for example, the elaboration of hypotheses was a mere activity that had to be done because the teachers asked so but lacked any value for students (and therefore, for the scientific investigations performed). Similarly, students did not seem to understand the logical progression and relationship between knowing the parameters that determines the quality of a sample of water for human consumption and the different mixture separation techniques, or the relationship between the materials used during the water filter design and the mixture separation techniques.

Likewise, most of the time, teachers missed opportunities to make connections to anything beyond the lesson or the inquiry investigation itself. For example, teachers did not make any effort to connect the scientific inquiries to the initial problem of the unit or to other real-world problems, or at the end of each investigation, to discuss about the hypotheses earlier formulated and their relationship with the results obtained. Therefore, it was observed that during the knowledge application phase, when designing and testing water filters, students selected materials at random, without actually applying the knowledge acquired through the two previous inquiry investigations conducted during the knowledge development phase of the unit. All these limitations were reflected in the engagement of students. Thus, a great disparity in student participation was observed. While some students of each working group were performing the hands-on activities, the other member of the group did not actively participate.

Finally, some teachers referred to the lack of content knowledge as an important challenge, but they overcame this by preparing the lessons in advance, and from the experience acquired through the implementation of the same lessons week after week. There was no mention about issues related to inquiry investigations being time demanding, curricular restrictions, or to teaching materials (Pozuelos, González, & Cañal, 2010), as the lessons were already designed, and the needed material and resources were easily accessed and prepared in advance.

To overcome these limitations, the following changes were made to the second prototype. First, related to the lack of clear structure of the unit and the minimum attention given by the teacher to make connections between the different activities and phases of the unit, the inquiry, and the engineering design process cycles were introduced, and the 15 sessions were rearranged accordingly to each step from the cycles. More specifically, the inquiry cycle established by Pedaste et al. (2015) was adopted during both water quality parameters and mixture separation techniques inquiry investigations performed during the knowledge development phase of the unit. Therefore, the unit sessions have been re-programmed in such a way that each session was devoted to addressing similar aspects of the inquiry cycle: (i) orientation sessions, where the teacher stimulated students curiosity through introducing the real-world problem in a meaningful way; (ii) conceptualization sessions, where research questions and hypotheses were discussed and formulated; (iii) investigation sessions, where students engaged in data collection and analysis; (iv) conclusion sessions, where the teacher and students drawing conclusions from the data based on the research questions and hypotheses formulated, and (v) discussion sessions, where students communicate their findings.

As for the knowledge application phase of the unit, a five-step engineering design process (Cunningham, 2009) was adopted. During the first step (ask sessions), students analyzed the problem under investigation, got to know how other engineers have approached the problem, and examined the constraints of the design solution. The second step (imagine sessions), consisted of students brainstorming different possible solutions, discussing the advantages and disadvantages of each one, and then selecting the best one. Then, in the third step (plan sessions), students drew a diagram of their intended product or technological solution and analyzed the materials that they will need considering the constraints of the problem. The fourth step (create sessions) is the moment when students built their proposed product and tested their efficacy. Finally, the fifth and last step of the engineering design process (improve sessions) consisted of students analyzing the limitations of their products to discover what worked out and what could work better; then, they modified their design and improved their product. As already mentioned, these inquiry and engineering design phases were distributed throughout the different 15 sessions that lasted each unit, allowing specific time to each phase so that students can in-depth reflect on them.

Likewise, to improve student's active engagement during each inquiry and engineering step, specific roles for each group member were designed (Johnson, Johnson, Holubec, & Roy, 1994), ,establishing up to four-member roles which were rotated through the 15-weekly sessions so that each member of the group can adapt each role. The first role, called manager, had the task to remind to the group the steps to be performed to stay on track, organize and distribute the materials, resolve conflicts and help reach consensus, monitor the time and noise levels and, and review the order and cleanliness of the group's work area. The second role, called speaker, is in charge of communicating doubts, problems or the resolution of activities to the teacher when required, by gathering the information agreed upon by the group and communicating it to the whole class and to the teacher. The third role is the secretary, who is in charge of collecting data in the field notebook. Each student had their own personal field notebook; however, it was observed that sometimes students were focused on carrying out the hand-on activities but did not collect any data in the notebooks. Therefore, the secretary's responsibility is to keep track of the results, remind classmates of the importance of data collection, and check that everyone has recorded the results in their field notebooks. Finally, the role of the researcher has been established; this member was charged with the realization of the different hands-on activities. It should be noted that this does not imply that the other members cannot participate in the hands-on activities; what this role aims at is to avoid arguments that arise when the task to be performed requires the participation of a single person (e.g., introducing the microcontroller board into the water samples to measure the temperature).

5.1.3. Third prototype

The third prototype, with the improvements mentioned above, was implemented for 5 weeks (from January 8th to February 8th, 2018). Although the new lesson structure based on the inquiry cycle and engineering design process improved the coherence and organization of how the unit was delivered, and the new group roles fostered equal participation of the group members, (especially among girls that often did not participate in the hand-on activities), the third focus group and the observation of 75 lessons revealed challenges related with students' inquiry and engineering skills, and teachers' difficulties of fostering them (**Table 10**).

It was observed that teachers lacked propositional knowledge to foster critical discussion between students and in-depth reflection about the different inquiry skills pertaining to doing inquiry (e.g., formulating investable research questions; designing and conducting an investigation that tackles the formulated question). For example, teachers asked the class to make predictions, but these were not related to the procedure used for testing them, nor to the results obtained from the investigations performed. In addition, student talk, reflection, and negotiation were almost inexistent, especially between peers, because they were only offered opportunities to answer teacher questions.

Likewise, due to an overly structured approach adopted by the teachers, students had not real possibilities to think critically about the phenomenon under study, nor opportunity to develop a thoughtful hypothesis or to engage in thought-provoking activities, like devising the procedure, negotiating meanings or synthesizing the outcomes observed. In the end, the inquiry investigations and the engineering design process were delivered in a teacher-centered, lecture-based style.

In order to improve teacher's propositional and didactic knowledge to teach using inquiry and engineering design teaching pedagogies, different strategies from the literature were used as scaffolding to allow students' engagement in reflective practices. Based on existing literature about improving teacher questioning and scaffolding during science inquiry, several questions and strategies were developed for each inquiry and engineering design step that teachers would use to foster communication and active participation of students (**Table 11**).

In addition, constructive feedback was encouraged. Based on Chin (2006), instead of using a direct feedback stating that the student answer was right or wrong, teachers were instructed to use several strategies: (i) when students response were correct, affirmation-direction instruction type of feedback that reinforce the response and provide further examples; (ii) when there is a disagreement in the class about the correct answer, teachers would accept all responses followed by several related questions to make students reflect on their answer until all working groups understand what would be the correct answer and why; and (iii) when no working group reached a correct response (i.e., no hypotheses related to the research questions was formulated), teachers could reformulate the instructions and the question to foster more discussion between students, instead of revealing the correct answer.

5.1.4. Final prototype

The fourth and final prototype was implemented for four weeks (from February 26th to March 23rd, 2018). The changes introduced for the fourth prototype provided pedagogical significance to

Table 10. Challenges identified in the third prototype and educational measures

Codes representing challenges	Educational measures
Teachers' lack of propositional knowledge (Toma, Greca & Meneses Villagrá, 2017) Lack of connection between the knowledge development phase and the knowledge application phase of the unit	- Scaffolding questioning and strategies - Constructive feedback

The codes with a reference were derived using direct content analysis, and the ones without any reference using conventional content analysis.

Table 11. Scaffolding strategies developed to overcome challenges from the third prototype

Questions types	Scaffolding questions
[1]Interpretation (foster observation and inferences)	What happened to the sample of water…? Can you describe your results?
[1]Consequence (ask for explanations)	If we do… what will happen to….? What would happen if you change…?
[1]Expectational (help generate predictions)	How can we proceed to test this question? What can we do to….?
[1]Judgmental (reinforce main ideas)	So, what have we learned about….? What are the take-home messages of…?
[1]Monitoring (check progress)	What are you doing with ….? Are you advancing on…?
[2] Synthesis (relate existing knowledge)	How is this related to previous lessons? What do you already know about…?
[2]Motivation (focus attention)	Would you drink this water? [while showing a contaminated water sample] Do you think this design would be effective? [while showing a water filter sample]
[2]Eliciting (check preconceptions)	Why do human beings need to drink water? Why it is important for our problem to depurate the water found on Mars planet?
[2]Extending (foster new ideas)	Which material would be useful to…? What characteristics must our design meet? What problem can be face if…?
[2]Application (use existing knowledge)	Why would you use that material for your water filter? How can we improve the turbidity? [while showing a different sample of contaminated water]
[3]Exploring pre-requisites	What are our constraints? What are we investigating?
[3]Hinting (generate explanations)	How can we find out? Why did we do this practice? Is this practice related to the one that we have done yesterday?

[1]Based on Erdogan & Campbell (2008).
[2]Based on Yip (2004).
[3]Based on Kawalkar & Vijapurkar (2013).

the intervention unit and facilitated the teaching of the unit through inquiry and engineering design teaching strategies. It was observed that students have had more opportunities to share and develop their previous ideas and that the different phases of the inquiry and engineering process were delivered in a more significant and less anecdotal way. In this way, it seemed that the students acquired a better understanding of the importance of these steps, as well as of the relationship between the knowledge development and the knowledge application phases. Based on these results, it was decided to not conduct another focus group, but just make minor adjustments based on the classroom observation and informal conversation between the researcher and the teachers after each session.

After two weeks of implementing the fourth prototype, it was observed that given the new focus on reflective thinking during each phase of the inquiry cycle and the engineering-design process, the implementation of 15 sessions in a week was too demanding for students. Therefore, it was decided to reduce the didactic unit from 15 lessons to only 12 total weekly lessons, which in addition, would resemble more faithfully the temporalization of a teaching unit in a conventional school. Therefore, some inquiry cycle and engineering design steps were delivered in the same session, as reported in Chapter 3 and Appendix 1 of this dissertation.

Likewise, it was decided to trial different amounts of guidance offered by the teacher during the knowledge development phase[10]. This decision was driven by the experience of the first three prototypes in which it has been observed that teachers who are unfamiliar with these active teaching strategies tend to turn them into structured, teacher-centered activities. Given that this has been the case for the teachers implementing the intervention unit at the CRIEB, it is assumed that this result could also be expected in conventional school centers, where there are other limitations such as the lack of time, the need to design and produce the teaching materials, curricular pressures, or workload (Cañal, 2007; Pozuelos, González, & Cañal, 2010).

Based on how much information and guidance is provided to students, Herron (1971) proposed a four-level continuum of inquiry, from confirmation, where students confirm a phenomenon following specific instruction with the research question and procedure already provided by the teacher and where the potential results are known in advance, to an open-type inquiry where students follow self-directed investigations by formulating research questions, design procedures and generate explanation based on evidences collected. Therefore, a structured inquiry and a guided inquiry group were created. Students in the structured inquiry group participated in the didactic unit where the research question was already established by the teacher, and where there was a specific procedure for carrying out the hands-on activities, without opportunity for proposing new procedures. On the other hand, students from the guided inquiries participated in the same didactic unit; however, the scaffolding questions and strategies reported in the third prototype section were used during the whole process of the inquiry implementation. Thus, in the orientation and conceptualization steps, while in the structured group students were limited to reading the problem statement and the research question, in the guided group, students were encouraged to identify the initial real-world problem and formulate research questions and hypotheses. In the investigation step, the structured group just followed the procedure given by the teacher and the guided group students were first

[10] It should be noted that it was decided to test it only with upper elementary graders because lower graders, aged 8-9 years old, were found to be not cognitively ready to work in such an autonomous manner.

given an opportunity to design their own procedure until reaching a similar procedure as the one used in the structured group. In the conclusion step, the teacher summarized the results obtained for the structured group; however, in the guided group this summary was done by reflecting on the research questions and hypotheses formulated. Finally, the discussion step was identical in both pedagogical approaches.

5.2. Quantitative strand

5.2.1. Pre-test data analysis

The aim of this study was to identify the interest of elementary girls and boys towards physics or biology science content and analyze whether there are differences in their interests according to gender and age variables (Toma & Meneses Villagrá, 2019b). In addition, the aim is to examine whether there are differences in attitudes towards school science according to their science interests' profiles. In this section, a summary of the analysis of CRIEB pretest data is reported. The full-text published article is attached in Appendix 8.

For data collection, the Spanish-School Science Attitudes Survey (S-SSAS), described in the previous section, was used. Interest in biology and physics content was determined from the analysis of two items from the "relevance of school science to personal life" construct: item 8, "I want to learn about the plants in my area" (biology interest) and item 10, "I want to learn about electricity and how it is used in a house" (physics interest). A non-hierarchical cluster analysis, called K-Means, was used for the extraction of the conglomerates that have determined the interest profiles for each student. The sample was composed of 733 students from 3rd to 6th grade of elementary education, enrolled in 29 schools from the province of Burgos that assisted the CRIEB school center during the formative phase of the main study.

The findings revealed the existence of four interest profiles:

(i) interest in physics, with students reporting significantly greater interest in physics contents than in biology;
(ii) interest in biology, with students reporting greater interest in biology contents than in physics;
(iii) low interest, with students reporting disinterest in both types of science contents; and
(iv) high interest, with students reporting high interest in both disciplines.

In relation to school-level variable participants were found to be uniformly represented in all four profiles. However, with respect to the gender variable, there was a significant over-representation of the girls with respect to the boys in the "interest in biology" profile and a significant under-representation of the girls in the "interest in physics" profile with respect to the boys. In addition, there were significantly more girls than boys grouped in the 'low interest' profile and significantly fewer girls than boys in the "high interest" profile. Finally, students in the "high interest" group are those who have reported better attitudes towards school science than those in other interest profiles.

Taken together, these results indicate that students' interest in scientific disciplines of physics or biology is shaped from elementary levels of the educational system. These results have direct implication for this dissertation, as they highlight that the aim of this dissertation (i.e., design and develop educational interventions from an early age to improve students' attitudes towards school science) is highly relevant and timely for the Spanish context.

5.2.2. Expectancies of success in school science

5.2.2.1. Lower grades

The 2 x 2 mixed between-within subjects ANOVA revealed that there was no significant interaction effect for time*gender, Wilks Lambda = 1.00, F (1, 63) = .01, p = .939. These results suggest that there is no significant gender difference in pre and posttest scores. Therefore, the main effects of time can be interpreted with confidence. There was a non-significant main effect for time, Wilks Lambda = .958, F (1, 63) = 2.816, p = .098. Inspection of mean test for girls and boys revealed that, following participation in the intervention unit, there was a negligible increase in student's expectancies of success in school science (**Table 12**).

Likewise, the main effect comparing gender was neither statistically significant, F (1, 63) = 1.471, p = .230, suggesting that the intervention unit has had no gender differential effect in lower graders expectancies of success in school science. These results were also supported when controlling for pretest scores. Thus, ANCOVA revealed that the interaction effect for gender was not significant, F (1, 62) = .443, p = .508. There was a strong relationship between the covariate (pretest scores) and the dependent variable (posttest scores), as indicated by η_p^2 value of .509, suggesting that even when controlling for student's expectancies of success scores prior to the intervention, the STEM-based unit has had no differential effect in lower grades girls' and boys' expectancies of success in school science (**Table 13**).

Table 12. Expectancy of success pretest and posttest results

					Multivariate[a]		
Construct	Time	Gender	*M*	*SD*	*F*	*p*	η_p^2
Expectancy of success	Pretest	Girls	3.79	0.73	2.816	.098	.042
		Boys	3.58	0.94			
	Posttest	Girls	3.92	0.81			
		Boys	3.69	0.67			

[a]Multivariate results are reported for pretest and posttest scores comparison.

Table 13. ANCOVA results for expectancy of success variable

			95% Confidence Interval	
Gender	*M*[a]	Std. Error	Lower Bound	Upper Bound
Girls	3.85	.09	3.68	4.03
Boys	3.77	.10	3.58	3.96

[a]Covariates appearing in the model are evaluated at the following pretest values: Expectancy = 3.69.

5.2.2.2. Upper grades

The 2 x 2 x 2 ANOVA revealed no significant interaction effect for time*pedagogical approach*gender, Wilks Lambda = .993, F (1, 171) = 1.17, p = .281; time*gender, Wilks Lambda = .996, F (1, 171) = .72, p = .397; and time*pedagogical approach, Wilks Lambda = .999, F (1, 171) = .13, p = .718. These results suggest that there is no significant gender and pedagogical approach difference in pre and posttest scores. Therefore, main effects for time can be interpreted with confidence.

There was a significant main effect for time, Wilks Lambda = .917, F (1, 171) = 15.411, p < .001, with both pedagogical approaches improving upper grades students' expectancies of success (**Table 14**). The inspection of effect sizes for each pedagogical approach suggests slightly greater improvements in the guided (η_p^2 = .129) compared to the structured approach (η_p^2 = .059). However, the main effect comparing the two types of pedagogical approaches was not significant, F (1, 171) = .444, p = .506, revealing that there is no difference in the effectiveness of both approaches (**Figure 7**).

Likewise, the main effect comparing gender was also not significant for upper grades students' expectancy of success, F (1, 171) = .216, p = .643, revealing that the intervention unit had no gender differential effect, in both the structured or guided approach. Consequently, following Cohen's (1988) guidelines, these results suggest that both approaches have a moderate effect on upper grades girls' and boys' expectancies of success in school science.

In addition, these results do hold when controlling for pretest scores. Consequently, ANCOVA reveals that the interaction effect for pedagogical approach*gender was not significant, F (1, 170) = 2.283, p = .133, and that there was a strong relationship between the covariate (pretest scores) and the dependent variable (posttest scores), as indicated by η_p^2 value of .545. Neither of the main effects for pedagogical approach, F (1, 170) = .324, p = .570, and gender, F (1, 170) = .905, p = .343, were statistically significant. These results suggest that when controlling for students' expectancies of success levels prior to the intervention, the two types of approaches had no differential effect for the girls and boys included in this study (**Table 15**). In other words, girls and boys responded similarly to the structured and guided approach.

Taken together, the results suggest that the intervention unit, designed according to the pedagogical framework for STEM-based science education developed in this dissertation, was only effective in improving 5th and 6th graders expectancies of success in school science. Moreover, the results suggest that students' expectancies of success improved in both the structured and the guided approach to inquiry investigation, suggesting no added benefit of reducing teacher guidance (i.e., using a guided inquiry instead of a structured one) during the knowledge development phase of the unit.

Table 14. Expectancy of success pretest and posttest scores

			Approach			
			Structured		Guided	
Construct	Time	Gender	*M*	*SD*	*M*	*SD*
Expectancy of success	Pretest	Girls	3.70	0.59	3.85	0.71
		Boys	3.81	0.73	3.77	0.62
	Posttest	Girls	3.76	0.64	4.02	0.71
		Boys	4.02	0.62	3.91	0.74

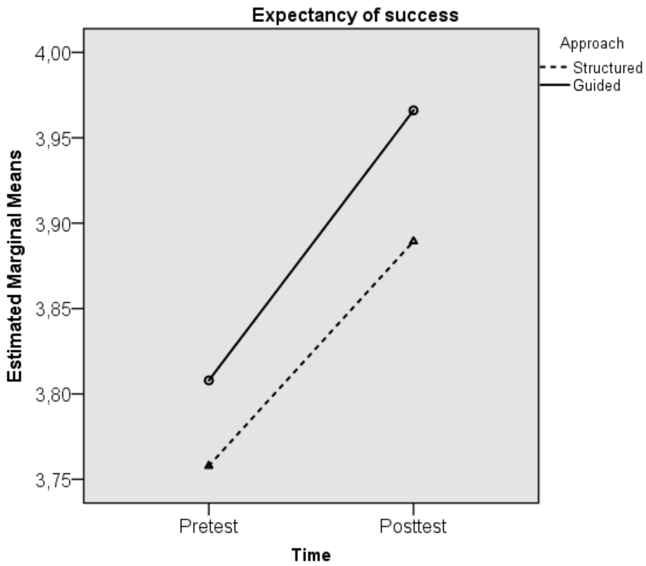


Figure 7. Secondary interaction effects of time of test and pedagogical condition on expectancy of success in school science variable.

Table 15. 2x2 ANCOVA results for expectancy of success variable

				95% Confidence Interval	
Approach	Gender	*M*[a]	Std. Error	Lower Bound	Upper Bound
Structured	Girls	3.82	.07	3.69	3.95
	Boys	3.99	.06	3.87	4.12
Guided	Girls	3.97	.07	3.82	4.11
	Boys	3.93	.07	3.78	4.07

[a]Covariates appearing in the model are evaluated at the following pretest values: Expectancy = 3.78.

5.2.3. Attitudes towards school Science

5.2.3.1. Lower grades

The 2 x 2 MANOVA with repeated measures revealed that there was no secondary significant interaction effect for time*gender, Wilks Lambda = .944, F (6, 59) = .594, p = .734, suggesting a non-significant gender difference in pre and posttest scores. Therefore, the main effects of time for each attitudinal factor can be interpreted with confidence.

Since there was a significant main effect for time, Wilks Lambda = .684, F (6, 59) = 4.615, p = .001, η_p^2 = .316, showing that the STEM-based intervention unit improved the combined six dependent attitudes towards school science variables, univariate test were performed and interpreted for each S-SSAS individual attitudinal variable or factors.

When considering the univariate time results for each factor, there was a substantial main effect for intention, F (1, 64) = 7.012, p = .010, η_p^2 = .097, enjoyableness, F (1, 64) = 25.664, p < .001, η_p^2 = .283, and self-efficacy F (1, 64) = 4.350, p = .041, η_p^2 = .063 attitudinal factors (**Table 16**). These results suggest that the intervention unit was effective in increasing lower grades girls' and boys' intentions to further enroll in school science, enjoyableness of school science lessons and self-efficacy in school science, but did not improve students perceived difficulty, usefulness and relevance of school science.

The main effect comparing gender was not significant, Wilks Lambda = .933, F (6, 59) = .589, p = .738, revealing that the intervention unit has had no gender differential effect in lower grades elementary students' attitudes towards school science. Therefore, the intervention unit was effective in improving three (i.e., intentions, enjoyableness, and self-efficacy) out of the six attitudinal constructs measures by the S-SSAS instrument, in both girls and boys.

These results are further supported when controlling for pretest scores. Therefore, MANCOVA revealed that the interaction effect for gender was not significant, Wilks Lambda = .32, F (6, 53) = 54.00, p = .683. There was a medium to the large relationship between the covariates (pretest scores) and the dependent variables (posttest scores), as indicated by η_p^2 values ranging from .103 to .304 when controlling for intention, enjoyableness and self-efficacy pretest scores, respectively. These results suggest that even when controlling for attitudes towards school science prior to intervention, the intervention unit had no differential effect based on gender variable (**Table 17**).

Table 16. Attitudes pretest and posttest univariate results

					Univariate[a]		
Constructs	Time	Gender	*M*	*SD*	*F*	*p*	η_p^2
Intention	Pretest	Girls	2.72	1.06	7.01	.010*	.097
		Boys	2.48	1.29			
	Posttest	Girls	2.97	1.08			
		Boys	3.03	1.33			
Enjoyableness	Pretest	Girls	3.17	1.40	25.66	.000*	.283
		Boys	3.19	1.47			
	Posttest	Girls	4.03	0.84			
		Boys	3.81	0.87			
Difficulty	Pretest	Girls	2.56	1.21	.61	.438	.009
		Boys	2.71	1.35			
	Posttest	Girls	2.33	1.35			
		Boys	2.68	1.14			
Self-efficacy	Pretest	Girls	3.44	1.27	4.35	.041*	.063
		Boys	3.32	1.47			
	Posttest	Girls	3.89	1.09			
		Boys	3.58	1.15			
Usefulness	Pretest	Girls	3.86	0.83	.31	.580	.005
		Boys	3.87	0.87			
	Posttest	Girls	3.90	0.82			
		Boys	3.95	0.86			
Relevance	Pretest	Girls	3.72	0.74	1.78	.187	.027
		Boys	3.50	0.95			
	Posttest	Girls	3.85	0.71			
		Boys	3.67	0.76			

*Statistically significant at p = .05 level.
[a]Univariate results are reported for pretest and posttest scores comparison.

Table 17. MANCOVA results for attitude towards science variable

				95% Confidence Interval	
Dependent Variable	Gender	*M*[a]	Std. Error	Lower Bound	Upper Bound
Intention posttest	Girls	2.92	.17	2.57	3.26
	Boys	3.10	.19	2.72	3.47
Enjoyableness posttest	Girls	4.01	.11	3.80	4.22
	Boys	3.83	.12	3.60	4.06
Difficulty posttest	Girls	2.34	.19	1.95	2.72
	Boys	2.67	.21	2.26	3.09
Self-efficacy posttest	Girls	3.87	.16	3.55	4.20
	Boys	3.60	.18	3.24	3.95
Usefulness posttest	Girls	3.88	.13	3.63	4.14
	Boys	3.97	.14	3.70	4.25
Relevance posttest	Girls	3.81	.12	3.57	4.05
	Boys	3.71	.13	3.45	3.96

[a]Covariates appearing in the model are evaluated at the following pretest values: intention = 2.61, enjoyableness = 3.18, difficulty = 2.63, self-efficacy = 3.39, usefulness = 3.87, relevance = 3.62.

5.2.3.2. *Upper grades*

The 2 x 2 x 2 MANOVA revealed that there was no significant interaction effect for time*pedagogical approach*gender, Pillai's Trace = .028, F (6, 161) = .786, p = .582; time*gender, Pillai's Trace = .019, F (6, 161) = .518, p = .794; and time*pedagogical approach, Pillai's Trace = .054, F (6, 161) = 1.544, p = .167. These results suggest that there is no significant gender and pedagogical approach difference in pre and posttest scores. Therefore, main effects for time can be interpreted with confidence. Since there was a significant main effect for time, Pillai's Trace = .153, F (6, 161) = 4.844, $p < .001$, $\eta_p^2 = .153$, with both pedagogical approaches improving the combined six dependent attitudes variables, univariate test were performed and interpreted for each attitudinal factor. There was no tertiary time*gender*pedagogical approach, nor secondary time*gender significant interaction effect for any attitudinal factor. This was also the case for the secondary time*approach interaction effect for all but the perceived difficulty factor, F (1, 166) = 5.45, p = .021, $\eta_p^2 = .032$. Examination of the means (**Tables 18** and **19**) and the plots (**Figure 8**) suggest that while the structured approach was effective in lowering students perceived difficulty of school science, F (1, 94) = 3.75, p = .04, $\eta_p^2 = .038$, the guided approach slightly increased students perceived difficulty, although this last result was not significant, F (1, 74) = 1.89, p = .174.

Table 18. Attitudes pretest and posttest univariate results for the structured pedagogical approach

					Univariate[a]		
Construct	Time	Gender	*M*	*SD*	*F*	*p*	η_p^2
Intention	Pretest	Girls	2.93	1.07	.227	.634	.002
		Boys	2.80	1.07			
	Posttest	Girls	3.07	0.75			
		Boys	2.78	1.04			
Enjoyableness	Pretest	Girls	3.16	1.26	6.81	.011*	.068
		Boys	3.46	1.25			
	Posttest	Girls	3.49	0.97			
		Boys	3.74	0.99			
Difficulty	Pretest	Girls	2.53	1.25	3.75	.050*	.038
		Boys	2.46	1.28			
	Posttest	Girls	2.18	0.89			
		Boys	2.32	1.28			
Self-efficacy	Pretest	Girls	3.33	1.04	3.49	.065	.036
		Boys	3.78	1.02			
	Posttest	Girls	3.60	0.81			
		Boys	3.88	0.98			
Usefulness	Pretest	Girls	3.82	0.89	.26	.610	.003
		Boys	3.92	0.98			
	Posttest	Girls	3.84	0.68			
		Boys	4.00	0.80			
Relevance	Pretest	Girls	3.44	0.83	2.71	.103	.028
		Boys	3.81	0.76			
	Posttest	Girls	3.66	0.79			
		Boys	3.85	0.84			

*Statistically significant at p = .05 level.
[a]Univariate results are reported for pretest and posttest scores comparison.

Univariate results for the remaining attitudinal factors revealed that both approaches significantly increase student's enjoyableness of school science lessons, with a slightly higher effect size for the guided (η_p^2 = .090) than the structured approach (η_p^2 = .068), although this difference was not statistically significant, F (1, 166) = 1.75, p = .188, suggesting no difference between approaches in terms of effectiveness. Consequently, following Cohen's (1988) guidelines, these results suggest that both approaches have a moderate effect on upper grades girls' and boys' enjoyableness of school science lessons.

Finally, univariate results revealed a significant main effect for time only in the guided approach to inquiry for the intention to further enroll in school science attitudinal factor (**Tables 18** and **19**). The main effect comparing approach was significant, F (1, 166) = .7.407, p = .007, η_p^2 = .043, suggesting that the guided approach was more effective in increasing students' intentions to further enroll in school science than the structured one.

The main effect comparing gender was also not significant for upper grades students' attitudes towards school science, Pillai's Trace = .936, F (6, 161) = .1.838, p = .095, suggesting that the intervention unit, in both the structured or guided inquiry approach, has had no gender differential effect in upper grades elementary students attitudes towards school science.

Table 19. Attitudes pretest and posttest univariate results for the guided pedagogical approach

					Univariate[a]		
Construct	Time	Gender	*M*	*SD*	*F*	*p*	η_p^2
Intention	Pretest	Girls	3.33	0.96	3.91	0.052*	.050
		Boys	2.97	1.06			
	Posttest	Girls	3.42	1.16			
		Boys	3.36	1.18			
Enjoyableness	Pretest	Girls	3.72	0.94	7.29	.009*	.090
		Boys	3.31	1.30			
	Posttest	Girls	3.97	0.91			
		Boys	3.64	1.16			
Difficulty	Pretest	Girls	1.86	1.13	1.89	.174	.025
		Boys	2.36	1.33			
	Posttest	Girls	2.14	1.15			
		Boys	2.54	1.19			
Self-efficacy	Pretest	Girls	3.61	0.90	3.39	.070	.044
		Boys	3.54	1.14			
	Posttest	Girls	3.94	0.83			
		Boys	3.64	1.06			
Usefulness	Pretest	Girls	4.18	0.73	2.279	.135	.030
		Boys	3.95	1.04			
	Posttest	Girls	3.97	0.74			
		Boys	3.86	0.87			
Relevance	Pretest	Girls	3.76	0.71	1.711	.195	.023
		Boys	3.68	0.79			
	Posttest	Girls	3.81	0.72			
		Boys	3.82	0.71			

*Statistically significant at p = .05 level.

[a]Univariate results are reported for pretest and posttest scores comparison.

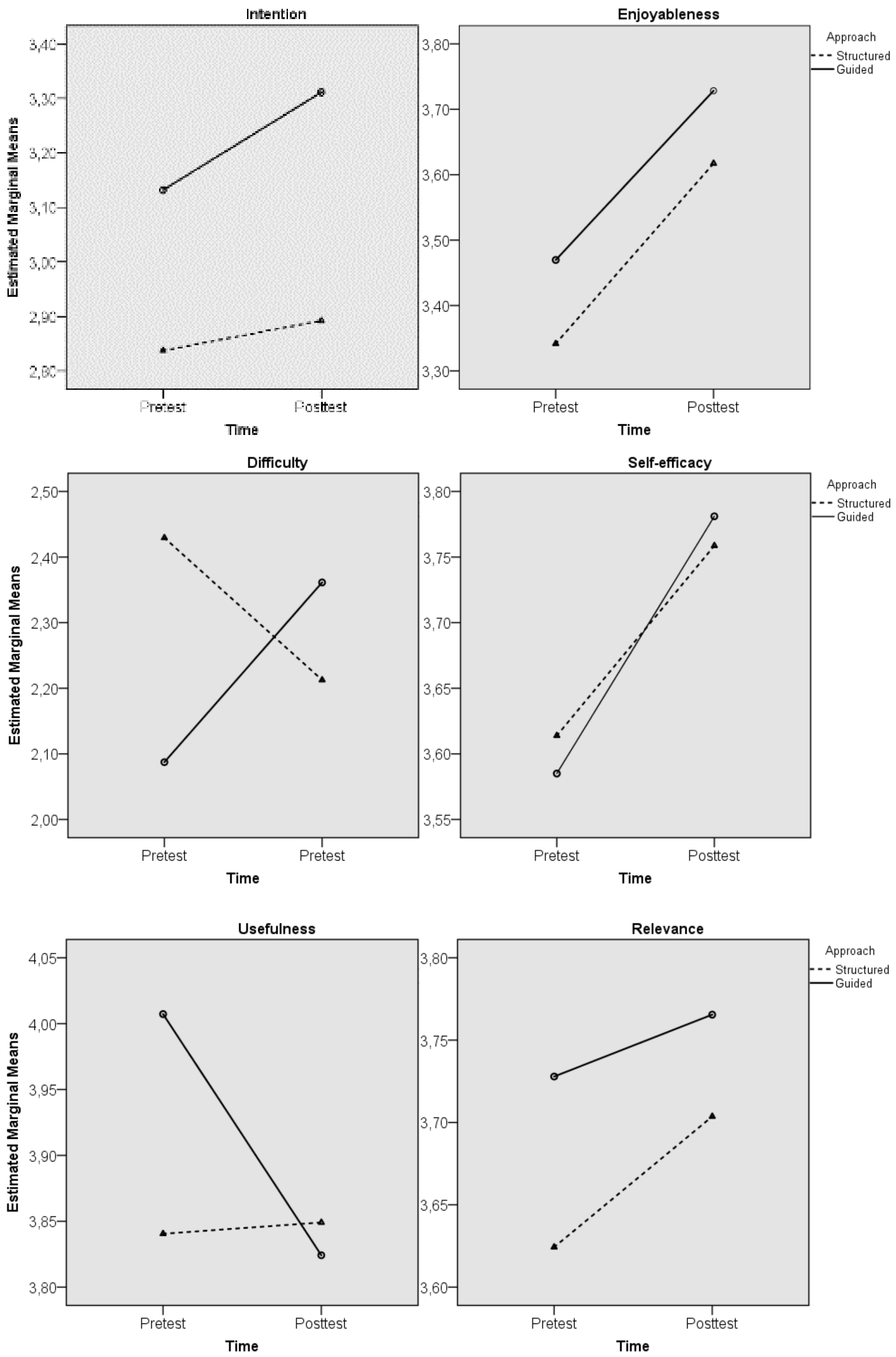


Figure 8. Secondary interaction effects of time of test and pedagogical condition on attitudinal variables

Table 20. 2 x 2 MANCOVA results for attitude towards science variable

Dependent Variable	Approach	*M*[a]	Std. Error	95% Confidence Interval Lower Bound	95% Confidence Interval Upper Bound
Intention	Structured	2.98	.09	2.80	3.17
	Guided	3.31	.11	3.10	3.52
Enjoyableness	Structured	3.68	.08	3.51	3.84
	Guided	3.73	.09	3.55	3.92
Difficulty	Structured	2.19	.11	1.98	2.40
	Guided	2.41	.12	2.18	2.65
Self-efficacy	Structured	3.75	.08	3.59	3.92
	Guided	3.79	.09	3.60	3.97
Usefulness	Structured	3.95	.07	3.81	4.09
	Guided	3.88	.08	3.72	4.04
Relevance	Structured	3.79	.06	3.67	3.92
	Guided	3.77	.07	3.63	3.92

[a]Covariates appearing in the model are evaluated at the following pretest values: Intention = 2.99; Enjoyableness = 3.40; Difficulty = 2.33; Self-efficacy = 3.57; Usefulness = 3.96; Relevance = 3.67.

Finally, after adjusting for pretest scores, MANCOVA results revealed that the interaction effect for the pedagogical approach*gender was not significant for any attitudinal factor. There was a medium to the large relationship between the covariates (pretest scores) and the dependent variables (posttest scores), as indicated by η_p^2 values ranging from .132 to .300. Similarly, the main effect for gender was neither statistically significant for any attitude construct, suggesting that girls and boys responded similarly to the two inquiry approaches. Finally, the only main effect for approach to reach statistically significant results was the intention factor, $F\,(1, 160) = 4.23$, $p = .025$, $\eta_p^2 = .031$, with students in the structured approach showing higher mean scores (**Table 20**). Thus, the MANCOVA results support the findings previously reported related to the differential effect of both pedagogical approaches (i.e., structured vs. guided). More specifically, MANCOVA reveals that even when controlling for pretest scores, the two types of integrative-STEM approaches had a differential effect in students' intention to further enroll in school science, favoring the guided approach.

Taken together, the attitudinal results suggest that the STEM-based intervention unit was effective in improving 3rd and 4th graders intentions to further enroll in school science, perceived enjoyableness of school science lessons, and self-efficacy in school science. Likewise, the intervention unit was effective in improving the 5th and 6th grader's enjoyableness of school science lessons, regardless of whether a guided or structured approach has been used during the inquiry investigations developed in the knowledge development phase of the unit. In addition, while the guided approach was more effective in improving student's intention to further enroll in school science, it also increased students' perceived difficulty of school science, which in fact significantly decreased when structured inquiries were used instead of guided ones. In other words, the findings suggest that delivering the STEM unit using a structured approach for the inquiry investigations, performed during the knowledge development phase, was more effective in reducing student's perceived difficulty of school science than when the guided approach to inquiry was used.

In short, while these results suggest that the intervention unit was effective in improving students' attitudes towards school science, the results are inconclusive as to what type of inquiry investigations (structured or guided) should be used during the knowledge development phase when engaging 5th and 6th graders in this type of educational paradigm.

Chapter 6. Discussion

This chapter reflects on this thesis dissertation journey as a whole. Section 6.1. discusses the main findings of the intervention study. Next, section 6.2. and 6.3. summarizes the significance and educational and research implications of these studies for science education. Finally, section 6.4. reflects on the limitation of this dissertation and closes with recommendations for future endeavors in STEM-based science education and attitude towards science research.

6.1. Discussion of the findings

The purpose of this thesis dissertation was to advance understandings of the STEM educational movement by proposing a pedagogical framework for designing valid and plausible teaching units for STEM-based science education at elementary school level, in Spain. Likewise, this study intended to evaluate the effectiveness of such educational paradigm (i.e., STEM as teaching science through interdisciplinary approaches) on improving elementary students' expectancies of success and attitudes towards school science, which would contribute to fixing the leaky science pipeline (i.e., disengagement in science studies and careers). To attain these aims, the mixed-methods multiphase design was adopted (Creswell & Clark, 2011), that combined several qualitative, quantitative and mixed-methods studies within an overall research project.

The overarching research question (i.e., What are the characteristics of a valid, practical, and effective pedagogical framework that facilitates the implementation of STEM-based teaching units in elementary science education for promoting students' expectancies of success and attitudes towards school science?) was subsequently divided into two specifics ones. The first specific research question addressed how to conceptualize STEM education and what didactic strategies should be used to implement valid and plausible-STEM based units during science education lessons. To this end, constructivist and socio-constructivists learning theories served as a guiding psychological framework that helped to establish design principles to be considered when developing STEM-based teaching units.

The second specific research question aimed at examining whether the teaching unit designed following the pedagogical framework developed according to constructivist and socio-constructivists learning theories, was effective in improving elementary school students expectancies of success and positive attitudes towards school science. To this effect, two systematic reviews were performed to select the measurement instruments most appropriated for the CRIEB context. Given the lack of valid and reliable instruments in the existing literature, two instruments, rooted in the tripartite model of attitudes (Khine, 2015; Simpson et al., 1994).) and the expectancy-value theory of achievement motivation (Eccles et al., 1983), were developed and assessed in terms of classical test psychometric properties using the population that participated in the CRIEB school center during the 2017-18 school year.

For the main study, this second research question was further divided into two tertiary research questions aimed at (i) examining if the level of guidance offered by the teacher (i.e., structured - teacher centered and low levels of students autonomy- or guided approach -student centered and low assistance from the teacher) would affect the dependent variables differently (Herron, 1971; Kirschner, Sweller, & Clark, 2006), and (ii) controlling the results by gender and grade level variables, consistent with existing literature revealing great disparity between girls and boys (Denessen et al., 2015; DeWitt & Archer, 2015; George, 2006; Hacıeminoğlu, 2016), and lower and upper graders (Ali et al., 2013; Said et al., 2016), with respect to the dependent variables examined.

Each of these topics is discussed, in-depth, in the sections that follow.

6.1.1. What teaching strategies should be used to implement valid and plausible STEM-based teaching units for science education in elementary grades?

In this dissertation, STEM was conceived as an educational paradigm aimed at learning science through its integration with technology, engineering and mathematics disciplines. While this conceptualization is consistent with other studies linking STEM to interdisciplinary education (e.g., Barak & Assal, 2018; Lou et al., 2017; Psycharis & Kotzampasaki, 2019), it differs in two critical aspects.

First, there is a common agreement in the STEM literature on categorizing a project under the STEM umbrella when it promotes the "(…) teaching and learning between/among any two or more of the STEM subject areas, and/or between a STEM subject and one or more other school subjects" (Sanders, 2009, p. 21). This conceptualization, that still persists in recent studies (see Martín-Páez et al., 2019 for a review of STEM studies), gives rise to several issues. On the one hand, if STEM is conceptualized as the integration of two disciplines, one might consider if STEM, therefore, adds anything new to science education, given that this approach has been promoted four decades ago when Science & Mathematics integration perspectives got popularized. Likewise, it is also worth considering why an approach similar to that of Science & Mathematics integration, whose effectiveness for science education has been limited at best (Czerniak et al., 1999), is expected to be more effective nowadays. On the other hand, this conceptualization is too simplistic and leads to a proliferation of studies that fall under the STEM umbrella without, in fact, promoting the teaching and learning of more than one STEM discipline. Thus, Martín-Páez et al. (2019) concluded that most STEM proposals are inconsistent in that the connection between disciplines is not explicitly described or does not exist. In this regard, this dissertation advances in conceptualizing STEM education by considering that in order for a project to be considered STEM, it should integrate the contents and procedures of the four STEM disciplines through the same teaching unit by explicitly making connections between these disciplines, being this conceptualization more coherent and novel than those present in the literature.

Second, the STEM definition adopted in this dissertation also differs from existing ones in that, while STEM is usually defined as a teaching approach (see van den Hurk, Meelissen, & van Langen, 2019 for a review of STEM interventions), this dissertation conceives STEM as an educational paradigm that can be delivered using different teaching strategies, such as the ones used in the intervention unit described in Chapter 3: inquiry-based science education and engineering design process. This helps to clarify even more what is STEM education and how it should be delivered in the classroom.

Based on this conceptualization, a pedagogical framework was developed, characterized by (i) embedding the STEM unit in real-world problems of educational relevance that are motivational for students; (ii) that coherently blends science and mathematics curricular content with technology and engineering practices; and that (iii) uses inquiry investigations and engineering design process teaching approaches to deliver the curricular content and to explicitly establish the similarities, differences, and connections among the four STEM disciplines. More specifically, it is proposed that this connection be made in two phases that attempt to simulate the process through which knowledge and technological and engineering solutions are developed in real life. Therefore, the first phase is focused on the development of scientific knowledge through students engaging in inquiry investigations, and the second phase is devoted to the design and testing of a process or technological solution through engineering practices by applying the knowledge developed in the first phase.

The formative evaluation phase (i.e., the qualitative strand from Chapter 5) revealed that such an approach could be implemented at the elementary school level, however, not without issues that need to be considered. Thus, the teachers needed up to fifteen weeks of implementing the unit to develop the necessary skills in order to implement it as designed (i.e., following the pedagogical approach developed in this dissertation). The main problems experienced by teachers were the ones already reported in the literature about teaching science through renovated pedagogies (i.e., Aragüés et al., 2014; King et al., 2001; Romero-Ariza et al., 2019). Thus, teachers presented difficulties in terms of (i) classroom management, due to an increased messiness and noise levels derived from hands-on activities such as inquiry investigations, engineering design, and computer coding of microcontroller boards; (ii) lack of pedagogical content knowledge for teaching science using active strategies, which made teachers to adopt a teacher-centered and structured approach for delivering the unit; (iii) lack of propositional knowledge for teaching in an integrative way by making connections between the STEM disciplines.

To this end, the restructuration of the classroom into collaborative groups and the use of the strategies reported by Keeley (2008) and Qablan & DeBaz (2015) were effective in facilitating the implementation of the unit and by improving the classroom learning environment. Likewise, it was found that teachers benefitted heavily from dividing the unit sessions according to the steps of the inquiry cycle (Pedaste et al., 2015) and the engineering design process (Cunningham, 2009). In this way, the different phases of the inquiry cycle (e.g., Formulating research questions and hypotheses, designing an experiment, etc.), as well as those of the engineering design (e.g., Identifying a problem; determining the constraints; design a prototype, etc.) were not merely anecdotal but became essential steps for the development of the didactic unit and the resolution of the initial real-world problem that was introduced at the start of the teaching unit. Likewise, the use of the scaffolding questions

Table 21. A summary of the quantitative strand findings

	Lower grades	Upper grades	
Dependent variables		Structured approach	Guided approach
Expectancy of success	=	+	+
Intention	+	=	+
Enjoyableness	+	+	+
Difficulty	=	+	-
Self-efficacy	+	=	=
Usefulness	=	=	=
Relevance	=	=	=

Note: the = symbol represents no change from pretest to posttest scores; + represents an increase in the posttest scores, thus, an increase in positive attitudes; and - represents a decrease in the posttest, thus, a decrease in positive attitudes.

and strategies proposed by Erdogan & Campbell (2008), Yip (2004), and Kawalkar & Vijapurkar (2013) proved to be effective in increasing the active participation of students, and in establishing an explicit connection between the four STEM disciplines.

Therefore, it can be concluded that while this framework was valid in terms of how STEM is conceptualized and how it should be delivered (i.e., considering learning principles from socio-constructivists theories), its implementation in elementary classrooms is not without issues and raises concerns about the practicality and plausibility of such an approach to science education.

6.1.2. Does STEM-based teaching units promote students' expectancies of success and attitudes towards school science?

After the 15-weeks formative phase, the efficacy of the proposed pedagogical framework in improving elementary school students' expectancy of success and attitudes towards school science was assessed (i.e., the quantitative strand from Chapter 5). Taken together, the findings revealed that STEM-based education could be effective in improving some dimensions of the dependent variables examined, although some mixed-findings should be considered (**Table 21**).

Taken together, the STEM-based intervention unit seemed to be effective in improving 3rd and 4th students' self-efficacy, enjoyableness, and intentions to further enroll in school science studies, and 5th and 6th graders expectancies of success, enjoyableness, and intentions. These findings are consistent with the literature showing that active-teaching methodologies are more motivational for students than teacher-centered and lecture-type ones. For example, a meta-analysis revealed empirical evidence supporting the use of problem-based teaching strategies for improving students' attitudes towards science courses (Demirel & Dağyar, 2016). Similarly, enjoyableness findings are in line with previous studies using information-communication technologies (ICT) in classroom (Kara & Yesilyrut, 2008) or by using active-type of teaching strategies (Lee & Erdogan, 2007; Akınoğlu & Tandoğan, 2007).

Likewise, these findings are consistent with previous studies promoting STEM-based science education. Thus, in relation to fostering interest in further scientific studies, Shahali et al. (2017) study concluded that exposure to STEM-based approach, with emphasis on the engineering design process, has a positive impact on students' interest in science subjects and careers. Similarly, Kim et al. (2015) highlighted that STEM education, delivered using robotics improved emotional engagement variables, such as interest and enjoyment. Nonetheless, these comparisons should be interpreted with caution because STEM conceptualization, as well as the approach to integration, varied significantly between these studies and this dissertation.

Likewise, an increase in intentions to further enroll in science can be explained by the fact that affective measures such as self-efficacy and expectancies of success have been shown to be linked to engagement and, therefore, related to decision-making to pursue science education studies or careers (Eddy & Brownell, 2016; Kelly, 2016; Wang & Degol, 2013). As such, the use of inquiry and engineering design teaching strategies in the intervention unit allowed students to increase their levels of self-efficacy in school science, and presumably, their intention to further enroll was affected by this aspect, as well as by an increase in enjoyableness.

As for the lack of improvement in the usefulness and relevance of school science-dependent variables, these results are not surprising and can be explained by the lack of explicit focus on targeting these aspects during the intervention unit (Gaspard et al., 2015). Previous studies suggest that perceived relevance can be fostered through explicit, relevance-inducing tasks, such as directly communicating the utility or value information (Shechter, Durik, Miyamoto, & Harackiewicz, 2011) or self-generating reflections through writing a text about the relevance of the subject (Gaspard et al., 2015; Hulleman, Godes, Hendricks, & Harackiewicz, 2010; see Harackiewicz, Tibbetts, Canning, & Hyde, 2014 for a review of interventions aimed at promoting utility-value of STEM courses). Therefore, while during the intervention unit students had the opportunity to understand the relevance of inquiry investigations and engineering design processes for solving real-world problems, these aspects did not tackle the usefulness and relevance of science as a school subject. In other words, students could have learned about the usefulness and relevance of science, technology, engineering and mathematics as disciplines and enterprises, but they could not relate these aspects to the relevance of studying science at school; therefore, the "Why should I learn school science?" question could have remained unanswered.

As mentioned earlier, some mixed findings were obtained. More specifically, while the structured approach to inquiry investigation was effective in reducing students' perceived difficulty of school science, the guided approach, characterized by less guidance from the teacher, increase students perceived difficulty of school science. These results are in line with challenges the common assumption in science education research about the need to deliver inquiry investigations through open-minimal guidance teaching strategies, aspect that has already been warned by Kirschner et al. (2006). These findings can be explained using constructivist theories principles. Thus, Ausubel (1982) theory highlight the importance of previous concepts and ideas that the learner already possesses; similarly, Vygotsky (1981) argued that in order for the students to learn, the learning episode should be adapted to the zone of proximal development. Therefore, given that the students from this study had no previous experience with inquiry teaching methodology, it is not surprising that they were not been able to engage significantly during the unit. Thus, by reducing the amount of teacher guidance, this teaching strategy stepped outside of student's zone of proximal development and therefore did not connect with their previous ideas and experiences, which could have led to increasing theirs perceived difficulty of school science.

6.2. Significance and implications of this dissertation

The significance of this dissertation appears to be threefold. First, this dissertation has advanced in the conceptualization of STEM education and offered a research-based pedagogical framework for developing teaching units for STEM-based science education. Second, the findings of the intervention study have added to existing knowledge about the difficulties experienced by teachers when implementing STEM-based units, and about the effectiveness of such educational approaches on expectancies and attitudes. And third, this dissertation has contributed to highlighting those psychometric practices that need to be improved when developing measurement instruments and added to the Spanish literature two valid and reliable instruments that can be used to assess the effectiveness of interventions aimed at repairing the STEM pipeline.

6.2.1. Conceptualization of STEM and pedagogical transposition

As discussed previously, this dissertation proposed a conceptualization of STEM education and exemplified how STEM can be translated into classroom practices. Therefore, this dissertation advances in filling the gap for common language when STEM is addressed in educational research and practice. This conceptualization of STEM as an educational paradigm aimed at teaching science, technology, engineering and mathematics in an integrative way, as well as the pedagogical framework describe for designing STEM-based teaching units, can be used in future studies to test the effectiveness of such educational proposals that are nowadays overly promoted and too frequently misinterpreted.

This aspect has several implications for science education research. First, while this thesis recognizes the multiplicity of STEM conceptions and approaches, it highlights the need to abandon the detrimental practice of coining the term STEM when only one or two disciplines are being addressed. Martín-Páez et al., (2019), after reviewing a total of 27 studies self-identified as STEM, concluded that a common error is to associate STEM with the treatment of a single discipline, or with the treatment of two disciplines in which one of them is only used to support for the first one. Therefore, in order to facilitate progress in this line of research, projects should only be categorized as STEM when it addresses the four STEM disciplines by making explicit connections between them. In this way, for example, interventions focused on scientific content and in which technology or mathematics are used sporadically (e.g., studying pulleys through an investigation in which mathematics is used only for data collection), or interventions that use engineering design process in an isolated manner and without establishing connections with science and technology (e.g., using engineering design process to build parachutes without first addressing the scientific concept of gravity or air resistance force), should not fit within the STEM umbrella.

The second implication is related to the instruction methodologies to be used. Thus, from the formative phase of this dissertation, it is derived that active methodologies involving students in science, engineering, and computational programming practices are the cornerstone for achieving the integration of the four STEM discipline within a single unit since they allow reflecting explicitly on the similarities and differences between these disciplines.

Third, using this common conceptualization of STEM, future studies should address several critical aspects stemming from this dissertation. Although the idea of teaching and learning science in an integrated way by establishing connections with other disciplines is appealing and intuitively valid, there are many factors that should be considered before blindly promoting such an approach. More specifically, given the results of the formative phase of the main study, in-service teachers seem to not possess the necessary training to effectively and operationally implement STEM education. Likewise, although the organizational system of Spanish primary education (in which the same teacher teaches several subjects) could favor the implementation of integrated approaches, it is worth asking to what extent the design of a curriculum based on these STEM integration principles is feasible.

Therefore, if STEM is to be fully adopted in science education, the proposed pedagogical framework for STEM-based science education has implications for in-service and prospective teacher professional development. While there is much research focused on improving science teacher curricula (Capps & Crawford, 2013), there is a gap in the literature on how to train teachers with the knowledge, skills, and strategies required to implement STEM-based education, especially since engineering and computer programming practices are to be promoted. Therefore, this implication presents perhaps one of the greatest challenges for STEM education as proposed in this study, as this new educational paradigm aimed at curricula integration of four disciplines raise the need to completely revise and adapt teacher professional development plans to address the needs rooted in such novel approaches.

As can be observed, there are many questions about STEM that cannot be answered unless STEM is clearly conceptualized. In this regard, it is hoped that the conceptualization advanced in this thesis could add to reversing the current trend in science education research consisting in incorporating the STEM acronym into all projects, regardless of whether or not the integration of the four STEM disciplines is proposed.

6.2.2. Effectiveness of STEM proposals

Besides advances in the conceptualization of STEM education, this dissertation piloted the proposed framework to assess whether such an educational paradigm is effective. So far, findings suggest that STEM-based science education could improve students' expectancies of success, self-efficacy, and enjoyableness of school science, and enhance intention to further enroll in science studies. The results of this study solidify the notion that student-centered teaching strategies are more beneficial than lecture-based lessons, in terms of psychological variables such as attitudes.

Nonetheless, a novel result challenges some assumptions about teaching approaches that uses minimum teacher guidance. Thus, the finding of the guided approach increasing students' perceived difficulty of school science should be further studied under controlled conditions. If using inquiry investigations as a teaching strategy aims at improving students' attitudes, it may be the case that the common assumption that the more open, the better (i.e., open inquiry approach) is not the most beneficial. Likewise, future studies should examine how value-based strategies, aimed at improving students perceived value (i.e., usefulness and relevance) of school science could be introduced in the pedagogical framework advanced in this dissertation.

6.2.3. Psychometric properties of measurement instruments

Finally, this dissertation adds a major contribution to measurement methods in science education research. Thus, both systematic reviews reveal that the quality of instruments focused measuring attitudes towards science, and other psychological variables related to attitudes, such as motivation or self-efficacy, is of major concern. Thus, the instruments reviewed are characterized by the absence of a theoretical framework for the conceptualization of what constitutes attitudes towards science, which leads to the inclusion under the same umbrella, and without justification, of some many diverse constructs whose relevance for the study of attitudes towards science is, in the best of cases, doubtful. Likewise, the absence of empirical test-retest reliability evidence makes these instruments inappropriate for longitudinal studies or studies focused on the evaluation of educational interventions (pretest-posttest designs). Since the stability and reproducibility of results is not guaranteed, researchers evaluating educational interventions would use these instruments blindly and without any possibility of knowing with certainty to what extent the educational intervention has been effective.

Moreover, it is of major concern the misuse of psychometric test, mainly those related to conducting exploratory factor analysis, or the fact that the vast majority of instruments have been developed using items that lack an empirical basis, so they have been formulated according to the criteria of the test developers or from adapting existing instruments. Therefore, it is not clear to what extent these items reflect the ideologies of the researchers rather than accurately measuring the real attitudes of the individuals under study.

The implications of these practices are in-depth discussed in Toma (in press)[11] and Toma, Lederman, & Meneses Villagrá (under review)[12]. Therefore, only a summary of them is included in this section. Consequently, due to the poor psychometric quality of the instruments available for measuring attitudes towards science, one might wonder about the degree of confidence that can be placed in the results derived from the use of these instruments whose validity and reliability are at stake. This requires the development of measurement tools that have undergone a rigorous process of design and validation, which would help to obtain valid and reliable results that would support or refute the assumptions established and agreed upon in this line of research.

Likewise, it is necessary to consolidate a robust body of instruments that report valid and reliable results that allow comparisons between international contexts, as well as between different educational interventions focused on promoting favorable attitudes. It is worth highlighting the contribution of this dissertation to this aspect by developing two instruments whose psychometric properties have been thoroughly examined and established in three empirical studies. On the one hand, the Spanish School-Science Attitude Survey (S-SSAS) has been cross-validated for the Spanish

[11] See Appendix 3
[12] See Appendix 4

context. Using a multistage cross-cultural translation approach and multiple psychometric evaluations, empirical evidence has been advanced to consider the S-SSAS as a valid and reliable instrument for measuring elementary students' attitudes toward school science. On the other hand, two psychometric studies provided empirical evidence that supports the SUCCESS instrument as a valid and reliable measurement tool for advancing understandings about elementary school student's expectancies of success in school science. The implication of these two studies is in-depth discussed in Appendices 5, 6, and 7.

6.3. Limitations

Such and comprehensive endeavor as a doctoral thesis is not without limitations. Therefore, the findings reported in this dissertation should be interpreted considering the following research design threats: (i) short intervention time, and (ii) lack of control group.

6.3.1. Short intervention time

One of the main aspects that could have had affected the findings reported in this dissertation is the intervention total time. Given that attitudes are a fairly stable construct, a 12-hour intervention, especially when delivered in a one-week timeframe, may fall too short to discover the real impact that STEM education might have. Therefore, the lack of improvement in some attitudinal dimensions should be interpreted considering this aspect. It may be the case that the same intervention, extended over several weeks[13] would have resulted in greater improvement in students' attitudes towards school science.

Nonetheless, considering existing literature on attitudes, it may be the case that the lack of improvement reported in this dissertation may be the reflection of the outcome that STEM education would produce. Thus, considering the tripartite model of attitudes, the attitudes factors that did improve after the intervention represent the affective (i.e., perceived enjoyableness and self-efficacy in school science) and the behavioral (i.e., intentions to further enroll in school science) dimensions of the attitude construct, which are considered to be more stable, profound and difficult to change (Oppenheim, 1992) than the cognitive dimension (i.e., perceived usefulness and relevance of school science factors) that did not improve after the intervention unit.

Likewise, in the attitude literature, there are many studies that report the benefits of extremely short interventions in improving students' attitudes pertaining to the cognitive dimension. For example, Gaspard et al. (2015) concluded that a 90-min intervention was effective in improving students' relevance and usefulness beliefs about mathematics subject. Similarly, Schmid & Bogner (2017) reported on the effectiveness of a 3-hour structured inquiry lesson in improving student's self-determination in science. Kosovich, Flake, & Hulleman (2017) concluded that expectancy and utility value can decrease over the course of a single college class. Likewise, Alexander, Fives, Buehl,

[13] Since a conventional school teaches a maximum of two science classes each week, this intervention would have been extended for a total of six weeks.

& Mulhern (2002) examined the effect of a single science session on students' beliefs, with positive results. Weisgram & Bigler (2007) study revealed that a 1-hour session about gender discrimination in the scientific field led to improvement in students' self-efficacy and beliefs about the value (e.g., similar to usefulness and relevance studied in this dissertation) of science. Finally, Prokop, Tuncer, & Kvasničák (2007) improved students toward biology by conducting a one-day field trip, followed by a short research-based activity.

To conclude, while the short duration of the intervention may be considered a limitation, it may also be the case that that STEM education is not effective in improving cognitive-based attitude constructs; therefore, there may be a need for a more explicit focus on the relevance and usefulness of science classes, an aspect that will need to be studied in the future.

6.3.2. Lack of control group

Similarly to the short intervention time limitation, the research design adopted for the quantitative strand is likely to be unappropriated for causal effects conclusions. More specifically, a one-group pretest-posttest quasi-experimental design was chosen (Campbell & Stanley, 1963). Therefore, the lack of control group does not control for history type of extraneous variables and consequently could prevent from concluding that the effects were caused by the intervention and not by other aspects taking place between the pretest and the posttest data collection. Hence, it may be the case that the pre and post-test scores have differed for reasons that are unrelated to the STEM-based intervention unit. However, this design was used for ecological validity reasons. In other words, the use of a control group unrelated to the intervention (i.e., that did not assist the CRIEB school center) would have also been an unacceptable practice because this group would not be comparable to the experimental one. Therefore, the use of a non-comparable control group would have raised issues about whether the outcomes are a selection bias effect.

This dissertation intended to reduce the limitation of this design by using random assignment to the different experimental conditions. Therefore, before assisting the CRIEB center, participating schools have been randomly assigned to one of the 21 intervention weeks. Therefore, some students participated during the formative evaluation phase, while others assisted in the CRIEB during the effectiveness evaluation phase; likewise, students assisting in the effectiveness evaluation phase have been randomly assigned to the structured of guided approach to inquiry investigations. Moreover, given that the dependent variables examined (i.e., attitudes towards science) have been consistently shown to decrease over the course of the end of elementary education and start of secondary education (Akpinar et al., 2009; Ali et al., 2013; de Pro Bueno & Pérez, 2014; Denessen et al., 2015; DeWitt & Archer, 2015; George, 2006; Marbá-Tallada & Márquez, 2010; Said et al., 2016; Vázquez & Manassero, 2008), there is no valuable reason to expect that the experimental students' attitudes would have naturally improve over the course of the intervention week.

Another measure was taken to reduce the threats to internal validity; thus, the intervention unit was implemented by in-service teachers (instead of the researcher itself) that implemented the unit for 15 weeks prior to the effectiveness evaluation phase. These 15-weeks formative evaluation was crucial for establishing implementation fidelity and therefore, reducing extraneous variables related to teacher training of different styles of implementing the STEM-based intervention unit.

Notwithstanding this, it should be acknowledged that the evaluation phase of this dissertation represents only a pilot study. That being so, there is a need to further conduct studies using a natural classroom context. In addition, these future studies could benefit from the adoption of a more rigorous methodological design, such as the pretest-posttest control group or the Solomon four-group design (Campbell & Stanley, 1963).

Dissemination of research

This section includes information on all the research publications stemming from this doctoral thesis project. Likewise, the three preliminary studies that provided the background for this dissertation are included.

Indexed journal articles

- Toma, R. B., Lederman, N. G., & Meneses Villagrá, J. A. (under peer review). Still on the wrong path: A systematic review and evaluation of attitude test development practices in science education.

This article is under peer review in a Journal Citation Report (JCR) indexed journal.

- Toma, R. B., & Meneses Villagrá, J. A. (under peer review). Development and validation of the SUCCESS instrument: Towards a valid and reliable measure of expectancies of success in school science.

This article is under the second round of peer review in a Journal Citation Report (JCR) indexed journal, quartile Q3 - IF = 1.280 (2017), quartile Q2 - IF = 1.468 (2018).

- Toma, R. B. (in press). Revisión sistemática de instrumentos de actitudes hacia la ciencia (2004-2016). *Enseñanza de las Ciencias*.

This article is accepted for publication in a Journal Citation Report (JCR) indexed journal, quartile Q4 - IF = 0.672 (2017), quartile Q4 - IF = 0.814 (2018).

- Toma, R. B., & Meneses Villagrá, J. A. (2019a). Validation of the single-items Spanish-School Science Attitude Survey (S-SSAS) for elementary education. *PLoS ONE, 14*(1), e0209027. doi: 10.1371/journal.pone.0209027

This article is published in a Journal Citation Report (JCR) indexed journal, quartile Q1 - IF = 2.766 (2017), quartile Q2 - IF = 2.776 (2018).

- Toma, R. B., Greca, I. M., Orozco Gómez, M. L. (2019). Attitudes towards Science and views of Nature of Science among elementary school students in terms of gender, cultural background and grade level variables. *Research in Science & Technological Education, 37*(4), 492-515. doi:10.1080/02635143.2018.1561433

This article is published in a Journal Citation Report (JCR) indexed journal, quartile Q4 - IF = 0.513 (2017), quartile Q4 - IF = 0.875 (2018)

- Toma, R. B. (2019). Measuring elementary students' expectancies of success in school science: Psychometric evaluation of the SUCCESS instrument. *EURASIA Journal of Mathematics, Science and Technology Education, 15*(8), em1733. doi:10.29333/ejmste/104695

This article is published in a journal that was indexed in the Journal Citation Report (JCR) in 2016, quartile Q3 - IF = 0.903. Nowadays it is a SCOPUS indexed journal, quartile Q2 (2017), quartile Q3 - H index = 25 (2018).

- Toma, R. B., & Meneses Villagrá, J. A. (2019b). Preferencia por contenidos científicos de física o de biología en Educación Primaria: un análisis clúster. Revista Eureka sobre Enseñanza y Divulgación de las Ciencias, 16(1), 1104. doi:10.25267/Rev_Eureka_ensen_divulg_cienc.2019.v16.i1.1104

This article is published in a journal indexed in the Emerging Sources Citation Index (Thomson Reuters), and in SCOPUS, quartile Q2 (2017), quartile Q2 – H index = 7 (2018).

- Toma, R. B., & Greca, I. M. (2018). The effect of integrative STEM instruction on elementary students attitudes towards science. *EURASIA Journal of Mathematics, Science and Technology Education, 14*(4), 1383-1395. doi:10.29333/ejmste/83676

This article is published in a journal that was indexed in the Journal Citation Report (JCR) in 2016, quartile Q3 - IF = 0.903. Nowadays it is a SCOPUS indexed journal, quartile Q2 (2017), quartile Q3 - H index = 25 (2018).

- Toma, R. B., Greca, I. M., & Orozco Gómez, M. L. (2018). Una revisión del protocolo Draw-a-Scientist-Test (DAST). *Revista Eureka sobre Enseñanza y Divulgación de las Ciencias, 15*(3), 3104. doi: 10.25267/Rev_Eureka_ensen_divulg_cienc.2018.v15.i3.3104

This article is published in a journal indexed in the Emerging Sources Citation Index (Thomson Reuters), and in SCOPUS, quartile Q2 (2017), quartile Q2 – H index = 7 (2018).

- Toma, R. B., Greca, I. M., Meneses-Villagrá, J. A. (2017). Dificultades de maestros en formación inicial para diseñar unidades didácticas usando la metodología de indagación. *Revista Eureka sobre Enseñanza y Divulgación de las Ciencias, 14*(2), 442-457. doi: 0.25267/Rev_Eureka_ensen_divulg_cienc.2017.v14.i2.11

This article is published in a journal indexed in the Emerging Sources Citation Index (Thomson Reuters), and in SCOPUS, quartile Q2 (2017), quartile Q2 – H index = 7 (2018).

Book chapters

- Toma, R. B., & Meneses Villagrá, J. A. (2019). Preferencia por contenidos de ciencias de la naturaleza en educación primaria. In P. Membiela, Mª I. Cebreiros, & M. Vidal (Eds.), Nuevos retos en la enseñanza de las ciencias (pp. 457-462). Ourense: Educación Editora.
- Toma, R. B., & Meneses Villagrá, J. A. (2019). Validación convergente de una escala de ítem único de actitudes hacia la ciencia. In P. Membiela, Mª I. Cebreiros, & M. Vidal (Eds.), Panorama actual de la enseñanza de las ciencias (pp. 395-400). Ourense: Educación Editora.
- Toma, R. B., & Greca, I. M., & Orozco Gómez, M. L. (2019). Concepciones sobre los científicos en alumnado de educación primaria. In P. Membiela, Mª I. Cebreiros, & M. Vidal (Eds.), Panorama actual de la enseñanza de las ciencias (pp. 401-406). Ourense: Educación Editora.
- Toma, R. B., & Meneses Villagrá, J. A. (2018). Prueba piloto de un modelo STEM integrado con programación computacional. In C. Martínez Losada & S. García Barros (Eds.), *28º Encuentros de Didáctica de las Ciencias Experimentales. Iluminando el cambio educativo* (pp. 811-816). A Coruña: Universidade da Coruña Servizo de Publicacións.
- Toma, R. B., & Meneses Villagrá, J. A. (2018). Estructura latente de una escala de actitudes hacia la ciencia escolar basada en el modelo motivacional Expectativa-Valor de Eccles. In C. Martínez Losada & S. García Barros (Eds.), *28º Encuentros de Didáctica de las Ciencias Experimentales. Iluminando el cambio educativo* (pp. 953-958). A Coruña: Universidade da Coruña Servizo de Publicacións.
- Toma, R. B. (2018). Integrando la programación computacional en el enfoque STEM: un ejemplo sobre la calidad del agua. In I. M. Greca Dufranc & J. Á. Meneses Villagrá (Eds.), *Proyectos STEAM para la escuela primaria. Fundamentos y aplicaciones prácticas* (pp. 116-136). Madrid: Dextra Editorial S.L.
- Toma, R. B., & Greca, I. M. (2017). Modelo interdisciplinar de educación STEM para la etapa de educación primaria. In Mª I. Cebreiros, P. Membiela, N. Casado, & M. Vidal (Eds.), *La enseñanza de las ciencias en el actual contexto educativo* (pp. 391-395). Ourense: Educación Editora.

Conference presentations

- Toma, R. B., Lederman, N. G., Jiménez, J., & Meneses Villagrá, J. Á. (2019). Exploring students' acceptance of coding activities during integrative STEM lessons. *Paper presented at the NARST 2019 Annual International Conference*. Baltimore: USA.

- Toma, R. B., Lederman, N. G., & Meneses Villagrá, J. Á. (2019). A systematic review of science education attitude instruments psychometric properties. *Paper presented at the ESERA 2019 Annual International Conference*. Bologna: Italy.
- Toma, R. B., Lederman, N. G., & Meneses Villagrá, J. Á. (2019). Development and validation of the negative appraisals of studying school science (NASSS) scale. *Paper presented at the ESERA 2019 Annual International Conference*. Bologna: Italy.
- Toma, R. B., & Meneses Villagrá, J. A. (2019). Segregation in Science begins in Elementary Education. Paper presented at the *GIREP-ICPE-EPEC-MPTL 2019 International Conference*. Budapest: Hungary.
- Toma, R. B., & Meneses Villagrá, J. A. (2018). Integrating computer programming into STEM: an EVT analysis. *Paper presented at the GIREP-MPTL 2018 International Conference*. San Sebastian: Spain.

References

AAAS. (1993). *Benchmarks for science literacy*. New York, NY: Oxford University Press.

Abd-El-Khalick, F., Boujaoude, S., Duschl, R. A., Lederman, N. G., Mamlok-Naaman, R., Hofstein, A., … Tuan, H. (2004). Inquiry in science education: International perspectives. *Science Education*, *88*(3), 397–419.

Abd-El-Khalick, Fouad, Summers, R., Said, Z., Wang, S., & Culbertson, M. (2015). Development and large-scale validation of an instrument to assess Arabic-speaking students' attitudes toward science. *International Journal of Science Education*, *37*(16), 2637–2663. doi:10.1080/09500693.2015.1098789

AERA, APA & NCME. (2014). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association.

Aikin, W. M. (1942). *The story of the Eight-Year study: With conclusions and recommendations*. New York, NY: Harper & Brothers.

Ainley, M., Ainley, J. (2011a). A cultural perspective on the structure of student interest in science. *International Journal of Science Education*, *33*(1), 51-71. doi:10.1080/09500693.2010.518640

Ainley, M., Ainley, J. (2011b). Student engagement with science in early adolescence: The contribution of enjoyment to students' continuing interest in learning about science. *Contemporary Educational Psychology*, *36*, 4-12. doi:10.1016/j.cedpsych.2010.08.001

Ajzen, I. (1991). The theory of planned behavior. *Organizational Behavior and Human Decision Processes*, *50*(2), 179–211.

Ajzen, I., & Fishbein, M. (2005). The Influence of Attitudes on Behavior. In D. Albarracín, B. T. Johnson, & M. P. Zanna (Eds.), *The handbook of attitudes* (pp. 173-221). Mahwah, NJ, US: Lawrence Erlbaum Associates Publishers.

Akpinar, E., Yildiz, E., Tatar, N., & Ergin, Ö. (2009). Students' attitudes toward science and technology: an investigation of gender, grade level, and academic achievement. *Procedia - Social and Behavioral Sciences*, *1*(1), 2804–2808. doi:10.1016/j.sbspro.2009.01.498

Akinoglu, O., & Tandogan, O. R. (2007). The effects of problem-based active learning in science education on students' academic achievement, attitude and concept learning. *EURASIA Journal of Mathematics, Science & Technology Education*, *3*, 71-81.

Ali, M. M., Yager, R., Hacieminoglu, E., & Caliskan, I. (2013). Changes in student attitudes regarding science when taught by teachers without experiences with a model professional development program. *School Science and Mathematics*, *113*(3), 109–119. doi:10.1111/ssm.12008

Alexander, P. A., Fives, H., Buehl, M. M., & Mulhern, J. (2002). Teaching as persuasion. *Teaching and Teacher Education, 18*(7), 795-813.

Allen, M. (2017). *The SAGE encyclopedia of communication research methods*. SAGE Publications.

Andersen, L., & Ward, T. J. (2014). Expectancy-value models for the STEM persistence plans of ninth-grade, high-ability students: A comparison between black, Hispanic, and white students. *Science Education*, *98*(2), 216–242. doi:10.1002/sce.21092

Anderson, R. D. (2002). Reforming science teaching: What research says about inquiry. *Journal of Science Teacher Education, 13*, 1-12. doi:10.1023/A:1015171124982

Aragües, A., Gil Quílez, J., & de la Gándara, M. (2014). Análisis del papel de los maestros en el desarrollo de actividades de indagación en el practicum de primaria. *Didáctica de las Ciencias Experimentales y Sociales, 28*, 135-151. doi:10.7203/DCES.28.3523

Arandia, E., Zuza, K., & Guisasola, J. (2016). Actitudes y motivaciones de los estudiantes de ciencias en bachillerato y universidad hacia el aprendizaje de la Física. *Revista Eureka Sobre Enseñanza y Divulgación de Las Ciencias*, *13*(3), 558–573. doi:10.25267/rev_eureka_ensen_divulg_cienc.2016.v13.i3.04

Archer, L., Dewitt, J., Osborne, J., Dillon, J., Willis, B., & Wong, B. (2010). "Doing" science versus "being" a scientist: Examining 10/11-year-old schoolchildren's constructions of science through the lens of identity. *Science Education*, *94*(4), 617–639. doi:10.1002/sce.20399

Ausubel, D. P. (1963). *The psychology of meaningful verbal learning*. Oxford: Grune & Stratton.

Ausubel, D. P. (1968). *Educational psychology: A cognitive view*. New York: Holt, Rinehart & Winston.

Ausubel, D. P. (1982). *Psicología educativa: un punto de vista cognoscitivo*. México: Trillas.

Bandura, A., Barbaranelli, C., Caprara, G. V., & Pastorelli, C. (2001). Self-efficacy beliefs as shapers of children ' s aspirations and career. *Child Development*, *72*(1), 187–206.

Barab, S. A., & Squire, K. D. (2004). Design-Based Research: Putting a stake in the ground. *Journal of the Learning Sciences, 13*(1), 1-14.

Barak, M., & Assal, M. (2018). Robotics and STEM learning: Students' achievements in assignments according to the P3 Task Taxonomy—practice, problem-solving, and projects. *International Journal of Technology and Design Education, 28*(1), 121–144. doi:10.1007/s10798-016-9385-9

Barak, M., & Zadok, Y. (2009). Robotics projects and learning concepts in science, technology and problem-solving. *International Journal of Technology and Design Education*, *19*(3), 289–307. doi:10.1007/s10798-007-9043-3

Beane, J. A. (1997). *Curriculum integration: Designing the core of democratic education*. New York, NY: Teachers College Press.

Beck, M., Bryman, A. y Futing, L. (2004). *The Sage Encyclopedia of Social Science Research Methods*. New Delhi: SAGE Publications.

Berland, M., & Wilensky, U. (2015). Comparing virtual and physical robotics environments for supporting complex systems and computational thinking. *Journal of Science Education and Technology*, *24*(5), 628–647. doi:10.1007/s10956-015-9552-x

Blalock, C. L., Lichtenstein, M. J., Owen, S., Pruski, L., Marshall, C., & Toepperwein, M. A.

(2008). In pursuit of validity: A comprehensive review of science attitude instruments 1935-2005. *International Journal of Science Education*, *30*(7), 961–977. doi:10.1080/09500690701344578

Bogdan, R.C. & Biklen, S.K. (2007) *Qualitative Research for education: An introduction to theory and methods* (5th Edition). Boston, USA: Allyn & Bacon.

Breiner, J. M., Harkness, S. S., Johnson, C. C., & Koehler, C. M. (2012). What is STEM? A discussion about conceptions of STEM in education and partnerships. *School Science and Mathematics*, *112*(1), 3–11. doi:10.1111/j.1949-8594.2011.00109.x

Bruner, J. S. (1960). *The process of education*. Cambridge: Harvard University Press.

Bruner, J. S. (1961). The act of discovery. *Harvard Educational Review*, *31*, 21–32.

Bruner, J. S. (1966). *Toward a theory of instruction*. Cambridge: Belkapp Press.

Bruner, J. S. (1988). *Desarrollo educativo y educación*. Madrid: Morata.

Buser, T., Niederle, M., & Oosterbeek, H. (2014). Gender, competitiveness, and career choices. *The Quarterly Journal of Economics*, *129*, 1409–1447.

Bybee, R. W. (2013). *The Case for STEM education*. Arlington: NSTA press.

Bybee, R., & McCrae, B. (2011). Scientific literacy and student attitudes: perspectives from PISA 2006 science. *International Journal of Science Education, 33*(1), 7-26. doi:10.1080/09500693.2010.518644

Camus, A. (1942). Le Mythe de Sisyphe. Paris: Les Éditions Gallimard.

Capps, D. K., & Crawford, B. A. (2013). Inquiry-based professional development: What does it take to support teachers in learning about inquiry and nature of science? *International Journal of Science Education, 2013*(12), 1947-1978. doi:10.1080/09500693.2012.760209

Caleon, I. S., & Subramaniam, R. (2008). Attitudes towards science of intellectually gifted and mainstream upper primary students in Singapore. *Journal of Research in Science Teaching*, *45*(8), 940–954. doi:10.1002/tea.20250

Campbell, D. T., & Stanley, J. C. (1963). Experimental and quasi-experimental designs for research. Chicago: Rand McNally.

Carter, L. (2017). Neoliberalism and STEM education: Some australian policy discourse. *Canadian Journal of Science, Mathematics and Technology Education*, *17*(4), 247–257. doi:10.1080/14926156.2017.1380868

Cañal, P. (2007). La investigación escolar, hoy. *Alambique*: *Didáctica de las Ciencias Experimentales*, 52, 9-19.

Chachashvili-Bolotin, S., Milner-Bolotin, M., & Lissitsa, S. (2016). Examination of factors predicting secondary students' interest in tertiary STEM education. *International Journal of Science Education*, *38*(3), 366–390. doi:10.1080/09500693.2016.1143137

Chalmers, C. (2018). Robotics and computational thinking in primary school. *International Journal of Child-Computer Interaction*, *17*, 93–100. doi:10.1016/j.ijcci.2018.06.005

Chambers, D. W. (1983). Stereotypic images of the scientist: The draw-a-scientist test. *Science Education*, *67*(2), 255–265. doi:10.1002/sce.3730670213

Chin, C. (2006). Classroom interaction in science: Teacher questioning and feedback to students' responses. *International Journal of Science Education, 28*(11), 1315-1346. doi:10.1080/09500690600621100

Crawford, B. A. (2014). From inquiry to scientific practices in the science classroom. In N. G. Lederman & S. K. Abell (Eds.), *Handbook of research on science education, Volume II* (pp. 515–541). New York, NY: Routledge.

Creswell, J. W., & Clark, V. L. P. (2011). *Designing and conducting mixed methods research* (2nd edition). Sage publications.

Creswell, J. W., & Clark, V. L. P. (2018). *Designing and conducting mixed methods research* (3rd edition). Sage publications.

Cunningham, C. M. (2009). Engineering is elementary. *The Bridge, 30*(3), 11-17.

Cunningham, C. M. (2014). Precollege engineering education. In N. G. Lederman & S. K. Abell (Eds.), *Handbook of research on science education, Volume II* (pp. 747–758). New York, NY: Routledge.

Cunningham, C. M. (2018). *Engineering in elementary STEM education: Curriculum design, instruction, learning, and assessment*. New York, NY and Boston, MA: Teachers College Press and Museum of Science Driveway.

Czerniak, C. M., Weber, W. B., Sandmann, A., & Ahern, J. (1991). A literature review of science and mathematics integration. *School Science and Mathematics*, *99*(8), 421–430.

Dare, E. A., Ellis, J. A., & Roehrig, G. H. (2018). Understanding science teachers ' implementations of integrated STEM curricular units through a phenomenological multiple case study. *International Journal of STEM Education*, *5*(4), 1–19. doi:10.1186/s40594-018-0101-z

de Pro Bueno, A., & Pérez Manzano, A. (2014). Actitudes de los alumnos de Primaria y Secundaria ante la visión dicotómica de la ciencia. *Enseñanza de Las Ciencias*, *32*(3), 111–132. doi:10.5565/rev/ensciencias.1015

Demirel, M., & Dağyar, M. (2016). Effects of problem-based learning on attitude: A meta-analysis study. *Eurasia Journal of Mathematics, Science and Technology Education*, *12*(8), 2115–2137. doi:10.12973/eurasia.2016.1293a

Denessen, E., Vos, N., Hasselman, F., & Louws, M. (2015). The Relationship between Primary School Teacher and Student Attitudes towards Science and Technology. *Education Research International*, 1–7. doi:10.1155/2015/534690

Denscombe, M. (2007). *The good research guide for small-scale social research projects (3rd edition)*. Buckingham (UK): Open University Press.

Design-Based Research Collective (2003). Design-based research: An emerging paradigm for educational inquiry. *Educational Researcher, 32*(1), 5-8.

DeVellis, R. F. (2017). *Scale development. Theory and applications*. Los Angeles: SAGE.

DeWitt, J., & Archer, L. (2015). Who Aspires to a Science Career? A comparison of survey responses from primary and secondary school students. *International Journal of Science Education, 37*(13), 2170–2192. doi:10.1080/09500693.2015.1071899

Eagly, A. H., & Chaiken, S. (1995). Attitude strength, attitude structure, and resistance to change. In R. E. Petty & J. A. Krosnich (Eds.), *Ohio State University series on attitudes and persuasion, Vol. 4. Attitude strength: Antecedents and consequences* (pp. 413–432). New York: Lawrence Erlbaum Associates, Inc.

EC. (2013). *Special Eurobarometer 401: Responsible Research and Innovation (RRI), Science and Technology*. Brussels: European Commission.

Eccles, J., Adler, T., Futterman, R., Goff, S., Kaczala, C., Meece, J., & Midgley, C. (1983). Expectancies, values, and academic behaviors. In J. Spence (Ed.), *Achievement and achievement motivation* (pp. 75–146). San Francisco, CA: W. H. Freeman.

Eccles, J., Wigfield, A., Harold, R., & Blumenfeld, P. (1993). Age and gender differences in children's self and task perceptions during elementary school. *Child Development, 64*(3), 830–847.

Eddy, S. L., & Brownell, S. E. (2016). Beneath the numbers: A review of gender disparities in undergraduate education across science, technology, engineering, and math disciplines. *Physical Review Physics Education Research, 12*(2), doi:10.1103/PhysRevPhysEducRes.12.020106

English, L. D. (2016). STEM education K-12: perspectives on integration. *International Journal of STEM Education, 3*(3), 1–8. doi:10.1186/s40594-016-0036-1

Engineering, N. A. of, & Council, N. R. (2014). *STEM integration in K-12 education: Status, prospects, and an agenda for research.* Washington, DC: The National Academies Press.

Erdogan, I., & Campbell, T. (2008). Teacher questioning and interaction patterns in classrooms facilitated with differing levels of constructivist teaching practices. *International Journal of Science Education, 30*(14), 1891-1914. doi:10.1080/09500690701587028

Feenber, A. (2017). A critical theory of technology. In U. Felt, R. Fouché, C. A. Miller, & L. Smith-Doerr (Eds.), *Handbook of science and technology studies* (pp. 635–663). Cambridge MA: MIT Press.

Fernández-Cézar, R., & Pinto-Solano, N. (2017). Actitud hacia las clases de ciencias naturales en la educación primaria en España. *Revista Electrónica de Investigación Educativa, 19*(4), 112–123. doi:10.24320/redie.2017.19.4.1393

Field, A. (2009). *Discovering statistics using SPSS.* London: SAGE.

Fishbein, M., & Ajzen, I. (1975). *Belief, attitude, intention and behavior: an introduction to theory and research.* Reading, MA: Addison Wesley.

Fishbein, M., Ajzen, I. (2010). *Predicting and Changing Behavior.* New York: Psychology Press.

Flick, L., & Lederman, N. G. (2004). *Scientific inquiry and nature of science.* Dordrecht: Kluwer Academic Publishers.

Fogarty, R. (1991). Ten ways to integrate curriculum. *Educational Leadership, 49*(2), 61–65.

Fraser, B. J. (1981). *Test of Science Related Attitudes.* Melbourne: Australian Council for Educational Research.

Gardner, P. L. (1975). Attitudes to science: A review. *Studies in Science Education, 2,* 1–41.

Gaspard, H., Dicke, A. L., Flunger, B., Brisson, B. M., Hafner, I., Nagengast, B., & Trautwein, U. (2015). Fostering adolescents' value beliefs for mathematics with a relevance intervention in the classroom. *Developmental Psychology, 51*(9), 1226–1240. doi:10.1037/dev0000028

George, R. (2006). A Cross-domain analysis of change in students' attitudes toward science and attitudes about the utility of science. *International Journal of Science Education, 28*(6), 571–589. doi:10.1080/09500690500338755

Gresnigt, R., Taconis, R., van Keulen, H., Gravemeijer, K., & Baartman, L. (2014). Promoting science and technology in primary education: A review of integrated curricula. *Studies in Science Education, 50*(1), 47–84. doi:10.1080/03057267.2013.877694

Guo, J., Parker, P. D., Marsh, H. W., & Morin, A. J. S. (2015). Achievement, motivation, and educational choices: A longitudinal study of expectancy and value using a multiplicative perspective. *Developmental Psychology, 51*(8), 1163–1176. doi:10.1037/a0039440

Hacieminoglu, E. (2016). Elementary school students' attitude toward science and related variables. *International Journal of Environmental and Science Education, 11*(2), 35–52. doi:10.12973/ijese.2016.288a

Hacıeminoğlu, E., Yılmaz-Tüzün, Ö., & Ertepınar, H. (2012). Development and validation of nature of science instrument for elementary school students. *Education 3-13, 42*(3), 258–283. doi:10.1080/03004279.2012.671840

Haggis, S., & Adey, P. (1979). A review of integrated science education worldwide. *Studies in Science Education, 6*(1), 69–89. doi:10.1080/03057267908559869

Harackiewicz, J.M., Tibbetts, Y., Canning, E.A., & Hyde, J.S. (2014). Harnessing values to promote motivation in education. In S.A. Karabenick & T.C. Urden (Eds.), *Advances in motivation and achievement* (pp. 71-105). Bingley, UK: Emerald Group Publishing Limited

Herron, M. D. (1971). The nature of scientific enquiry. *The School Review, 79*(2), 171–212. doi:10.1057/9781137389831

Hillman, S. J., Zeeman, S. I., Tilburg, C. E., & List, H. E. (2016). My Attitudes Toward Science (MATS): The development of a multidimensional instrument measuring students' science attitudes. *Learning Environments Research, 19*, 203–219. doi:10.1007/s10984-016-9205-x

Hsieh, H.-F., & Shannon, S. E. (2005). Three approaches to qualitative content analysis. *Qualitative Health Research, 15*(9), 1277–1288. doi:10.1177/1049732305276687

Hulleman, C. S., Godes, O., Hendricks, B. L., & Harackiewicz, J. M. (2010). Enhancing interest and performance with a utility value intervention. *Journal of Educational Psychology, 102*(4), 880-895. doi:10.1037/a0019506

Hurley, M. M. (2001). Reviewing integrated science and mathematics: The search for evidence and definitions from new perspectives. *School Science and Mathematics, 101*(5), 259–268. doi:10.1111/j.1949-8594.2001.tb18028.x

INE. (2017). Mujeres y hombres en España. Educación. Retrieved from https://goo.gl/RC2UKd

Johnson, C. C. (2012). Conceptualizing integrated STEM education - Editorial. *School Science and Mathematics, 113*(8), 367–368.

Johnson, D. W., Johnson, R. T., Holubec, E. J., & Roy, P.(1984). *Circles of learning. Cooperation in the classroom*. Retrieved from: https://files.eric.ed.gov/fulltext/ED241516.pdf

Kara, Y., & Yesilyurt, S. (2008). Comparing the impacts of tutorial and edutainment software programs on students' achievements, misconceptions, and attitudes towards biology. *Journal of Science Education and Technology, 17*(1), 32–41. doi:10.1007/s10956-007-9077-z

Kawalkar, A., & Vijapurkar, J. (2013). Scaffolding science talk: The role of teachers' questions in the inquiry classroom. *International Journal of Science Education, 35*(11), 2004-2027. doi:10.1080/09500693.2011.604684

Keefe, B. (2010). *The perception of STEM: Analysis, issues and future directions. Entertainment and Media Communication Institute, Division of Entertainment Industries Council, Inc. (EIC).* Burbank, CA: EIC

Kelley, T. R., & Knowles, J. G. (2016). A conceptual framework for integrated STEM education. *International Journal of STEM Education, 3*(11), 1–11. doi:10.1186/s40594-016-0046-z

Kelly, A. M. (2016). Social cognitive perspective of gender disparities in undergraduate physics. *Physical Review Physics Education Research, 12*(2), doi:10.1103/PhysRevPhysEducRes.12.020116

Kennedy, J. P., Lyons, T., & Quinn, F. (2014). The continuing decline of science and mathematics enrolments in Australian high schools. *Teaching Science*, *60*(2), 34–46.

Kennedy, J. P., Quinn, F., & Taylor, N. (2016). The school science attitude survey: A new instrument for measuring attitudes towards school science. *International Journal of Research & Method in Education, 39*(4), 422–445. doi:10.1080/1743727X.2016.1160046

Keselman, A. (2003). Supporting inquiry learning by promoting normative understanding of multivariable causality. *Journal of Research in Science Teaching, 40*(9), 898–921. doi:10.1002/tea.10115

Khine, M. S. (2015). *Attitude measurements in science education: Classic and contemporary approaches*. Charlotte, NC: Information Age Publishing.

Kim, H., & Chae, D.H. (2016). The development and application of a STEAM program based on traditional Korean culture. *EURASIA Journal of Mathematics, Science and Technology Education, 12*(7), 1925-1936. doi:10.12973/eurasia.2016.1539a

Kim, C., Kim, D., Yuan, J., Hill, R. B., Doshi, P., & Thai, C. N. (2015). Robotics to promote elementary education pre-service teachers' STEM engagement, learning, and teaching. *Computers and Education, 91*, 14–31. doi:10.1016/j.compedu.2015.08.005

Kim, M. K., Lee, J. Y., Yang, H., Lee, J., Jang, J. N., & Kim, S. J. (2019). Analysis of elementary school teacer's perception of mathematics-focused STEAM education in Korea. *EURASIA Journal of Mathematics, Science and Technology Education, 15*(9), em1746. doi: 10.29333/ejmste/108482

Kind, P., Jones, K., & Barmby, P. (2007). Developing attitudes towards science measures. *International Journal of Science Education*, *29*(7), 871–893. doi:10.1080/09500690600909091

Kirschner, P. A., Sweller, J., & Clark, R. E. (2006). Why minimal guidance during instruction does not work: An analysis of the failure of constructivist, discovery, problem-based, experiential, and inquiry-based teaching. *Educational Psychologist, 31*(2), 75-86. doi:10.1207/s15326985ep4102_1

Klopfer, L. E. (1971). Evaluation of learning in science. In B. S. Bloom, J. T. Hastings, & G. F. Madaus (Eds.), *Handbook of formative and summative evaluation of student learning*. London: McGraw-Hill.

Koballa, T. R. Jr. (1988). The determinants of female junior high school students' intentions to enroll in elective physical science courses in high school: Testing the applicability of the theory of reasoned action. *Journal of Research in Science Teaching, 25*(6), 479-492.

Kosovich, J. J., Flake, J. K., & Hulleman, C. S. (2017). Short-term motivation trajectories: A parallel process model of expectancy-value. *Contemporary Educational Psychology, 49*, 130-139. doi:10.1016/j.cedpsych.2017.01.004

Lamb, R., Akmal, T., & Petrie, K. (2015). Development of a cognition-priming model describing learning in a STEM classroom. *Journal of Research in Science Teaching, 52*(3), 410–437. doi:10.1002/tea.21200

Lee, M. K., & Erdogan, I. (2007). The effect of science-technology-society teaching on students' attitudes toward science and certain aspects of creativity. *International Journal of Science Education, 29*(11), 1315–1327. doi:10.1080/09500690600972974

Leech, N.L., Barrett, K.C. and Morgan, G.A. (2005) *SPSS for Intermediate Statistics, Use and Interpretation (2nd edition)*. Mahwah: Lawrence Erlbaum Associates Inc.

Lederman, J. S. (2009). Teaching scientific inquiry: Exploration, directed, guided, and opened-endedlevels. *National Geographic Science: Best Practices and Research Base.*

Lederman, Judith S., Lederman, N. G., Bartos, S. A., Bartels, S. L., Meyer, A. A., & Schwartz, R. S. (2014). Meaningful assessment of learners' understandings about scientific inquiry - The views about scientific inquiry (VASI) questionnaire. *Journal of Research in Science Teaching*, *51*(1), 65–83. doi:10.1002/tea.21125

Lederman, Norman G. (1992). Students' and teachers' conceptions of the Nature of Science: A review of the research. *Journal of Research in Science Teaching*, *29*(4), 331–359. doi:10.1002/tea.3660290404

Ley Orgánica 1/1990, de 3 de octubre. Boletín Oficial del Estado, núm. 238, 4 de octubre de 1990.

Ley Orgánica 2/2006, de 3 de mayo. Boletín Oficial del Estado, núm. 106, 2006, 4 de mayo.

Ley Orgánica 8/2013, de 9 de diciembre. Boletín Oficial del Estado, núm. 295, 2013, 10 de diciembre.

Liberati, A., Altman, D., Tetzlaff, J., Mulrow, C., Gøtzsche, P. C., Ioannidis, J. P. A., … Moher, D. (2009). The PRISMA statement for reporting systematic reviews and meta-analyses of studies that evaluate health care interventions: explanation and elaboration. *PLoS Med 6*(7), e1000100. doi:10.1371/journal.pmed.1000100

Lou, S.-J., Chou, Y.-C., Shih, R.-C., & Chung, C.-C. (2017). A study of creativity in CaC2 steamship-derived STEM project-based learning. EURASIA Journal of Mathematics, Science and Technology Education, 13(6), 2387–2404. doi:10.12973/eurasia.2017.01231a

Marbà-Tallada, A., & Márquez, C. (2010). ¿Qué opinan los estudiantes de las clases de ciencias? Un estudio transversal de sexto de primaria a cuarto de ESO. *Enseñanza de Las Ciencias*, *28*(1), 19–30.

Martin-Hansen, L. (2002). Defining inquiry. *The Science Teacher*, *69*(2), 34–37.

Mau, W. C. (2003). Factors that influence persistence in science and engineering career aspirations. *Career Development Quarterly*, *51*(3), 234–243. doi:10.1002/j.2161-0045.2003.tb00604.x

MECD. (2016). *Datos y cifras del sistema universitario español. Curso 2015/2016.*

Meece, J. L., Wigfield, A., & Eccles, J. S. (1990). Predictors of math anxiety and its consequences for young adolescents' course enrollment intentions and performances in mathematics. *Journal of Educational Psychology*, *82*(1), 60–70.

MINEDUC (2012). *Bases Curriculares. Educación Básica*. Santiago: Ministerio de Educación.

Martín-Páez, T., Aguilera, D. Perales-Palacios, F. J., & Vílchez-González, J. M. (2019). What are we talking about when we talk about STEM education? A review of literature. *Science Education, 103*(4), 799-822. doi:10.1002/sce.21522

McKenney, S., Nieveen, N. & Van den Akker, J. (2006). Design research from a curriculum perspective. In J. Van den Akker, K. Gravemeijer, S. McKenney & N. Nieveen (Eds), *Educational design research* (62-90). London: Routledge.

Minner, D. D., Levy, A. J., & Century, J. (2010). Inquiry-based science instruction - What is it and does it matter? Results from a research synthesis years 1984 to 2002. *Journal of Research in Science*, *47*(4), 474–496. doi:10.1002/tea.20347

Mulhall, A. (2003). In the field: Notes on observation in qualitative research. *Journal of Advanced Nursing, 41*(3), 306-313. doi:10.1046/j.1365-2648.2003.02514.x

Munby, H. (1983). Thirt Studies Involving the "Scientific Attitude Inventory": What confidence can we have in this instrument? *Journal of Research in Science Teaching, 20*(2), 141–162.

Munby, H. (1997). Issues of validity in science attitude measurement. *Journal of Research in Science Teaching, 34*(4), 337–341.

NAE & NRC (2014). *STEM integration in K-12 education: Status, prospects, and an agenda for research*. Washington, DC: The National Academies Press.

NAS, NAE, & IOM. (2010). *Rising above the gathering storm, revisited: Rapidly approaching category 5*. Washington, DC: The National Academies Press.

Newell, A. D., Tharp, B. Z., Vogt, G. L., Moreno, N. P., & Zientek, L. R. (2015). Students' attitudes toward science as predictors of gains on student content knowledge: Benefits of an after-school program. *School science and mathematics, 115*(5), 216-225. doi:10.1111/ssm.12125

NGACPB. (2010). *Common Core State Standards for Mathematics*. Retrieved from http://www.corestandards.org/wp-content/uploads/Math_Standards1.pdf

NGSS Lead States. (2013). *The Next Generation Science Standards: For states, by states*. Washington, DC: The National Academies Press.

Nieveen, N. (1999). Prototyping to reach product quality. In J. van den Akker, R.M. Branch, K. Gustafson, N. Nieveen, & T. Plomp (Eds), *Design approaches and tools in education and training* (pp. 125-136). Boston: Kluwer Academic.

NRC. (2000). *Inquiry and the national science education standards: A guide for teaching and learning*. Washington, DC: The National Academies Press.

NRC. (2012). *A Framework for K-12 science education: Practices, crosscutting concepts and core ideas*. Washington, DC: The National Academies Press.

Nugent, G., Barker, B., Grandgenett, N., & Adamchuk, V. I. (2010). Impact of robotics and geospatial technology intervention on youth STEM learning and attitudes. *Journal of Research on Technology in Education*, *42*(4), 391–408.

Nunnally, J. C., & Bernstein, I. H. (1994). *Psychometric theory*. New York: McGraw-Hill.

Oppenheim, A. N. (1992). *Questionnaire design, interviewing and attitude measurement*. London: Pinter.

Osborne, J., Simon, S., & Collins, S. (2003). Attitudes towards science: A review of the literature and its implications. *International Journal of Science Education*, *25*(9), 1049–1079. doi:10.1080/0950069032000032199

Palmer, T.-A., Burke, P. F., & Aubusson, P. (2017). Why school students choose and reject science: a study of the factors that students consider when selecting subjects. *International Journal of Science Education*, *39*(6), 645–662. doi:10.1080/09500693.2017.1299949

Patton, M. Q. (2002). *Qualitative Research and Evaluation Methods*. Thousand Oaks, CA: Sage.

Pedaste, M., Mäeots, M., Siiman, L. A., de Jong, T., van Riesen, S. A. N., Kamp, E. T., … Tsourlidaki, E. (2015). Phases of inquiry-based learning: Definitions and the inquiry cycle.

Educational Research Review, *14*, 47–61. doi:10.1016/j.edurev.2015.02.003

Pedaste, M., & Sarapuu, T. (2006). Developing an effective support system for inquiry learning in a Web-based environment. *Journal of Computer Assisted Learning*, *22*(1), 47–62. doi:10.1111/j.1365-2729.2006.00159.x

Pérez-Franco, D., de Pro Bueno, A., & Pérez Manzano, A. (2018). Actitudes ambientales al final de la ESO. Un estudio diagnóstico con alumnos de Secundaria de la Región de Murcia. *Revista Eureka Sobre Enseñanza y Divulgación de Las Ciencias*, *15*(3), 3501. doi:10.25267/Rev_Eureka_ensen_divulg_cienc.2018.v15.i3.3501

Piaget, J. (1964). Part I: Cognitive development in children: Piaget development and learning. *Journal of Research in Science Teaching*, *2*(3), 176–186. doi:10.1002/tea.3660020306

Piaget, J. (1974). *Psicología de la inteligencia*. Argentina: Psique.

Pinar, W. F. (2010). The Eight-Year study. An essay review of Stories of the Eight-Year study: Reexamining secondary education in America. *Curriculum Inquiry*, *40*(2), 295–316.

Plomp, T., & Nieveen, N. (2007). *An introduction to Educational Design Research*. Enschede, The Netherlands: SLO.

Porlán, R., Martín del Pozo, R., Rivero, A., Harres, J., Azcárate, P., & Pizzato, M. (2010). El cambio del profesorado de ciencias I: Marco teórico y formativo. *Enseñanza de las Ciencias, 28*(1), 31-46.

Potvin, P., & Hasni, A. (2014). Interest, motivation and attitude towards science and technology at K-12 levels: a systematic review of 12 years of educational research. *Studies in Science Education*, *50*(August 2016), 85–129. doi:10.1080/03057267.2014.881626

Pozuelos, F., Gonzélez, G. T., & Cañal, P. L. (2010). Inquiry-based teaching: Teachers' conceptions, impediments and support. *Teaching Education, 21*(2), 131-142. doi:10.1080/10476210903494507

Prokop, P., Tuncer, G., & Kvasničák, R. (2007). Short-term effects of field programme on students' knowledge and attitude toward biology: A Slovak experience. *Journal of Science Education and Technology, 16*(3), 247-255.

Psycharis, S., & Kotzampasaki, E. (2019). The impact of a STEM inquiry game learning scenario on computational thinking and computer self-confidence. *EURASIA Journal of Mathematics, Science and Technology Education, 15*(4), em1689. doi:10.29333/ejmste/103071

Qablan, A., & DeBaz, T. (2015). Facilitating elementary science teachers' implementation of inquiry-based science teaching. *Teacher Development, 19*(1), 3-21. doi: doi:10.1080/13664530.2014.959552

Rapoport, B., & Thibout, C. (2018). Why do boys and girls make different educational choices? The influence of expected earnings and test scores. *Economics of Education Review*, *62*, 205–229. doi:10.1016/j.econedurev.2017.09.006

Romero-Ariza, M., Quesada, A., Abril, A. M., Sorensen, P., & Oliver, M. C. (2020). Highly recommended and poorly used: English and Spanish science teachers' views of inquiry-based learning (IBL) and its enactment. *EURASIA Journal of Mathematics, Science and Technology Education, 16*(1), em1793. doi:10.29333/ejmste/109658

Reeves, T.C. (2006). Design research from a technology perspective. In J. Van den Akker, K., Gravemeijer, S. McKenney, & N. Nieveen (Eds), *Educational design research* (pp. 52-66). London: Routledge.

Said, Z., Summers, R., Abd-El-Khalick, F., & Wang, S. (2016). Attitudes toward science among grades 3 through 12 arab students in Qatar: findings from a cross-sectional national study. *International Journal of Science Education*, *38*(4), 621–643. doi:10.1080/09500693.2016.1156184

Sanders, M. (2009). STEM, STEM education, STEM mania. *Technology Teacher*, *68*(4), 20–26.

Schmid, S., & Bogner, F. X. (2017). How and inquiry-based classroom lesson intervenes in science efficacy, career-orientation and self-determination. *International Journal of Science Education, 39*(17), 1-19. doi:10.1080/09500693.2017.1380332

Schwarz C. V., & Gwekwerere, Y. N. (2007) Using a guided inquiry and modeling instructional framework (EIMA) to support preservice K-8 science teaching. *Science Education 91*(1),158-186.

Shahali, E. H. M., Halim, L., Rasul, M. S., Osman, K., & Zulkifeli, M. A. (2017). STEM learning through engineering design: Impact on middle secondary students' interest towards STEM. *EURASIA Journal of Mathematics, Science and Technology Education, 13*(5), 1189–1211. doi:10.12973/eurasia.2017.00667a

Shaw, M. E., & Wright, J. M. (1968). *Scales for the measurement of attitude*. New York: McGraw-Hill.

Shavelson, R. & Towne, L. (2002). *Scientific research in education*. Washington DC: National Academic Press

Shechter, O. G., Durik, A. M., Miyamoto, Y., & Harackiewicz, J. M. (2011). The role of utility value in achievement behavior: The importance of culture. *Personality and Social Psychology Bulletin, 37*(3), 303-317. doi:10.1177/0146167210396380

Seung, E., Park, S., Jung, J. (2014). Exploring preservice elementary teachers' understanding of the essential features of inquiry-based science teaching using evidence-based reflection. *Research in Science Education, 44*(4), 507-529. doi:10.1007/s11165-013-9390-x

Schibeci, R. A. (1983). Selecting appropriate attitudinal objectives for school science. Science Education, 67(5), 595-603. doi.org/10.1002/sce.3730670508

Sawada, D., & Piburn, M. (2000) R*eformed teaching observation protocol (RTOP) training guide*. Arizona State University.

Schroeder, C. M., Scott, T. P., Tolson, H., Huang, T.-Y., & Lee, Y.-H. (2007). A meta-analysis of national research: Effects of teaching strategies on student achievement in science in the United States. *Journal of Research in Science Teaching*, *44*(10), 1436–1460. doi:10.1002/tea.20212

Schwartz, R. S., Lederman, N. G., & Crawford, B. A. (2004). Developing views of nature of science in an authentic context: An explicit approach to bridging the gap between nature of science and scientific inquiry. *Science Education*, *88*(4), 610–645. doi:10.1002/sce.10128

Scott, R., Goddard, D., Kinberg, S., Schaefer, M., Sood, A., Huffam, M., Damon, M., ... Twentieth Century Fox Home Entertainment, Inc. (2015). The Martian.

Sellami, A., El-Kassem, R. C., Al-Qassass, H. B., & Al-Rakeb, N. A. (2017). A path analysis of student interest in STEM, with specific reference to Qatari students. *EURASIA Journal of Mathematics, Science and Technology Education*, *13*(9), 6045–6067. doi:10.12973/eurasia.2017.00999a

Shute, V. J., Sun, C., & Asbell-Clarke, J. (2017). Demystifying computational thinking. *Educational Research Review*, *22*, 142–158. doi:10.1016/j.edurev.2017.09.003

Simpson, R. D., Koballa, T. R., Oliver, J. S., & Crawley, F. E. (1994). Research on the affective dimension of science learning. In D. L. Gabel (Ed.), *Handbook of research on science teaching and learning* (pp. 211–234). New York: Macmilla.

Sjøberg, S., & Schreiner, C. (2010). *The ROSE project: An overview and key findings*. Norway: University of Oslo.

STEM Task Force Report. (2014). *Innovate: A blueprint for science, technology, engineering, and mathematics in California public education*. California: Californians dedicated to education foundation. Retrieved from https://www.cde.ca.gov/pd/ca/sc/documents/innovate.pdf

Tabachnick, B. G., & Fidell, L. S. (2007). *Using multivariate statistics (5th edition)*. Boston: Pearson Education.

Toma, R. B., Greca, I. M., Orozco Gómez, M. L. (2019). Attitudes towards Science and views of Nature of Science among elementary school students in terms of gender, cultural background and grade level variables. *Research in Science & Technological Education, 37*(4), 492-515. doi: 10.1080/02635143.2018.1561433

Toma, R. B., & Greca, I. M. (2018). The effect of integrative STEM instruction on elementary students attitudes towards science. *EURASIA Journal of Mathematics, Science and Technology Education, 14*(4), 1383-1395. doi:10.29333/ejmste/83676

Toma, R. B., Greca, I. M., & Orozco Gómez, M. L. (2018). Una revisión del protocolo Draw-a-Scientist-Test (DAST). *Revista Eureka sobre Enseñanza y Divulgación de las Ciencias, 15*(3), 3104. doi:10.25267/Rev_Eureka_ensen_divulg_cienc.2018.v15.i3.3104

Toma, R. B., Greca, I. M., & Meneses-Villagrá, J. A. (2017). Dificultades de maestros en formación inicial para diseñar unidades didácticas usando la metodología de indagación. *Revista Eureka sobre Enseñanza y Divulgación de las Ciencias, 14*(2), 442-457. doi: 0.25267/Rev_Eureka_ensen_divulg_cienc.2017.v14.i2.11

Toma, R. B., Lederman, N. G., & Meneses Villagrá, J. A. (under peer review). Still on the wrong path: A systematic review and evaluation of attitude test development practices in science education.

Toma, R. B., & Meneses Villagrá, J. A. (under peer review). Development and validation of the SUCCESS instrument: Towards a valid and reliable measure of expectancies of success in school science.

Toma, R. B. (in press). Revisión sistemática de instrumentos de actitudes hacia la ciencia (2004-2016). *Enseñanza de las Ciencias*.

Toma, R. B., & Meneses Villagrá, J. A. (2019a). Validation of the single-items Spanish-School Science Attitude Survey (S-SSAS) for elementary education. *PLoS ONE, 14*(1), e0209027. doi: 10.1371/journal.pone.0209027

Toma, R. B. (2019). Measuring elementary students' expectancies of success in school science: Psychometric evaluation of the SUCCESS instrument. *EURASIA Journal of Mathematics, Science and Technology Education, 15*(8), em1733. doi:10.29333/ejmste/104695

Toma, R. B., & Meneses Villagrá, J. A. (2019b). Preferencia por contenidos científicos de física o de biología en Educación Primaria: un análisis clúster. *Revista Eureka sobre Enseñanza y Divulgación de las Ciencias, 16*(1), 1104. doi:10.25267/Rev_Eureka_ensen_divulg_cienc.2019.v16.i1.1104

Tsai, H., Chung, C., Lou, S. (2018). Construction and Development of iSTEM Learning Model. *EURASIA Journal of Mathematics, Science and Technology Education, 14*(1), 15-32. doi:10.12973/ejmste/78019

Tseng, K. H., Chang, C. C., Lou, S. J., & Chen, W. P. (2013). Attitudes towards science, technology, engineering and mathematics (STEM) in a project-based learning (PjBL) environment. *International Journal of Technology and Design Education, 23*(1), 87–102. doi:10.1007/s10798-011-9160-x

Tytler, R., & Osborne, J. (2012). Attitudes and aspirations towards Science. In B. J. Fraser, K. Tobin, & C. J. McRobbie (Eds.), *Second International Handbook of Science Education* (pp. 597–625). Dordrecht: Springer Netherlands.

Vasquez, J. A., Sneider, M., & Comer, M. (2013). *STEM lesson essentials: Grades 3-8. Integrating science, technology, mathematics and education*. Portsmouth, NH: Heinemann.

Van den Akker, J., Gravemeijer, K, McKenney, S. & Nieveen, N. (2006). Educational design research. London: Routledge.

Van den Hurk, A., Meelissen, M., & van Langen, A. (2019). Interventions in education to prevent STEM pipeline leakage. *International Journal of Science Education, 41*(2), 150-164. doi:10.1080/09500693.2018.1540897

Vázquez, A., & Manassero, M. A. (2009). La vocación científica y tecnológica: predictores actitudinales significativos. *Revista Eureka Sobre Enseñanza y Divulgación de Las Ciencias*, *6*(2), 213–231. doi:10.25267/rev_eureka_ensen_divulg_cienc.2009.v6.i2.03

Vázquez, A., & Manassero, M. A. (2008). El declive de las actitudes hacia la ciencia de los estudiantes: un indicador inquietante para la educación científica. *Revista Eureka Sobre Enseñanza y Divulgación de Las Ciencias*, *5*(3), 274–292.

Vygotsky, L. S. (1979). *El desarrollo de los procesos psicológicos superiores*. España: Grijalbo.

Vygotsky, L. S. (1981). *Pensamiento y lenguaje*. Buenos Aires: La Pléyade.

Wang, L., Bruce, C., & Hughes, H. (2011). Sociocultural theories and their application in information literacy research and education. *Australian Academic and Research Libraries*, *42*(4), 296–308. doi:10.1080/00048623.2011.10722242

Wang, M. T., & Degol, J. (2013). Motivational pathways to STEM career choices: Using expectancy value perspective to understand individual and gender differences in STEM fields. *Developmental Review, 33*(4), 304–340. doi:10.1016/j.dr.2013.08.001

Weisgram, E. S., & Bigler, R. S. (2007). Effects of learning about gender discrimination on adolescent girls' attitudes toward and interest in science. *Psychology of Women Quarterly, 31*(3), 262-269.

Wigfield, A. (2004). Motivation for reading during the early adolescent years. In D. S. Strickland & D. E. Alvermann (Eds.), *Bridging the literacy achievement gap in grades 4–12* (pp. 56–69). New York: Teachers College Press.

Wigfield, A., & Cambria, J. (2010). Students' achievement values, goal orientations, and interest: Definitions, development, and relations to achievement outcomes. *Developmental Review*, *30*(4), 1–35. doi:10.1016/j.dr.2009.12.001

Wigfield, A., & Eccles, J. S. (2000). Expectancy-value theory of motivation. *Contemporary Educational Psychology*, *25*, 68–81.

Wing, J. M. (2006). Computational thinking. *Communications of the ACM*, *49*(3), 33–35. doi:10.1145/1118178.1118215

Wing, J. M. (2008). Computational thinking and thinking about computing. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, *366*(1881), 3717–3725. doi:10.1098/rsta.2008.0118

Zhang, D., & Campbell, T. (2011). The psychometric evaluation of a three-dimension elementary science attitude survey. *Journal of Science Teacher Education*, *22*, 595–612. doi:10.1007/s10972-010-9202-3

Yip, D. Y. (2004). Questioning skills for conceptual change in science instruction. *Journal of Biological Education, 38*(2), 76-83. doi:10.1080/00219266.2004.9655905

Appendices: Under review, accepted, and published studies

Appendix 1: A pilot model of STEM education

This appendix includes the full-text of the published article about the pilot model of STEM education that was further developed in this dissertation.

Toma, R. B., & Greca, I. M. (2018). The effect of integrative STEM instruction on elementary students attitudes towards science. *EURASIA Journal of Mathematics, Science and Technology Education, 14*(4), 1383-1395. doi:10.29333/ejmste/83676

MODESTUM
OPEN ACCESS
EURASIA Journal of Mathematics, Science and Technology Education
ISSN: 1305-8223 (online) 1305-8215 (print)
2018 14(4):1383-1395
DOI: 10.29333/ejmste/83676

The Effect of Integrative STEM Instruction on Elementary Students' Attitudes toward Science

Radu Bogdan Toma [1*], Ileana M. Greca [1]

[1] Departamento de Didácticas Específicas, Universidad de Burgos, SPAIN

Received 21 September 2017 ▪ Revised 18 December 2017 ▪ Accepted 18 December 2017

ABSTRACT

An inquiry-based integrative STEM education approach was implemented in two fourth grade Elementary Education classes, in Spain, through a module on simple machines. The viability of this science education model in the official Spanish curriculum and its influence on students' attitudes towards science and learning of STEM subjects has been studied through an adapted Test of Science Related Attitude scale, achievement tests, and teachers' interviews. Students participating in the integrative STEM project reported significantly more favourable attitudes toward science than students from traditional classrooms. Although attitude scale and achievement test results seem to show that an integrative STEM education may be feasible in 4th grade of the Spanish elementary education, interviews revealed reluctance among teachers to use integrative STEM education and more directive instructions on implementing such an educational model are demanded. Implications for science education and future studies are discussed.

Keywords: integrative STEM, attitudes toward science, inquiry-based science education, TOSRA

INTRODUCTION

Numerous reports warn of failings in the educational system to provide suitable science and technological training for young people in relation to the employment needs of the 21st (e.g. EC, 2016; UNESCO, 2015). Over recent years, fewer and fewer students appear to be interested in problems of a scientific-technological nature (Osborne & Dillon, 2008). In Europe, the number of students that end their formal studies with no scientific qualification has increased (EC, 2015), and the number of enrolments in scientific courses has fallen (OECD, 2005, 2006). The reasons for this constant reduction have been previously investigated (e.g. Ali, Yager, Hacieminoglu & Caliskan, 2013; Jones, Howe & Rua, 2000; Osborne, Simon & Collins, 2003). Evidence seem to show a failure in reaching the proposal of science education reform; as Osborne and Dillon (2008) pointed out, "The challenge therefore, is to re-imagine science education (…) for the modern world and how it can meet the needs of all students." (p. 5). Apparently, students begin elementary education with a spontaneous interest in nature, however, by the end of this educational stage they perceive science to be irrelevant, boring and too difficult to learn. Thus, while students interest in science is high at 10 years old, regardless of gender (Haworth, Dale & Plomin, 2008), by the time students are 14 years old, their interest has decreased considerably (Osborne et al., 2003). Lindahl (2007), after a longitudinal study with students of 12-to-16 years old, concluded that career aspirations and interest in science became evident at 13 years old and that the probability of engaging students in science related activities at later ages was progressively more difficult.

Although these findings suggest a need for greater emphasis on science education at the elementary education stage, there is little knowledge on how to best do it. Thus, there is a need to develop science education initiatives that address the generalized development of unfavourable attitudes toward science in elementary students. For this reason, the aims of this paper are to present a proposal for integrative STEM education and to assess its impact on elementary students' attitudes toward science and to study its viability in the Spanish education system. The results may be useful in decreasing the gap in the literature related to science education efforts focused on reducing the factors that leads to students' rejection of science, especially in Spain.

✉ rtx0001@alu.ubu.es **(*Correspondence)** ✉ imgreca@ubu.es

Contribution of this paper to the literature

- This paper presents a frame of reference for the design of integrative STEM projects for the elementary level.
- The proposed model seems feasible to be used for science teaching, particularly in the Spanish education system.
- The proposed model seems to improve 4th grade students' attitudes toward science in comparison to a traditional science teaching approach.

BACKGROUND

Attitudes toward Science

Concern over students' attitudes toward science is not new (Osborne et al., 2003) and different factors involved have been investigated. Different studies have established that the environment in which the learning of science takes place has a direct relation with the attitudes that students will develop at a later stage in life (e.g. Aldridge & Fraser, 2000; Puacharearn & Fisher, 2004). Thus, unfavourable attitudes seem to be related to traditional science teaching approaches (Oh & Yager, 2004) and rote learning (Hacieminoglu, 2016). In addition to the methodological factor, other aspects seem to impact students' attitudes, such as academic results, gender, and grade level (Ali et al., 2013). In relation to the first aspect, Caleon and Subramaniam (2008) indicate that students with low performance have shown a more unfavourable attitude towards science in comparison with high achievers. In terms of gender, there are contradictory evidences. While some studies report that boys have more favourable attitudes in comparison to girls (e.g. De Pro Bueno & Pérez Manzano, 2014; Jarvis & Pell, 2005), others have noted similar attitudes towards science among both male and female students (Akpinar, Yildiz, Tatar & Ergin, 2009).

Finally, related to grade level, although traditionally the age of 14-16 years has been established as the focal-age to improve interest in the attitudes toward science, Haworth et al. (2008) suggest that those vocations are already defined at 14 years old. Research into elementary students' attitudes toward science confirmed that girls tend to report more unfavourable attitudes than boys. More specifically, girls enjoy less science lessons (Denessen, Vos, Hasselman, & Louws, 2015) and boys have more aspiration in science than girls (DeWitt & Archer, 2015). Also, it seems that as grade level increase, favourable attitudes toward science decrease drastically (Ali et al., 2013; Said, Summers, Abd-El-Khalick, & Wang, 2016).

STEM Education

In order to increase students' positive attitudes toward science and foster their aspiration for a science related career, STEM (Science, Technology, Engineering & Mathematics) proposals have gained momentum. The idea of STEM education is the conceptualization of these disciplines as a cohesive entity, the teaching of which is integrated and coordinated as they are applied to problem-solving in the real world (Sanders, 2009). Therefore, STEM education is a model that ought to promote and improve the learning of the disciplines to which the acronym refers. Nevertheless, the conceptualizations of what STEM implies often vary between authors. While some proposals focus on the renovation of each STEM subject (Bybee, 2013), others support a multidisciplinary and integrative STEM education in which the teaching of the contents of Science, Technology, Engineering, and Mathematics subjects is similar to the treatment of these disciplines in real life (Ritz & Fan, 2014).

Nevertheless, this approach is complex and there is a lack of consensus over establishing how the content should be organized, taught and evaluated, and at what educational stage its implementation would be more convenient and beneficial (Pitt, 2009). In the comprehensive review of STEM programs conducted by Heil, Pearson & Burger (2013), the authors indicate that there is an absence of empirical studies and theoretical frameworks that guide the design and implementation of STEM programs; that the clear majority of proposals are implemented in an extracurricular schedule; and that STEM proposals are generally designed and implemented in secondary school. In the Spanish educational system, Cañal's (1998, 2007) and Charrier Mellillán, Cañal & Rodrigo Vega's (2006) studies report that teachers are reluctant to implement active teaching methodologies (like inquiry-based Science teaching) in their classrooms. In general, Spanish teachers consider that inquiry teaching is a very slow teaching method and think that it is not viable; teachers consider that it will be impossible to implement the curricula objectives and contents that are necessary if contents are taught according to student's interests. Finally, teachers tend to consider that the quality of the knowledge that students achieve through an inquiry-based education can be low or very low (Cañal, 2007).

At the institutional level, several governments have emphasized the need to develop specialized STEM programmes at all educational stages, with several of them based on the inquiry teaching approach. Thus, for

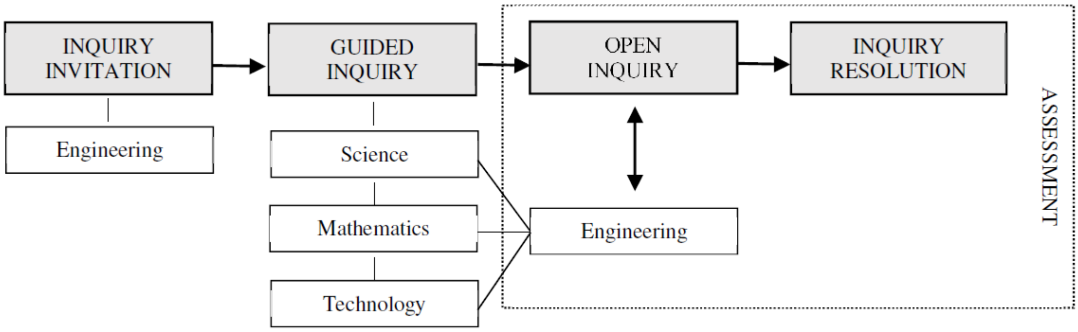


Figure 1. Integrative STEM framework for elementary education

example, Australia (Marginson, Tytler, Freeman & Roberts, 2013) and Scotland (Science and Engineering Education Advisory group, 2012) have published national reports with recommendations for the implementation of STEM education and have proposed an integrated curriculum in the final year of Secondary Education (ASEE, 2011; Ritz & Fan, 2014; Pitt, 2009; Australian Government, 2013). However, for Elementary Education there is an absence of educational projects of these characteristics (Heil et al., 2013).

THEORETICAL UNDERPINNINGS OF THE PROPOSED MODEL

In this article, we propose an integrative STEM education model for elementary school. We consider that an integrative STEM education is more pertinent for elementary school because disinterest in science among students starts in this level (Keeley, 2009), especially in girls (Abell & Lederman, 2006). Moreover, a strong positive relationship has been found between students' science-related experiences at elementary school and the choice of future studies in STEM disciplines (Tai, Qi Liu, Maltese & Fan, 2006). Regarding viability, in elementary education, teachers teach most of the subjects to the same class. So, an interdisciplinary and integrated treatment would not be a drastic change at that educational level. At least in Spain, it would be more difficult to implement an integrative STEM model in secondary education, due to the curricular organization and, even though teachers at this level possess a higher conceptual knowledge of the subject matter, all the necessary teaching and pedagogical skills to implement an integrative STEM education at that stage may be lacking.

The proposed model of integrative STEM education uses, as a main teaching methodology, an inquiry-based science teaching approach. The term "inquiry teaching" has been used to characterize good practices in both the teaching and the learning of STEM disciplines (NRC, 2011, 2012; Rocard et al. 2007), and science teachers and researchers have scattered various interpretations of the effective forms of inquiry-based science education. In our model, we understand inquiry teaching as defined by the National Research Council (NRC, 2012): as a set of activities that seek to assimilate the learning of science and the processes and strategies that scientists follow to resolve problems in real world situations. It is a strategy that seeks to facilitate self-learning through students' interactions with the objects of the environment that stimulate them, awake their curiosity, and drive the development of thoughts of a higher order and problem-solving skills, which are competencies demanded in many science curricula reforms and in the Spanish educational system (LOMCE, 2013). For us, learning science and about science requires inquiry teaching that should involve activities that include the analysis of scientific questions through the use and development of numerous skills (identification of variables related to the problem that needs to be investigated; design and realization of experiments; data interpretation; development of explanations, and the communications of results and conclusions). More specifically, our proposal is based on a coupled inquiry (Martin-Hansen, 2002) that combines a guided and an open inquiry investigation. Through this model, different aspects of Science, Mathematics, Technology and Engineering are addressed (**Figure 1**). It begins with an invitation to inquiry in which teacher select a first problem to investigate that is connected to a specific science standard or content. Next, an open-inquiry is implemented where students generate questions related to the first problem and "specific concepts can be explored in a more didactic fashion allowing students to connect their concrete experiences to abstract concepts" (Martin-Hansen, 2002, p.35).

The proposed model consists of four phases that seeks to encompass each STEM discipline. Thus, in the first phase (inquiry invitation), the teacher proposes an engineering-based real-world problem that serves as a context to teach science-related content matter. During the second phase students perform a guided inquiry in which they conduct different experiments using scientific practices, technology, and interpret data using mathematics. The third phase consists of an open inquiry during which students should discuss the results obtained in the guided inquiry and propose new research questions necessary to solve the initial problem. The fourth and final phase (inquiry resolution) requires the design or implementation of a solution which could be technological in nature. In this way, students begin to explore engineering design, linking engineering and science, as proposed in NRC (2012). **Table 1** report how STEM disciplines are emphasized during the four phases of the proposed model.

Table 1. Integrative STEM phases and its relationship with STEM disciplines

Coupled inquiry	**STEM disciplines**
INQUIRY INVITATION Science content is introduced through a real world problem	SCIENCE - ENGINEERING Real world problem related to an engineering challenge
GUIDED INQUIRY Students perform guided experiment following teacher instruction	SCIENCE Application of scientific methodologies in order to address the scientific concepts needed to solve the problem MATHEMATICS Data analysis and interpretation TECHNOLOGY Handling of devices and instruments for the design of experiments, data gathering and analysis
OPEN INQUIRY Students keep addressing the initial problem through experiments that are not guided by the teacher	SCIENCE, TECHNOLOGY, ENGINEERING, MATHEMATICS Students discuss the results obtained and they identify better ways to improve their design in order to solve the initial problem
INQUIRY RESOLUTION Solving the initial problem	ENGINEERING Students design or implement the technological device that solves the initial problem, using the scientific concepts developed previously and, in this way, linking engineering and science TECHNOLOGY Students propose possible technological applications in real world situations of the scientific concepts addressed throughout the inquiry. Students communicate their results and offer a possible resolution of the initial problem

Programme Description: The Proposed Model as Applied to Simple Machine Unit for 4th Graders

The following is an example of the application of this model on the topic of simple machines, as designed for fourth graders.

1st phase: Invitation to inquiry. The teacher introduces the following engineering-based problem: How did the Egyptians transport the stone blocks for building the pyramids without using modern machines?

2nd phase: Guided inquiry. After the discussion, the teacher indicates students how to construct three simple machines (pulleys, inclined planes and levers) using LEGO™ material. To perform the experiments, students tie a small-sized bag into each simple machine. Then, they insert small-sized coins until the block to be transported is raised. Next, students determine the force applied by the coins to move the stone block using a dynamometer. The experiment is repeated changing the size and the weight of the stone block, the fulcrum in the case of the lever, the angle of slope in the case of the inclined plane, and the type of pulley (fixed or mobile and its size). In each case, students have to predict what will happen. The experiments are conducted three times with each simple machine, in groups. When the experiments are done, students discuss the results obtained and the laws that govern each machine are deduced with teacher facilitation. It´s worth stressing that in this unit the mathematical concepts are adding, subtracting, multiplying, and dividing decimals.

3rd phase: Open inquiry. Students report their results and discuss new variables that might have an impact on the force that is necessary to move the pyramid block using simple machines, such as the ruggedness of the inclined plane surface, the length of the inclined plane, or if a simple or compound pulley is used. Afterwards, in groups, they propose new hypotheses, design and complete experiments, and formulate their conclusions.

4th phase: Solution of the initial problem. With the results of the coupled inquiry and other support materials (e.g. class textbook), students create a model with LEGO™ materials in which the route of a stone block from its origin to the hypothetical pyramid is shown, pointing out the way in which simple machines would have been used. This solution implies an engineering thinking process, where students have to apply the scientific concepts studied and give a reasoned answer.

DESIGN AND METHODS

Purpose

The purpose of this study is to implement and evaluate, through a quasi-experimental study, the viability of the proposed integrative STEM model in the elementary stage of the Spanish educational system, and to assess its impact in 4th year students' attitudes toward science. Therefore, the following research questions were proposed:

1. What effect does the proposed STEM model have on girls' and boys' attitudes toward Science?
2. Is the proposed STEM model feasible in the Spanish educational system?

It´s worth stressing that for answering this second question it is necessary to address two aspects. One is related to student's science achievement when science is being taught using a hands-on approach, and the other is related to the main objections placed by Spanish teachers for not implementing active-teaching approaches (e.g. time consuming and student's low achievement).

Participants and Procedure

The research was carried out in two elementary schools from a metropolitan area situated in northern Spain (n = 96 students; 42% girls). Two classes have formed the treatment group (n = 55) and two classes the control group (n = 41). The treatment group was selected using convenience sampling and control group was selected based on similarities with the treatment school in terms of science teaching methodology, materials used, and students' performance.

All students included in this study passed science subject during the first trimester and most students were Spanish (foreign students were less than 10% in each school). Also, both schools followed similar science teaching approaches, mainly via traditional teaching strategies using textbooks: students neither conducted experiments, nor were engaged in problem solving activities.

The content of simple machines was selected. The treatment group studied this content through the proposed integrative inquiry-based STEM model (as detailed in the previous section), and the control group through traditional science teaching methodology. Teachers from treatment group were assisted by the first author of this study during the whole implementation. Both treatment and control group simple machine unit lasted for 12 sessions (60 minutes each), as initially expected.

Measures

Adapted Test of Science Related Attitudes (TOSRA)

An attitudinal scale adapted from Fraser's (1981) TOSRA instrument was used to assess the influence of the proposed model on students' attitudes toward science. Originally, TOSRA was designed to assess seven attitudes' dimensions through a total of 70 Likert-type items. For this study, we adapted the scale to fit our specific context and to be more viable for its application with four grade elementary students. Thus, items were reduced to a total of 14, two for each dimension. All items formulated negatively were discarded because of Spanish language characteristics, in which negatively worded sentences reduced understanding. Of the remaining positive items, authors selected four items per attitude sub-dimension until reaching complete consensus. Next, items were translated into Spanish by two bilingual translators using direct and back-translation (Callegaro Borsa, Figueiredo Damásio & Ruschel Bandeira, 2012). A pilot study was performed using a 28 Likert-type scale (n = 24; 4th graders). Preliminary results indicated student's fatigue and scale administration was time demanding. Thus, items with low reliability were excluded ($\alpha<.60$) and a final scale of 14 items was developed, two per dimension. Cronbach's α of the finale scale administered in this study was .78. The adapted TOSRA assess social implication of science (i.e. It is worth spending money on science; Science can help to make the world a better place), normality of scientists (i.e. Scientists are normal people who look like anyone else; Scientists are just as friendly as other people), attitudes toward scientific inquiry, (i.e. If I want to know something about science I prefer to do an experiment instead of receiving the answer from another person; It is better to discover things by experimentation rather than asking the teacher), adoption of scientific attitudes (i.e. I am curious about the world in which I live and the things surrounding me; I find it interesting to debate and hear opinions that are different than mine), enjoyment of science lessons (i.e. Science is the most interesting subject; I would like to have more science classes each week), leisure interest in science (i.e. I would like to receive scientific materials to do experiments at home; I like to talk about science during my leisure time), and career interest in science (i.e. When I grow up: I would like to work with people who makes scientific discoveries; I would like to study something related to science). A Spanish version of the scale is included in the Appendix 1.

During the post-test, once students complemented TOSRA scale, they were asked two open-ended questions (i.e. What are your thoughts about school science subject? Can you explain your feelings and thoughts about school science subject?)

Students achievement test

An achievement test with 25 questions was prepared following teachers' criteria and standards of evaluation. Two items asked for theoretical definition and the remaining were problem solving-based items that were answered through multiple choice or true/false answer. Examples of these questions are attached in Appendix 2.

Teacher structured interview

In order to know the vision of the regular teachers of the treatment group, a structured interview was prepared to gather information about what they believed to be important factors for science teaching and learning in elementary education (i.e. What factors do you consider to be essential for science teaching and learning at elementary education based on gender variables?), what methodology, materials and activities they used to develop for science teaching (i.e. How would you describe your approach to science teaching and what materials and activities do you use most?), and their thoughts about teaching science through an integrative STEM approach (i.e. Discuss the use of integrative STEM approaches for science teaching).

Data Collection and Analysis

The implementation of the proposal and data collection were carried out during the second semester of the 2015-2016 academic years. Data for the 1st research question (i.e. What effect does the proposed STEM model have on students' attitudes toward science?) were intended to be gathered using a control group pre-post-test design. However, because of school time restriction, the control group didn't allow pre-test data collection. Thus, available data for this research question were obtained based on a comparative post-test design using the adapted TOSRA scale. The actions adopted in order to prevent threats to internal validity were related to the context. The proposal was implemented in the same context as usual for students: time, teachers, classroom organizations, and the written support (the usual textbook).

In order to categorize students' attitudes toward science, cut-off points were established with equal percentages based on the cases explored ($n = 96$), obtaining three categories describing their attitudes (favourable, indifferent and unfavourable). Next, a statistical Student t-test for each TOSRA dimensions was used to determine differences in attitudes toward science between control and treatment group students.

Data for the second research question (i.e. Is the proposed STEM model feasible in the Spanish educational system?) were collected administering to the treatment group ($n = 55$) an achievement test (pre and post implementation) and by performing structured interviews with teachers ($n = 2$; only post-implementation). For analysing the achievement test, cut-off points were established (fail: correct answers <12; below average: 13-15 correct answers; average: 16-18 correct answers; better than average: 19-21 correct answers; excellent: correct answers > 22). Descriptive statistics and paired samples t-tests were used to analyse achievement test results. Interviews were analysed through conventional content analysis, which is a technique generally used for describing a phenomenon without imposing any existing theory (Hsieh & Shannon, 2005), in this case teacher's perception of the feasibility of the proposed STEM model. To do so, interviews were transcribed and edited to allow analysis at the sentence level. Interviews transcripts from both teachers were read comprehensively several times to get an overview of their answers. Next, transcripts were read in detail and key words that capture teachers' key opinions to each interview question were highlighted and codes were derived. Continuing this process, related codes were sorted into more general categories that reflect more than one linked key opinion. To ensure that the analyses was reliable, both authors discussed the coding scheme and categories derived, and the few discrepancies were solved by consensus after in-depth analysing the transcript jointly.

RESULTS

What Effect does the Proposed STEM Model have on Girls' and Boys' Attitudes toward Science?

Student t-student test indicated that students' attitudes from the treatment group differed significantly from students in control group, as measured post-test by the adapted TOSRA scale. On average, the treatment group reported better attitudes toward science (80% favourable; 18.2% indifferent; 1.8% unfavourable) than control group (19.5% favourable; 60.9% indifferent; 19.5% unfavourable; $p < .05$). More specifically, students studying simple machines through the integrative STEM approach proposed reported more favourable attitudes in five TOSRA sub-dimensions (**Table 2**) than students studying simple machines unit through a conventional approach. Up to 83.7% of the students from the treatment group expressed positive opinions about science subject, indicating that they saw science as useful (60%) and that they were able to use imagination and creativity during experiments (73.3%). On the contrary, 80.5% of control group students reported negative feelings about science subject, considering it was too hard because of many exercises and homework (82.2%) and boring due to continually reading from the textbook (46.3%).

Table 2. TOSRA sub-dimensions' results. Comparison between treatment and control group

	SIS[1]	NS	ASI	ASA	ESL	LIS	CIS
p value	.193	.008*	.000*	.001*	.000*	.000*	.082

* Significant at a level of .05 in favour of treatment group.

[1] Social implications of science (SIS); Normality of Scientist (NS); Attitude to Scientific Inquiry (ASI); Adoption of Scientific Attitudes (ASA); Enjoyment of Science Lessons (ESL); Leisure Interest in Science (LIS); Career Interest in Science (CIS)

Thus, the proposal seems to foster more favourable attitudes towards scientists, scientific inquiry, adoption of scientific attitudes, enjoyment of science lessons, and leisure interest in science. No significantly differences regarding gender where obtained between and within each treatment or control group (p>.05).

Is the Proposed STEM Model Feasible in the Spanish Educational System?

The achievement test supports the integrative STEM model proposed in this study in fostering learning in students: all students improved their knowledge about simple machines, as shown in **Figure 2**.

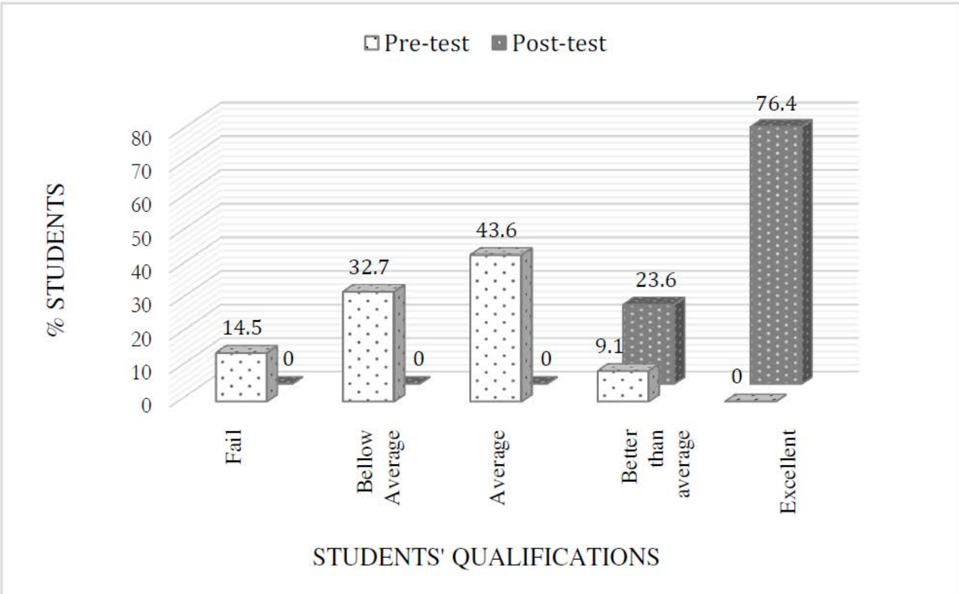


Figure 2. Treatment group achievement test results

Three questions reported significant differences between pre and post-STEM implementation (**Table 3**). No significant gender differences where obtained between boys' and girls' achievement results (p>.05).

Table 3. Achievement test questions that were correctly answered by most treatment group students

Achievement test questions	Pre-test % wrong answers		Post-test % wrong answers		P value
	Girls	Boys	Girls	Boys	
Only living beings have force	78.3	78.1	4.3	3.1	.03
The weight of an object influences the speed of its fall	95.7	87.5	34.8	15.6	.04
Don't know how to define simple machines	100	96.9	17.4	25	.02

Related to the interviews, both teachers interviewed were females with more than 20 years of teaching experience at elementary level. None of them had specific background or training in science. In relation to the important factors for science teaching, two themes emerged: practical approach, and teacher-student's interaction. Both teachers reported that science teaching needs to be practical to foster understandings of abstract concepts. One teacher discussed the importance of a positive teacher-student interaction, concluding that teachers must create a space of dialogue in which students can solve all their misunderstandings.

However, these teachers' conceptions weren't used during science lessons. Analysis of the second interview question (i.e. How would you describe your approach to science teaching and what materials and activities to you use most?) revealed that neither the teaching approach, nor the materials used to enhance students' understanding were practical in nature. Thus, both teachers mentioned the textbook as the principal material for science teaching and individual and group written activities as the most habitual practice during science lessons. Three themes emerged as reasons for these choices: educational authorities' demand, usefulness, and convenience. Relating to educational authority's demand, both teachers alluded that textbooks were imposed: "Textbooks are imposed by the school board"; "Given that parents spent a lot of money buying these books, we must use it and we cannot demand parents to buy extra material [that is necessary for performing experiments]". Also, textbooks were considered useful: "Using textbooks we can complete more easily the study program, and this is something that parents value most". Finally, regarding the convenience aspect of using textbooks as the main material, teachers said that "It is the only way to make sure that students don't make a lot of noise."; "Class management is easier [using textbooks]"; "It hardly requires a lot of preliminary preparation", "The publisher provides enough activities to accomplish the study program", and finally, that "parents want to see what their children learn in class. So, by having a textbook, they can review what have been taught each day".

In relation to the third interview question (i.e. Discuss the use of integrative STEM approaches for science teaching) both teachers argued that, although they were surprised by students' results and the classroom atmosphere during the STEM project, they wouldn't use it as the main approach for science teaching. Two themes emerged for this decision: lack of knowledge, and time demanding for the teachers. Thus, one teacher said that "I would be encouraged to use this kind of methodology if class lessons would be provided like the textbooks that we use. I need to know what to do in each moment and how to teach content". Also, both teachers agreed that "I see possibilities to use this approach as a complement to the textbook, but I can't imagine teaching science using STEM [referring to the approach introduced in this study]"; "I simply have no time for preparing experiments at home and then bring them to the classroom."; "I think that this kind of teaching is only viable in the case of having a program-guide and materials already prepared, like the textbook we use". So although the proposal seems to be feasible in Spanish educational, teachers went on being reluctant about its use.

DISCUSSION AND CONCLUSION

Results of this study reinforce the advantages of an integrative STEM education which uses inquiry teaching methodology. In general, it appeared to improve students' attitudes toward science and fostered science content learning. However, although the use of this approach has indicated to be viable in 4th grade of elementary education, teachers remained reluctant about its use and demanded more directive instructions, which requires future studies.

As shown by TOSRA results, students engaged in an integrative STEM approach reported more positive attitudes toward science than those studying science through a conventional (e.g. textbook-based) approach. These results confirm those from earlier studies stating that an active science teaching methodology improves students' attitudes (Hacieminoglu, 2016; Oh & Yager, 2004). Gender didn't have any significant influence in both treatment and control students' attitudes towards science, differing from the tendency showed in previous studies (i.e. Denessen et al. 2015; DeWitt & Archer, 2015; De Pro Bueno & Pérez Manzano, 2014; Jarvis & Pell, 2005), but being similar to those obtained by Akpinar et al. (2009). Overall, these results seem to indicate that the model proposed – using a coupled inquiry and clearly defining each of the phases needed for integrating the different STEM subjects – may be useful for the design of other STEMs teaching units for the elementary level.

This study has also shown positive results regarding the viability of an integrative STEM approach for elementary education. Thus, none of the limitations reported by Spanish teachers in Cañal's (2007) were identified. More specifically, the integrative STEM proposal was implemented without any need to make large-scale changes in the classroom distribution, subjects timing, nor school centre organization, contradicting common belief that "Inquiry teaching is a very slow teaching method. It is not viable"; "If content is taught according to students' interests it will be impossible to implement the curricula objectives and contents that are necessary" (Cañal, 2007, p. 10). Additionally, achievement test results seemed to show that students did learn the science-content proposed by following scientific and engineering practices instead of rote-learning, which clearly show that the common beliefs about the quality of knowledge achieved through active-teaching approaches being low, as reported by teachers in Cañal's (2007) study, doesn't seem to be true. However, the interviews showed that although teachers initially considered that science teaching and learning needs a practical approach, which tends to be a common conception within science teachers (Ireland, Watter, Brownlee & Lupton, 2012), they remained reluctant about the implementation of the STEM model presented in this study. Specifically, teachers reported to prefer a textbook based approach for science teaching due to convenience and logistical reasons, and they conceived the proposed STEM model as a complement to the theoretical lessons, which is similar to Spanish pre-service teachers' conceptions that consider practical activities to be used for clarifying and reinforcing the theoretical content (Hamed, Rivero & del Pozo, 2014).

The results of this study should be interpreted considering the following limitations. The application of the achievement test to both treatment and control groups could have provided more information about the relationship between scientific content learning and the teaching approach used. However, this was not an issue to be investigated in the present study. Another limitation is related to the study design. Although comparative post-test design can report useful results, its main downfall is not controlling the variables that could have impacted post-test results. Thus, it is not clear to what extent treatment students' attitudes toward science were more favourable due to the integrative STEM approach used or due to an already existing positive attitude prior to the project implementation. However, to counteract this downfall we selected schools with similar characteristics and, most importantly, that used the same teaching method and materials.

Results and limitations of this study point to more questions that are worth investigating. The extent to which this model could be adapted to other science content needs to be studied. Additionally, studies using this approach in higher or lower grades that 4th grade are encouraged. Finally, it is necessary to deepen in teacher's hesitation of shifting to a more active-based methodology for science teaching. If science teaching reforms are intended to be achieved, especially in the Spanish educational system, there seems to be an urgent need to study teacher's intention to use integrative STEM approaches and to meet their demands for more directive instructions, trainings and communities of practices (El-Hani & Greca, 2013) or mentoring programmes for the implementation of inquiries (Yoon, Kim, Kim, Young & Park, 2013) that guide science teaching at elementary level through integrative proposals. Additionally, there is a need for science textbooks to include coupled-inquiry laboratory activities. By this inclusion, teachers may be more likely to implement this particular hands-on learning strategy.

In conclusion, this study has sought to design, implement and asses and integrative STEM approach for science teaching at 4th grade elementary school level. The results of this study show that the model proposed may foster favourable attitudes toward science and offer a frame of reference for the design of integrative STEM projects.

REFERENCES

Abell, S., & Lederman, N. G. (2006). *Handbook of research on science education*. New Jersey: Lawrence Erlbaium Associates.

Akpinar, E., Yildiz, E., Tatar, N., & Ergin, Ö. (2009). Students' attitudes toward science and technology: an investigation of gender, grade level, and academic achievement. *Procedia Social and Behavioral Sciences*, 1(1), 2804-2808. https://doi.org/10.1016/j.sbspro.2009.01.498

Aldridge, J. M., & Fraser, B. J. (2000). A cross-cultural study of classroom learning environments in Australia and Taiwan. *Learning Environments Research*, 3, 101-134. https://doi.org/10.1023/A:1026599727439

Ali, M. M., Yager, R. E., Hacieminoglu, E., & Caliskan, I. (2013). Changes in student attitudes regarding science when taught by teachers without experiences with a model professional development program. *School Science and Mathematics*, 113(3), 109-119. https://doi.org/10.1111/ssm.12008

ASEE (American Society for Engineering Education). (2011). Preparing future engineers around the world. *PRISM*, 21(5), 26-34.

Australian Government. (2013). *Science, Technology, Engineering and Mathematics in the National Interest: A Strategic Approach*. Retrieved on 21 August 2017 from http://www.chiefscientist.gov.au/wp-content/uploads/STEMstrategy290713FINALweb.pdf

Bybee, R. W. (2013). *The case for STEM education: Challenges and opportunities*. National Science Teachers Association Press.

Caleon, I. S., & Subramaniam, R. (2008). Attitudes towards science of intellectually gifted and mainstream upper primary students in Singapore. *Journal of Research in Science Teaching*, 45(8), 940-954. https://doi.org/10.1002/tea.20250

Callegaro Borsa, J., Figueiredo Damásio, B., & Ruschel Bandeira, D. (2012). Cross-Cultural Adaptation and Validation of Psychological Instruments: Some Considerations. *Paidéia*, 22(53), 423–432. https://doi.org/10.1590/1982-43272253201314

Cañal, P. (1998) El origen de la investigación escolar: una alternativa se síntesis frente a la enseñanza tradicional. In G. Travé & F. J. Pozuelos (Eds) *Investigar en el aula*. Aportaciones para una didáctica innovadora. Universidad de Huelva: Servicio de Publicaciones.

Cañal, P. (2007). La investigación escolar, hoy. *Didáctica de las Ciencias Experimentales,* (52), 9-19.

Charrier Mellillán, M., Cañal, P., & Rodrigo Vega, M. (2006) Las concepciones de los estudiantes sobre la fotosíntesis y la respiración: una revisión sobre la investigación didáctica en el campo de la enseñanza y el aprendizaje de la nutrición de las plantas. *Enseñanza de las Ciencias, 24*(3), 401-410.

De Pro Bueno, A., & Pérez Manzano, A. (2014). Actitudes de los alumnos de Primaria y Secundaria ante la visión dicotómica de la Ciencia. *Enseñanza de las Ciencias*, 32(3), 111-132. https://doi.org/10.5565/rev/ensciencias.1015

Denessen, E., Vos, N., Hasselman, F., & Louws, M. (2015). The Relationship between Primary School Teacher and Student Attitudes towards Science and Technology. *Education Research International*, *2015*, 1-7. https://doi.org/10.1155/2015/534690

DeWitt, J., & Archer, L. (2015). Who Aspires to a Science Career? A comparison of survey responses from primary and secondary school students. International Journal of *Science Education*, *37*(13), 2170–2192. https://doi.org/10.1080/09500693.2015.1071899

EC (European Commision). (2016). *Horizon 2020. Monitoring Report 2014*. Luxembourg. Retrieved on 21 August 2017 from http://ec.europa.eu/research/evaluations/pdf/archive/h2020_monitoring_reports/second_h2020_annual_monitoring_report.pdf

EC (European Commission). (2015). *Science education for Responsible Citizenship*. Luxembourg: Publication Office of the European Union.

El-Hani, C. N., & Greca, I. M. (2013). ComPratica: A Virtual Community of Practice for Promoting Biology Teachers Professional Development in Brazil. *Research in Science Education*, 43, 1327-1359.

Fraser, B. J. (1981). *TOSRA: Test of science-related attitudes handbook*. Hawthorn: Australian Council for Educational Research.

Hacieminoglu, E. (2016). Elementary school students' attitude toward science and related variables. *International Journal of Environmental and Science Education*, *11*(2), 35–52. https://doi.org/10.12973/ijese.2016.288a

Hamed, S., Rivero, A., & del Pozo, R. M. (2016). El cambio en las concepciones de los futuros maestros sobre la metodología de enseñanza en un programa formativo. *Revista Eureka sobre Enseñanza y Divulgación de las Ciencias, 13*(2), 476-492.

Haworth, C. M., Dale, P., & Plomin, R. (2008). A Twin Study in the Genetic and Environmental Influences on Academic Performance in Science in nine-year-old boys and girls. *International Journal of Science Education*, *30*(8), 1003-1025. https://doi.org/10.1080/09500690701324190

Heil, D. R., Pearson, G., & Burger, S. E. (2013). Understanding Integrated STEM Education: Report on a National Study. Paper presented at *2013 ASEE Annual Conference & Exposition*, Atlanta, Georgia.

Hsieh, H-F., & Shannon, S. E. (2005). Three approaches to qualitative content analysis. *Qualitative Health Research*, *15*(9), 1277-1288. https://doi.org/10.1177/1049732305276687

Ireland, J. E., Waters, J. J., Brownless, J., & Lupton, M. (2012). Elementary teacher's conceptions of inquiry teaching: messages for teacher development. *Journal of Science Teacher Education, 23*, 159-175. https://doi.org/10.1007/s10972-011-9251-2

Jarvis, T., & Pell, A. (2005). Factors influencing elementary school children's attitudes toward science before, during, and after a visit to the UK National Space Centre. *Journal of Research in Science Teaching*, *42*(1), 53-83. https://doi.org/10.1002/tea.20045

Jones, G., Howe, A., & Rua, M. (2000). Gender differences in students' experiences, interests, and attitudes towards science and scientist. *Science Education*, *84*, 180-192. https://doi.org/10.1002/(SICI)1098-237X(200003)84:2<180::AID-SCE3>3.0.CO;2-X

Keeley, P. (2009). *Elementary Science Education in the K-12 system*. Retrieved from http://www.nsta.org/publications/news/story.aspx?id=55954

Lindahl, B. (2007). A longitudinal Study of Students' Attitudes towards Science and Choice of Career. *80th NARST International Conference*, New Orleans.

LOMCE (Ley orgánica para la mejora de la calidad educativa). (2013, 10 December). *Ley Orgánica 8/2013, 9 de diciembre*. Boletín Oficial del Estado, nº 295.

Marginson, S., Tytler, R., Freeman, B., & Roberts, K. (2013). *STEM: Country comparisions* (Final Report). Melbourne: Australian Council of Learned Academies.

Martin-Hansen, L. (2002). Defining inquiry. *The Science Teacher*, *69*(2), 34-37.

NRC (National Research Council) (2011). *Successful K-12 STEM Education: Identifying Effective Approaches in Science, Technology, Engineering, and Mathematics*. Washington, DC: The National Academies Press.

NRC (National Research Council) (2012). *A framework for K-12 science education: Practices, crosscutting concepts, and core ideas*. Washington, DC: National Academy Press.

NRC. (2011). *Successful K-12 STEM Education: Identifying Effective Approaches in Science, Technology, Engineering, and Mathematics*. Washington, DC: The National Academies Press.

NRC. (2012). *A framework for K-12 science education: Practices, crosscutting concepts, and core ideas*. Washington, DC: National Academy Press.

OECD (Organisation for Economic Co-operation and Development). (2005*). Conference on Declining Student Enrolment in Science and Technology*. Amsterdam: OECD Global Science Forum.

OECD (Organisation for Economic Co-operation and Development). (2006). *Evolution of Student Interest in Science and Technology Studies*. OECD Global Science Forum.

Oh, P. S., & Yager, R. E. (2004). Development of Constructivist Science Classrooms and Changes in Student Attitudes toward Science Learning. *Science Education International*, *15*(2), 105-113.

Osborne, J. F., Simon, S., & Collins, S. (2003). Attitudes towards Science: a review of the literature and its implications. *International Journal of Science Education*, *25*, 1049-1079. https://doi.org/10.1080/0950069032000032199

Osborne, J., & Dillon, J. (2008). *Science Education in Europe: Critical Reflections*. London: The Nuffield Foundation.

Pitt, J. (2009). Blurring the Boundaries - STEM Education and Education for Sustainable Development. *Design and Technology Education: an International Journal*, *14*(1), 37-47.

Puacharearn, P., & Fisher, D. (2004). *The effectiveness of cooperative learning integrated with constructivist teaching on improving learning environments in Thai secondary school science classrooms*. Australia: Curtin University of Technology.

Ritz, J. M., & Fan, S. C. (2014). International View of STEM education. *Proceedings PATT-28 conference*, Florida.

Rocard, M., Csermely, P., Jorde, D., Lenzen, D., Walberg-Henriksson, H., & Hemmo, V. (2007). *Science Education Now: A renewed Pedagogy for the Future of Europe*. Brussels: Directorate General for Research, Science, Economy and Society. Retrieved from https://ec.europa.eu/research/science-society/document_library/pdf_06/report-rocard-on-science-education_en.pdf

Said, Z., Summers, R., Abd-El-Khalick, F., & Wang, S. (2016). Attitudes toward science among grades 3 through 12 Arab students in Qatar: findings from a cross-sectional national study. *International Journal of Science Education*, *38*(4), 621–643. https://doi.org/10.1080/09500693.2016.1156184

Sanders, M. (2009). STEM, STEM education, STEM mania. *Technology Teacher*, *68*(4), 20-26.

Science and Engineering Education Advisory group. (2012). *Supporting Scotland's STEM education and culture*. Retrieved from http://www.gov.scot/resource/0038/00388616.pdf

Tai, R. H., Qi Liu, C., Maltese, A. V., & Fan, X. (2006). Planning early for careers in science. *Science, 312*, 1143-1145.

UNESCO (United Nations Educational, Scientific and Cultural Organization). (2015). Education for all 2000-2015. *Achievements and challenges*, France. https://doi.org/10.1126/science.1128690

Yoon, H-G., Kim, M., Kim, B. S., Joung, Y. J., & Park, Y. S. (2013). Pre-service teachers' views of inquiry teaching and their responses to teacher educators' feedback on teaching practice. *Eurasia Journal of Mathematics, Science & Technology Education*, *9*(4), 347-359. https://doi.org/10.12973/eurasia.2013.944a

APPENDIX 1

Adapted TOSRA. Spanish version

1. Merece la pena gastar dinero en la ciencia
2. Un científico se parece mucho a las demás personas
3. Prefiero resolver un problema haciendo un experimento en lugar de recibir una respuesta
4. Tengo curiosidad por las cosas que me rodean y por el mundo en el que vivo
5. Conocimiento del Medio es la asignatura más interesante
6. Me gustaría recibir materiales científicos para poder hacer experimentos en casa
7. Cuando sea mayor, me gustaría trabajar con personas que realizan descubrimientos científicos
8. La ciencia puede ayudar a que el mundo sea un lugar mejor
9. Los científicos son igual de simpáticos que las demás personas
10. Es mejor descubrir la respuesta mediante un experimento antes que preguntar al profesor
11. Me gusta escuchar a las personas que tienen una opinión diferente a la mía
12. Me gustaría tener más horas de Conocimiento del Medio a la semana.
13. Me gusta hablar sobre la ciencia fuera de clase.
14. Cuando sea mayor, quiero estudiar algo que tenga que ver con la ciencia.

APPENDIX 2

Achievement test. Sample questions

Theoretical questions:

Define what you think is a simple machine. Give an example.

Define what you think is a compound machine. Give an example.

Multiple choice questions:

- Two balls (one black, one white) are the same size. The black ball weighs 10 kg and the white ball weighs 5 kg. Both balls are thrown from a building at the same moment and from the same height. Which of the two balls reach the ground first? Select the correct option and explain your answer.

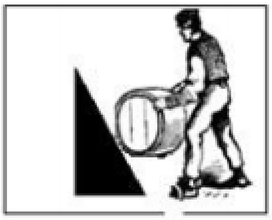
a) Both balls hit the ground at the same time.

b) The black ball arrives first.

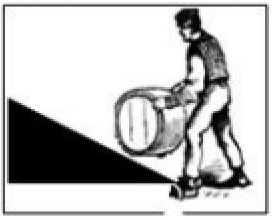
c) The white ball arrives first.

- Consider the following pictures. Sandro must load a truck with a heavy barrel. For this, he can use ramp A or ramp B. Indicate whether the following statements are true (T) or false (F). Explain the answers you think are false.

Ramp A



Ramp B



a) In both the ramp A and the ramp B, Sandro must apply the same force to transport the barrel.

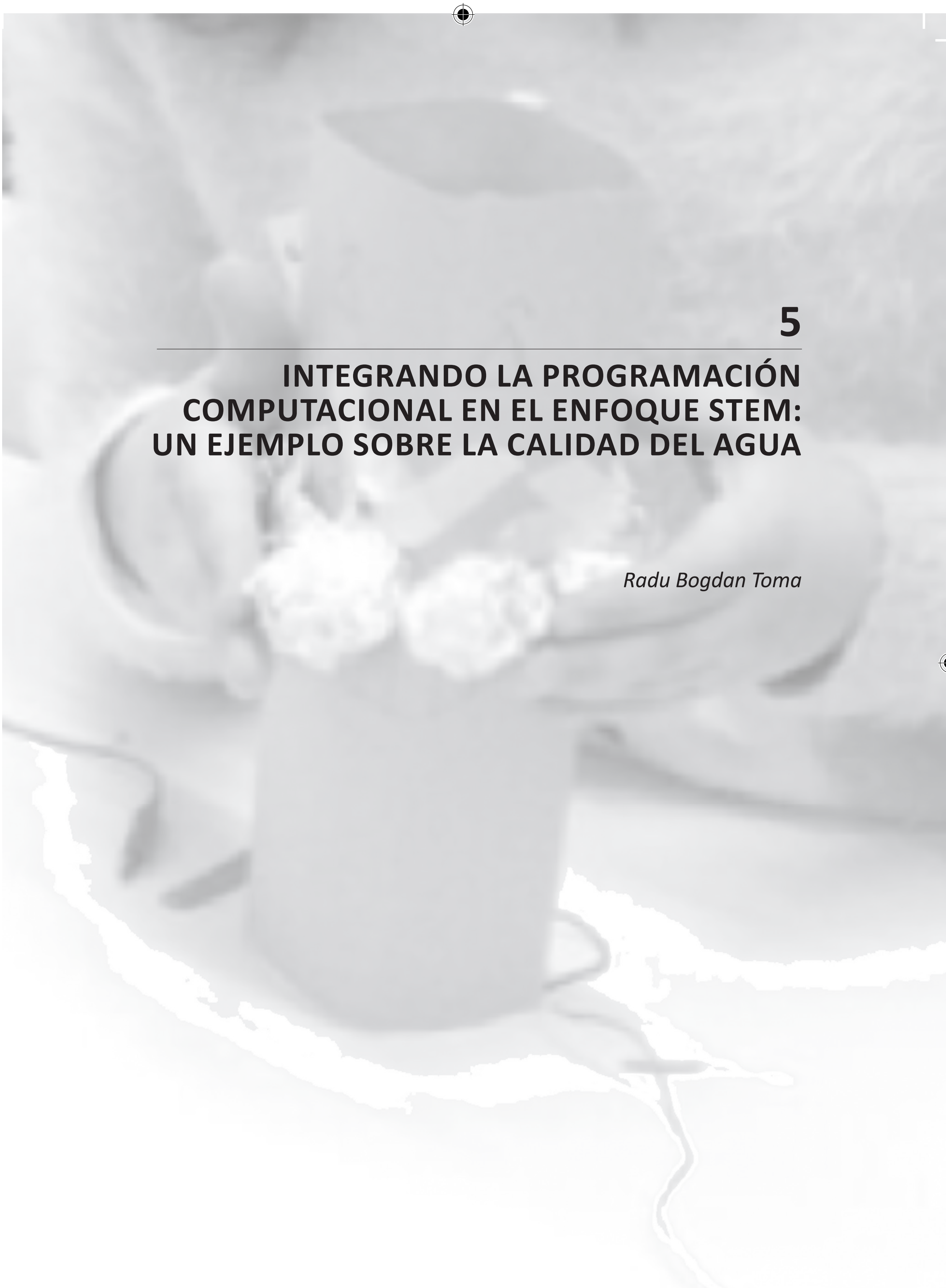
b) If Sandro chose ramp B, he will apply less force to transport the barrel in comparison to using ramp A.

c) Sandro must not use any ramp. He'd better pick up the barrel and transport it without any ramp.

Appendix 2: The intervention unit

This appendix includes the full-text of the published book chapter about the intervention unit implemented in this dissertation.

Toma, R. B. (2018). Integrando la programación computacional en el enfoque STEM: un ejemplo sobre la calidad del agua. In I. M. Greca Dufranc & J. Á. Meneses Villagrá (Eds.), *Proyectos STEAM para la escuela primaria. Fundamentos y aplicaciones prácticas* (pp. 116-136). Madrid: Dextra Editorial S.L.



5

INTEGRANDO LA PROGRAMACIÓN COMPUTACIONAL EN EL ENFOQUE STEM: UN EJEMPLO SOBRE LA CALIDAD DEL AGUA

Radu Bogdan Toma

5.1. Introducción

En este trabajo se propone un modelo de educación STEM interdisciplinar que integra los contenidos y procedimientos científicos, ingenieriles, matemáticos y los propios del pensamiento computacional a través de dos metodologías de enseñanza y aprendizaje: la indagación científica y el diseño ingenieril en el contexto escolar. Posteriormente, se ejemplifica la implementación de este modelo en el aula de 5º y 6º curso de Educación Primaria a través de una unidad didáctica (UD) que versa sobre los parámetros fisicoquímicos de la calidad del agua y las técnicas físicas de separación de mezclas.

5.2. Planteamiento STEAM

Se propone un modelo de Educación STEM (figura 5.1) que concibe la integración de las cuatro disciplinas a través de diferentes fases que conforman la UD (Toma y Meneses, 2018a).

Al comienzo de la UD, se plantea a los estudiantes un problema de relevancia social y educacional, basado en los contenidos curriculares de la asignatura de Ciencias de la Naturaleza. Posteriormente, se adopta un modelo de indagación acoplada (Martin-Hansen, 2002) para abordar los contenidos objeto de estudio de la asignatura de Ciencias, claves para la resolución del problema inicial, y así abordar la **S** (*Science*) del marco de educación STEM. La indagación acoplada, en el ejemplo que presentamos, consiste en la realización de dos indagaciones: la primera es muy asistida por el profesor, que es quien plantea las preguntas de investigación, guía la formulación de las hipótesis, diseña los experimentos y facilita instrucción para su realización; en la segunda indagación, el alumnado cobra autonomía y el profesor disminuye la cantidad de instrucciones guiadas, intentando facilitar un planteamiento más autónomo por parte de los estudiantes respecto a las preguntas de investigación, la emisión de hipótesis y el diseño y realización de los experimentos. Las fases propuestas y desarrolladas en la indagación científica se recogen en la figura 5.2

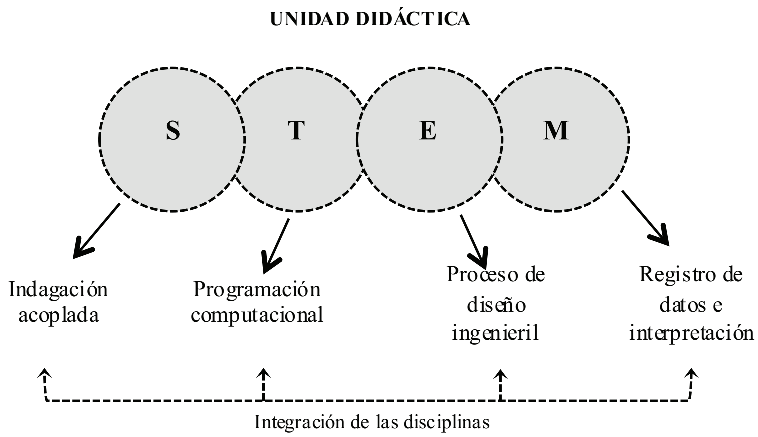


Figura 5.1. *Modelo propuesto de educación STEM integrada con programación computacional.*

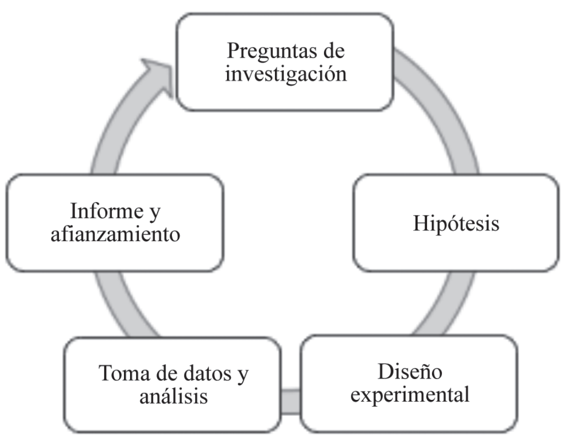


Figura 5.2. *Fases del método de iindagación científica.*

Finalizado el proceso de indagación acoplada, los estudiantes emplean el método de diseño ingenieril (Bybee, 2011) aplicando el conocimiento adquirido a través de las indagaciones anteriores, dando respuesta o solución al problema ingenieril planteado. Con ello se pretende trabajar la **E** (engineering) del marco STEM. Las fases del método ingenieril, muy similar al método de indagación científica, se recogen en la figura 5.3. A diferencia de una indagación en la que los estudiantes pretenden avanzar en el conocimiento teórico sobre una situación problemática, mediante el método de diseño ingenieril los estudiantes deben dibujar, construir y probar un prototipo de artefacto, aplicando el conocimiento teórico adquirido para resolver la situación problemática inicial. En estos dos procesos, de desarrollo de conocimiento (indagación científica) y elaboración y evaluación de un prototipo (diseño ingenieril) interviene las matemáticas (**M** del marco STEM), mediante la interpretación de los resultados; y la tecnología (**T** del modelo STEM), mediante la programación computacional de una placa microcontroladora que sirva como herramienta para la toma y registro de datos y la manipulación de diferentes materiales de laboratorio para la recogida de datos.

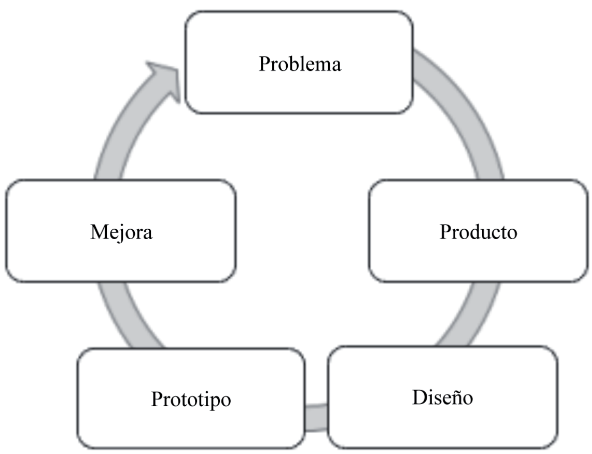


Figura 5.3. *Fases del método de diseño ingenieril.*

5.3. Contenidos y objetivos de la propuesta de enseñanza

La propuesta que se recoge en este capítulo aborda los principales contenidos establecidos internacionalmente como relevante para la educación científica. Más concretamente, se incide en las ideas fundamentales doce y trece de Harlen (2015) sobre la ciencia, pretendiendo generar comprensión por parte de los estudiantes en los siguientes aspectos:

- Las explicaciones, teorías y modelos científicos son aquellos que mejor reflejan las evidencias existentes en un determinado momento o tiempo (p. 31).

- Una teoría científica o modelo que representa relaciones entre variables de un fenómeno natural debe estar sujeta a las observaciones existentes en un determinado momento y dar lugar a predicciones que puedan ser probadas.
- Toda teoría y modelo es provisional y está sujeta a revisión a la luz de nuevas evidencias.
- Las explicaciones científicas se generan a partir de una investigación sistemática que supone la colección de datos mediante la observación o medida de los objetivos o fenómenos objeto de estudio, o el empleo de datos de otras fuentes.
- La efectividad de las explicaciones científicas depende de los datos recolectados y están guiadas por alguna otra teoría o hipótesis de partida sobre lo que podría estar sucediendo.

– El conocimiento producido por la ciencia es empleado en ingeniería y en tecnología para crear productos nuevos que satisfagan las necesidades de las personas (p. 32).

- El uso de ideas científicas en la tecnología ha dado lugar a cambios considerables en muchos aspectos de la actividad humana. Los avances tecnológicos permiten la realización de futuras investigaciones, y a su vez, esto mejora la comprensión del mundo natural.
- La ciencia, la ingeniería y la tecnología están estrechamente interrelacionadas. La aplicación de la ciencia en la fabricación de nuevos materiales es un ejemplo de cómo el conocimiento científico ha liderado los avances en la tecnología y ha proporcionado a los ingenieros una mayor variedad en el diseño de construcciones. Al mismo tiempo, los avances tecnológicos han contribuido al desarrollo científico mediante la mejora de los instrumentos de observación y medición, la automatización de procesos que, de otro modo, podrían resultar demasiado peligrosos o llevar demasiado tiempo, y en particular, mediante el suministro de ordenadores.
- En algunas áreas de la actividad humana la tecnología está más avanzada que las ideas científicas, sin embargo, en otras, las ideas científicas preceden a la tecnología.
- Las tecnologías han sido creadas por las personas para proveerse de cosas que necesitan o pueden usar, como la comida, herramientas, ropa, lugar para vivir o formas de comunicarse. Todo lo que nos rodea son ejemplos de cómo los materiales han sido modificados para cumplir un determinado propósito.
- Las tecnologías se desarrollan utilizando la ingeniería, lo que implica la identificación de problemas y el uso de ideas de la ciencia para diseñar y desarrollar la mejor solución posible. Siempre existen diferentes maneras de abordar los problemas, por lo que hay que probar varias posibilidades. Para decidir cuál es la mejor solución es necesario tener claro cuál es el resultado que se pretende obtener y, por lo tanto, cómo se ha de juzgar el éxito.

- Diseñar una solución a un problema generalmente implica hacer un dibujo o modelo. Los modelos físicos, matemáticos o informáticos permiten probar el efecto de los cambios en los materiales o el diseño y mejorar la solución. En general, hay muchos factores a considerar en la optimización de una solución, como el coste.

Concretando, para esta UD se abordan los objetivos más específicos y contenidos conceptuales y procedimentales que se señalan en las tablas 5.1 y 5.2. La primera recoge los aspectos tratados en cada disciplina STEM.

Tabla 5.1. *Aspectos abordados en cada disciplina STEM*

(S) Ciencia	(T) Tecnología	(E) Ingeniería	(M) Matemática
Parámetros fisicoquímicos de la calidad del agua. Procedimientos y competencias propias de la indagación escolar.	Empleo de instrumentos de medida y recoleccción de datos. Programación computacional de una placa microcontroladora.	Procedimientos y competencias propias del método de diseño ingenieril.	Unidades de medida. Gráficos Tablas.

Tabla 5.2. *Competencias propias de la indagación escolar y del método de diseño ingenieril en relación con las asignaturas Ciencias de la Naturaleza y Matemáticas*

Ciencias de la Naturaleza	Matemáticas
– Diseñar y aplicar experiencias sencillas para determinar la calidad del agua, planteando problemas, enunciando hipótesis, seleccionando el material necesario, realizando y extrayendo conclusiones, y comunicando los resultados. – Explicar los parámetros fisicoquímicos de la calidad del agua – Diseñar y aplicar experiencias sencillas para filtrar el agua planteando problemas, enunciando hipótesis, seleccionando el material necesario, realizando y extrayendo conclusiones, y comunicando los resultados. – Explicar procedimientos físicos de separación de mezclas. – Reflexionar y extraer conclusiones sobre el trabajo realizado. – Comunicar oralmente los resultados obtenidos en los experimentos. – Hacer un uso adecuado de las TIC como recurso de aprendizaje. – Identificar y describir algunos avances de la ciencia que mejoran la salud (depuración y potabilización del agua, etc.). – Elaborar un informe como técnica para el registro de un plan de trabajo, comunicando de forma escrita las conclusiones. – Efectuar búsquedas guiadas de información en la red.	– Interpretar los resultados obtenidos. – Leer, escribir e interpretar los números decimales de hasta dos cifras decimales. – Identificar y ordenar las unidades del sistema métrico decimal: longitud y volumen. – Realizar mediciones con instrumentos. – Interpretar datos a partir de tablas y gráficas sencillas. – Plasmar los resultados en tablas y gráficas sencillas.

La tabla 5.2 muestra con mayor detalle las competencias propias de la indagación científica escolar y del método de diseño ingenieril que se abordan, en relación con las asignaturas Ciencias de la Naturaleza y Matemáticas de la Educación Primaria.

5.4. Secuenciación de enseñanza

5.4.1. Indagación 1 sobre la calidad del agua

Para que los estudiantes comprendan los parámetros que determinan la calidad del agua, se presenta una situación problemática que requiere de la aplicación del método de indagación científica para su resolución. Durante el proceso de indagación, se introducen algunos de los parámetros fisicoquímicos de la calidad del agua: i) turbidez, ii) sólidos en suspensión, iii) temperatura, iv) nitratos, v) cloro, y vi) pH. Se averiguan los niveles óptimos de cada parámetro realizando una búsqueda de información en páginas especializadas de internet y se programa una placa microcontroladora como sensor de temperatura para la toma de datos. Posteriormente, se emplean diferentes muestras de agua para determinar la calidad de cada una y proponer una solución al problema inicial planteado.

SESIÓN 1: *FASE 1 (preguntas de investigación) y 2 (hipótesis) de la indagación científica*

Se comienza la UD abordando la primera y segunda fase del método de indagación científica, introduciendo la siguiente situación problemática a través del tráiler de la película *Marte* (disponible en https://goo.gl/T4Gg2D):

> *En una misión a Marte, la tripulación se ve obligada a abandonar el planeta rojo ante una peligrosa tormenta de arena. Durante la evacuación, el astronauta Mark queda atrapado en el planeta, solo y sin muchos recursos para sobrevivir. El astronauta Mark ha encontrado varios lugares de Marte con agua, pero no sabe si son aptas para consumo humano. Debe recurrir a la ciencia y a la ingeniería para poder descubrirlo y así sobrevivir hasta que sus compañeros consigan rescatarle.*

A partir de esta situación problemática, se inicia un diálogo y debate con los estudiantes (¿A qué problema nos enfrentamos?, ¿Qué retos debemos superar?), finalizando con la concreción de la siguiente pregunta de investigación: ¿Qué fuente de agua tiene más calidad? Mediante esta pregunta de investigación se pretende estudiar algunos parámetros fisicoquímicos de la calidad del agua. Posteriormente, los estudiantes observan tres muestras de agua, que previamente habrán sido preparadas por el docente (Figura 5.4). Las muestras de agua están compuestas por:

- Muestra A: 450 ml de agua a 60-70 ºC, 100 g de arena, 10 hojas y ramitas de árbol, una pastilla de nitratos, una pastilla de cloro y 50 ml de vinagre.
- Muestra B: 400ml de agua a más de 80 ºC, nada de arena, ninguna hoja ni ramita de árbol, una pastilla de nitratos, una pastilla de cloro y 100 ml de vinagre.
- Muestra C: 500 ml de agua a 20-25 ºC, 100 g de arena, 10 hojas y ramitas de árbol, ninguna pastilla de nitratos ni de cloro y tampoco contiene vinagre.

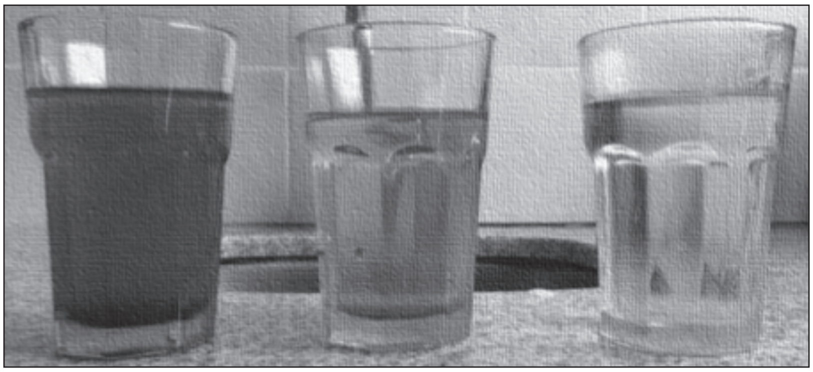


Figura 5.4. *Muestras de agua A, B y C.*

Los estudiantes observan estas muestras de agua y plantean sus hipótesis, justificando su respuesta ofreciendo argumentos. Algunos ejemplos de hipótesis planteadas por los estudiantes son:

- La muestra de agua A tiene más calidad que la B porque es más transparente.
- La muestra de agua B tiene más calidad porque no tiene cosas flotando.
- La muestra de agua C tiene menos calidad porque está muy oscura y sucia.
- La sesión finaliza con una puesta en común de las hipótesis formuladas por cada grupo de trabajo.

SESIÓN 2: *FASE 3 (diseño experimental) de la indagación científica*

En la segunda sesión, se introducen los parámetros fisicoquímicos que determinan la calidad del agua. Para ello, el maestro facilita a los estudiantes varios códigos QR que contienen información acerca de cuáles son los parámetros fisicoquímicos de la calidad del agua, con qué instrumentos y procedimientos se miden y cuáles son los rangos a los que se debe encontrar cada parámetro fisicoquímico para que el agua pueda considerarse apta para consumo humano.

Utilizando tabletas electrónicas, los estudiantes escanean los códigos QR, extraen y resumen la información necesaria, y completan el cuadro pertinente (figura 5.5 y tabla 5.3). Los parámetros que se analizan en esta propuesta didáctica son: la temperatura, los niveles de nitratos, el cloro, los sólidos en suspensión, la turbidez y el pH del agua. Para la recogida de datos, se emplean diferentes kits de medidas para determinar los niveles de nitratos, cantidad de cloro y pH del agua, se utiliza el disco *Secchi* para determinar la turbidez y los niveles de suspensión del agua, y se emplea una placa microcontroladora y un sensor para medir la temperatura del agua.

Figura 5.5. *Estudiantes completando la tabla 5.3.*

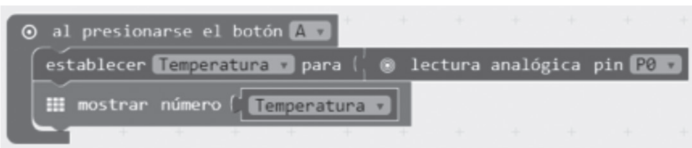
Tabla 5.3. *Tabla a rellenar tras la búsqueda de información*

Parámetros de la calidad del agua	Material para medir	Valores adecuados

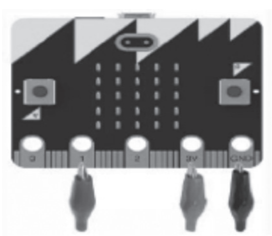
Una placa microcontroladora idónea para la etapa de Educación Primaria se denomina *BBC Micro bit*. Se trata de un microordenador que se programa empleando un código visual por bloques. Utilizando diferentes sensores, la placa microcontroladora puede ser programada para convertirse en una herramienta para medir diferentes parámetros, como por ejemplo la temperatura. Siguiendo las instrucciones facilitadas por el docente (Cuadro 5.1), los estudiantes programan la placa microcontroladora para que sirva de instrumento de medida de la temperatura del agua. La sesión finaliza con una puesta en común y discusión de la información recogida a partir de los códigos QR.

Cuadro 5.1. *Programación del microcontrolador para medir temperatura*

Paso 1: accede al enlace https://makecode.microbit.org/
Pase 2: selecciona el bloque de entrada *al presionar el botón A* y colócalo en la ventana de programación.
Paso 3: crea una nueva variable denominada *temperatura*. Para ello, abre la categoría variable, selecciona el bloque *ejecuta una variable*, e introduce el nombre de la nueva variable.
Paso 4: arrastra el bloque de variable *establecer objeto para* a la ventana de programación, conectándolo al bloque de entrada existente.
Paso 5: selecciona la variable *temperatura* pulsando encima del bloque *establecer objeto para*.
Paso 6: para registrar la temperatura, selecciona el bloque *lectura analógica pin 1* y conéctala al bloque *establecer temperatura para* existente en la ventana de programación.
Paso 7: Selecciona el bloque básico *mostrar número* y conéctalo a continuación de los bloques existentes en la ventana de programación.
Paso 8: selecciona la variable *temperatura* y conéctala al bloque situado en la ventana de programación llamado *mostrar número*.



Paso 9: Conecta el sensor de temperatura al micro:bit. El cable negro debe conectarse en *GND*, el cable rojo en *3V* y el cable azul en *1*.
Paso 10: Introduce el sensor de temperatura en el agua y pulsa el *botón A* del micro:bit. En la pantalla led aparecerá la temperatura registrada.



SESIÓN 3: *FASE 4 (toma de datos y análisis) de la indagación científica*

Se dedica la tercera sesión para la toma de datos, empleando el siguiente diseño experimental:

Variables

- Variable dependiente: calidad del agua.
- Variable independiente: turbidez, sólidos en suspensión, temperatura, nitratos, cloro, pH.
- Variable control: cantidad de muestra de agua.

Materiales

Kits de medición y muestras de agua.

Desarrollo del experimento

De forma grupal, los estudiantes realizan las mediciones oportunas para determinar la calidad de las tres muestras de agua, registrando los datos obtenidos (tabla 5.4) y elaborando gráficos. Para la utilización del material, los estudiantes emplearán la información recogida en la segunda sesión. La sesión finaliza con una puesta en común de los resultados obtenidos por cada grupo.

Tabla 5.4 *Toma de datos (sesión 3)*

Parámetros	Muestra A	Muestra B	Muestra C
Temperatura			
Nitratos			
Cloro			
pH			
Sólidos en suspensión			
Turbidez			

Los parámetros que se analizan en esta propuesta didáctica son la temperatura, los niveles de nitratos, cantidad de cloro, los sólidos en suspensión, la turbidez y el pH del agua. Para la recogida de datos, se emplea una placa microcontroladora para medir la temperatura, un kit de pH y cloro, un kit analizador de nitratos y un disco Secchi para analizar la turbidez y los sólidos en suspensión en el agua. Para medir la temperatura de las muestras, se introduce en el agua el sensor de temperatura conectado a la placa microcontroladora, se pulsa la tecla A y se anota el valor marcado en la pantalla LED de la placa microcontroladora. Los kits de pH, cloro y nitratos se han de emplear siguiendo las instrucciones que proporciona el fabricante correspondiente. Finalmente, el disco Secchi se introduce en la muestra de agua, comparando los niveles de turbidez entre las tres muestras.

SESIÓN 4: *FASE 5 (Informe y afianzamiento) de la indagación científica*

Se analizan y discuten los resultados obtenidos, extrayendo las siguientes conclusiones:

– La muestra de agua A es la que posee peores parámetros de calidad de agua debido a una elevada concentración de arena, hojas y ramitas de árbol, alta temperatura, gran cantidad de nitratos y cloro, y elevada acidez por la gran cantidad de vinagre. Por ello, ninguno de los parámetros posee valores recomendados.

- La muestra B posee valores poco recomendados en la temperatura, nitratos, cloro y pH. No obstante, posee valores óptimos en cuanto a la turbidez y los sólidos en suspensión.
- La muestra C posee valores poco indicados en la turbidez y los sólidos en suspensión. Sin embargo, la temperatura, nitratos, cloro y el pH alcanzan niveles óptimos para consumo humano.
- Se concluye que la muestra C es la que posee más parámetros fisicoquímicos con valores adecuados. Sin embargo, para que sea apta para consumo humano requiere de un proceso de filtrado y depuración.

Se interpretan los resultados obtenidos en relación con las hipótesis planteadas, discutiéndose en qué medida se confirman o no. Posteriormente, cada grupo de trabajo realiza un mural que recoge la información clave de cada una de las fases del método de indagación científica, prestando especial atención a relacionar la pregunta de investigación planteada, las hipótesis formuladas y los resultados obtenidos. Finaliza la sesión con la presentación del mural.

5.4.2. Indagación 2 sobre métodos físicos de separación de mezclas

Para que los estudiantes comprendan los métodos físicos de separación de mezclas, se retoman los resultados de la indagación sobre la calidad del agua. En la cuarta sesión se concluyó que es necesario un proceso de depuración de la muestra C para que pueda ser apta para consumo humano. De este modo, se presenta una segunda situación problemática que requiere de una nueva aplicación del método de indagación científica para su resolución. Durante el segundo proceso de indagación, se introducen diferentes métodos físicos de separación de mezclas, se realiza una búsqueda de información en páginas especializadas de internet, y se profundiza en las técnicas de i) filtrado y ii) tamizado. Se experimenta con diferentes materiales para determinar su capacidad de filtrado en base a tres variables: i) tiempo de filtrado, ii) cantidad de sólidos en suspensión removidos, y iii) turbidez del agua mejorada. Se programa una placa microcontroladora como sensor de volumen para determinar la rapidez de filtrado de cada material. Finalmente, se determinan las propiedades de separación de cada material y se discute el más apropiado para depurar la muestra de agua C de la primera indagación.

SESIÓN 5: *FASE 1 (preguntas de investigación) y 2 (hipótesis) de la indagación científica*

Se retoman los resultados obtenidos en la primera indagación, en la que la muestra C poseía niveles adecuados de temperatura, nitratos, cloro y pH, y sin embargo, presentaba altos niveles de turbidez y muchos sólidos en suspensión. A partir de esta situación problemática, se inicia un diálogo y debate con los estudiantes (¿Cómo podemos limpiar el agua?, ¿de qué forma?), concretándose una segunda pregunta de investigación: ¿Qué materiales debemos utilizar para filtrar el agua y mejorar su calidad? Los estudiantes observan y manipulan los diferentes materiales disponibles y plantean hipótesis argumentadas con respecto a las siguientes tres sub-preguntas de investigación: 1) ¿Qué material va a filtrar más cantidad de agua en un determinado periodo de tiempo?; 2) ¿Qué material va a quitar más sólidos en suspensión del agua?; y 3) ¿Qué material va a mejorar más la turbidez del agua? Algunos ejemplos de hipótesis planteadas son:

H.1 ¿Qué material va a filtrar más cantidad de agua en un determinado periodo de tiempo? La rejilla con mucha porosidad filtrará más cantidad de agua que el resto de los materiales porque el agua la atravesará muy rápidamente.

H.2 ¿Qué material va a quitar más sólidos en suspensión del agua? El filtro de café filtrará más sólidos en suspensión porque al ser poco porosa, las ramas y hojas de los árboles se quedarán atrapadas en el filtro.

H.3 ¿Qué material va a mejorar más la turbidez del agua? El algodón es el material que más mejorará la turbidez del agua porque absorberá toda la arenilla presente en el agua.

La sesión finaliza con una puesta en común de las hipótesis formuladas por cada grupo de trabajo.

SESIÓN 6: *FASE 3 (diseño experimental) de la indagación científica*

La sexta sesión se desarrolla de manera similar a la segunda sesión. La búsqueda de información es menos guiada que la anterior y la programación de la placa microcontroladora para medir la cantidad de agua filtrada se realiza de forma autónoma por cada grupo de trabajo, sin apoyo por parte del docente. Las técnicas de separación de mezclas incluidas en esta sesión son: i) filtración, ii) tamizado, iii) imantación y iv) evaporación. Se realiza una puesta en común de la información recabada, profundizándose en la filtración y el tamizado. La placa controladora se programa siguiendo las mismas instrucciones recogidas en el cuadro 1, cambiando la variable *temperatura* por *volumen*. De esta forma, la placa microcontroladora se emplea para medir la cantidad de agua filtrada en un determinado tiempo (10 segundos), lo que permite contestar a la primera sub-pregunta de investigación. Es decir, comenzado el filtrado se anota en la libreta la cantidad de agua filtrada (indicada por el micro:bit) durante 10 segundos.

SESIÓN 7: *FASE 4 (toma de datos y análisis) de la indagación científica*

Se emplean tres diseños experimentales diferentes, uno para cada sub-pregunta de investigación, completándose la tabla 5.5:

H.1 ¿Qué material va a filtrar más cantidad de agua?

Variables

Variable dependiente: tiempo de filtrado.
Variable independiente: materiales.
Variable control: cantidad de agua.

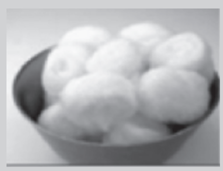
Materiales

Muestra de agua C, vaso de tubo, cronómetro, micro:bit, materiales (rejilla metálica muy porosa, rejilla metálica poco porosa, filtro de café, filtro de campana, papel de cocina, algodón).

Desarrollo del experimento

Se coloca el sensor de volumen, conectado al micro:bit, dentro del vaso de tubo. Posteriormente, se coloca cada material encima del vaso de tubo y se vierte la muestra de agua. Tras diez segundos, se anota el valor marcado del micro:bit, que representa la cantidad de agua filtrada en 10 segundos.

Tabla 5.5. *Capacidad de filtrado de los materiales (sesión 7)*

			
Tiempo (segundos)			
Turbidez (JTU)			
Sólidos en suspensión	Pocos (1) Medios (2) Muchos (3)	Pocos (1) Medios (2) Muchos (3)	Pocos (1) Medios (2) Muchos (3)

H.2 ¿Qué material va a quitar más sólidos en suspensión del agua?

Variables

Variable dependiente: sólidos en suspensión filtrados.
Variable independiente: materiales.
Variable control: cantidad de agua.

Materiales

Muestra de agua C, vaso de tubo, micro:bit, materiales (rejilla metálica muy porosa, rejilla metálica poco porosa, filtro de café, filtro de campana, papel de cocina, algodón).

Desarrollo del experimento

Se coloca cada material encima del vaso de tubo y se vierte 250 ml de la muestra de agua. Una vez filtrada toda la cantidad de agua, se examina la cantidad de sólidos en suspensión en la muestra filtrada.

H.3 ¿Qué material va a mejorar más la turbidez del agua?

Variables

Variable dependiente: turbidez de agua mejorada.
Variable independiente: materiales.
Variable control: cantidad de agua.

Materiales

Muestra de agua C, vaso de tubo, vaso con disco Secchi, micro:bit, materiales (rejilla metálica muy porosa, rejilla metálica poco porosa, filtro de café, filtro de campana, papel de cocina, algodón).

Desarrollo del experimento

Se coloca cada material encima del vaso de tubo y se vierte 250 ml de la muestra de agua. Una vez filtrada toda la cantidad de agua se vierte el agua filtrada en un vaso con disco Secchi, y se examina la turbidez del agua.

La sesión finaliza con una puesta en común de los resultados obtenidos por cada grupo.

SESIÓN 8: *FASE 5 (informe y afianzamiento) de la indagación científica*

Se analizan y discuten los resultados obtenidos, comparándose entre los distintos grupos. Se extraen varias conclusiones:

- La capacidad de filtrado de los materiales depende de su porosidad. A mayor porosidad, menos será el filtrado.

– Los materiales menos porosos (por ejemplo, rejilla metálica) filtran el agua más rápidamente, aunque los sólidos en suspensión finos no son filtrados.
– Algunos materiales son más resistentes que otros.

La sesión finaliza de manera similar a la 4ª sesión, relacionando las preguntas de investigación, las hipótesis formuladas y los resultados obtenidos.

5.4.3. Aplicación del método ingenieril para construir un filtro-depurador de agua

Se emplea el método ingenieril para otorgar la oportunidad a los estudiantes de aplicar los conocimientos adquiridos mediante la realización de las dos indagaciones anteriores. En las primeras ocho sesiones, los estudiantes han aprendido a determinar de qué parámetros depende la calidad del agua, y han experimentado con diferentes materiales para analizar su capacidad de filtrado a través de la técnica de filtración y tamizaje. Se concluye que la muestra de agua C es la que posee mejores parámetros fisicoquímicos; sin embargo, presenta altos niveles de turbidez y sólidos en suspensión. Además, se determina que algunos materiales son más útiles que otros en la depuración y filtración de los sólidos en suspensión presentes en el agua.

De este modo, se presenta una tercera situación problemática que requiere de la construcción del prototipo de un producto capaz de mejorar la turbidez y reducir los niveles de sólidos en suspensión existentes en la muestra de agua C. Así, en las cuatro sesiones restantes de la UD., los estudiantes emplean el método de diseño ingenieril para diseñar, construir y evaluar la eficacia de un prototipo de filtro-depurador.

Los estudiantes prueban sus filtros-depuradores y analizan aquellos aspectos de su diseño que pueden ser mejorados en términos de i) tiempo de filtrado, ii) coste, iii) cantidad de sólidos en suspensión filtrados y iv) mejora de la turbidez del agua. Finalmente, atendiendo a las conclusiones extraídas durante la primera (calidad del agua) y segunda indagación (métodos físicos de separación de mezclas) y, tras la aplicación del proceso ingenieril para la construcción de un filtro-depurador, los estudiantes retoman la situación problemática inicial y resuelven el problema planteado en la primera sesión.

SESIÓN 9: *FASE 1 (problema) y 2 (producto) del método ingenieril*

Se repasan los contenidos y las conclusiones extraídas en las dos indagaciones iniciales. Así, se concluye que la muestra de agua C aún no posee una calidad apta para consumo humano debido a los altos niveles de turbidez y sólidos en suspensión, y que algunos materiales poseen más capacidad de depuración que otros. En base a estos resultados, se plantea la última situación problemática, relacionada con la creación de un producto que mejore los niveles de turbidez y sólidos en suspensión de la muestra de agua. Se discuten y ejemplifican diferentes productos existentes para depurar el agua.

Finalmente, los estudiantes buscan información sobre el funcionamiento de una depuradora industrial y sus diferentes fases, y se relaciona el proceso de depuración a gran escala con la función que realiza un filtro-depurador comercial que se emplea en los grifos de los hogares, introduciéndose de este modo el producto que deberán diseñar y construir: un filtro-depurador casero. Esta búsqueda se realiza asimismo de forma autónoma por cada grupo de trabajo, bien a través de internet o mediante el empleo de códigos QR.

SESIÓN 10: *FASE 3 (diseño) del método ingenieril*

Se introducen las características que debe tener el prototipo para considerarse efectivo, así como los criterios que han de considerarse en su construcción. Los criterios están relacionados con el material, pudiéndose emplear solamente aquellos que han sido probados en la segunda indagación (p.ej.: rejilla metálica muy porosa, rejilla metálica poco porosa, filtro de café, filtro de campana, papel de cocina, algodón). Además, se limitará el número máximo de materiales que pueden ser empleados. En esta UD se ha limitado a cinco materiales en total.

En cuanto a la eficacia del filtro-depurador, esta se determinará empleando una rúbrica (tabla 5.6). Se otorgará una puntuación entre 1 y 3 puntos para cada uno de los siguientes criterios: coste del filtro depurador, tiempo de filtrado, turbidez, cantidad de sólidos en suspensión restantes. El coste de los materiales debe ser establecido por el docente. En esta UD se han establecido los siguientes precios: algodón (6€), filtro de campana (5€), filtro de café (4€), papel de cocina (2€), rejilla metálica poco porosa (4€), y rejilla metálica muy porosa (3€). Un filtro compuesto por un filtro de campana, un filtro de café, una rejilla metálica poco porosa y un papel de cocina ha obtenido la máxima puntuación en cada criterio. No obstante, existen una multitud de combinaciones de materiales que pueden obtener asimismo la máxima puntuación. Además, ha de tenerse en cuenta la colocación de los mismos, pues el orden en el que se coloca los materiales puede hacer variar la eficacia del filtro-depurador. Por ejemplo, el papel de cocina es el material menos resistente de todos, por lo que colocarlo el primero sería una decisión poco acertada. En definitiva, se trata de que los estudiantes apliquen en el diseño del filtro-depurador lo aprendido en la segunda indagación sobre las propiedades de los materiales.

De forma grupal, se dibujan varios diseños del filtro depurador (figura 5.6). En cada diseño, se deben indicar los materiales que se van a emplear, la cantidad de los mismos, y su colocación en el filtro depurador. Finalmente, cada grupo discute las ventajas y desventajas de cada diseño, eligiendo el que consideran más adecuado para su posterior construcción.

Tabla 5.6. *Rúbrica para determinar la eficacia del filtro-depurador*

Criterios	Puntuación		
	1	2	3
Coste	Más de 26 €	Entre 16 y 25 €	Menos de 15 €
Tiempo de filtrado de 250 ml de agua	Más de 181 segundos	Entre 120 y 180 segundos	Menos de 120 segundos
Turbidez	> 30 JTU	10 – 30 JTU	< 0.5 JTU
Sólidos en suspensión restantes en el agua	Muchos	Pocos	Ninguno

Figura 5.6. *Ejemplo de diseño de filtro-depurador.*

SESIÓN 11: *FASE 4 (prototipo) del método ingenieril*

Para construir el filtro depurador, cada grupo dispone de una botella de plástico de tamaño 1,5 litros, tijeras, y cinta adhesiva, que podrán ser utilizados sin límite alguno. Para cursos superiores, se ofrece libertad máxima tanto para el diseño y colocación de la botella como para la elección de los materiales que constituirán el filtro. Sin embargo, para cursos inferiores se facilitan varias instrucciones para recordar y colocar la botella de agua (cuadro 5.2). Posteriormente, se evalúa la eficacia del filtro, registrando en la ficha correspondiente (tabla 5.7) los valores obtenidos en los cuatro criterios de calidad.

Cuadro 5.2. *Creación del filtro depurador*

Paso 1: con la ayuda de un adulto, corta la botella por la mitad y agujerea el tapón de la botella.
Paseo 2: de la parte restante, corta el fondo de la botella de tal forma que obtengas un soporte sobre el que colocar tu filtro.
Paso 3: coloca el vaso en el que vas a filtrar el agua dentro del soporte.
Paso 4: coloca la botella boca-abajo encima del soporte.
Paso 5: rellena la botella con el material que va a constituir tu filtro depurador.

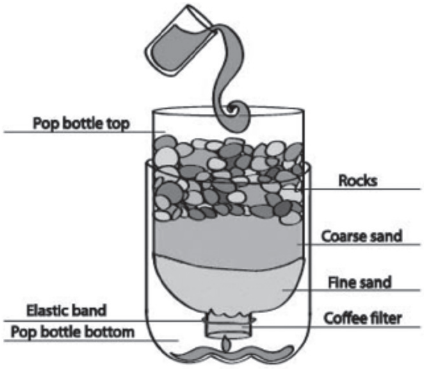


Tabla 5.7. *Eficacia del filtro depurador (sesión 11)*

Criterios	Valor obtenido	Puntuación obtenida
Coste (en €)		
Tiempo de filtrado de 250 ml de agua (en segundos)		
Turbidez (en JTU)		
Sólidos en suspensión restantes en el agua (valoración cualitativa: muchos, pocos, ninguno)		
Puntuación total =		

SESIÓN 12: *FASE 5 (mejora) del método ingenieril*

Se evalúa la eficacia del filtro y se señalan aquellos aspectos que han obtenidos menor puntuación. Se discute de forma grupal las limitaciones encontradas en el diseño propuesto y los aspectos que podrían mejorar aquellos criterios con peor puntuación. Se vuelve a realizar un diseño tratando de mejorar los aspectos poco eficaces. En el caso de que se haya obtenido una máxima puntuación en cada apartado, se procurará reducir el coste del nuevo filtro aún más o reducir el número total de materiales empleados. Posteriormente, se evalúa el nuevo diseño y se analizan los cambios realizados y cómo estas modificaciones han ayudado (o no) a mejorar la eficacia del filtro. Finalmente, se resuelve el problema inicial de la primera sesión, empleando los conocimientos adquiridos en las dos indagaciones y mediante la aplicación del método de diseño ingenieril.

5.5. Conclusiones

En este capítulo se introduce un modelo didáctico para la enseñanza integrada de las disciplinas STEM en la etapa de Educación Primara. El modelo propuesto concibe la educación STEM como la superposición y complementariedad entre las cuatro disciplinas a través de dos bloques diferenciados. El primero de ellos, enfocado en la adquisición de los conocimientos científicos propios del currículo de Ciencias de la Naturaleza de la etapa de Educación Primaria a través del empleo de la metodología de indagación escolar; y el segundo diseñado para la aplicación práctica de dicho conocimiento a través del uso del método de diseño ingenieril. Así, se pretende fomentar el aprendizaje de los contenidos objeto de estudio de una forma significativa, en conexión con otras disciplinas como la tecnología, la ingeniería y las matemáticas. De este modo, se abordan problemas socialmente relevantes y didácticamente atractivos y motivadores para los estudiantes, ofreciendo momentos para la experimentación y la adquisición y el desarrollo de los procedimientos y competencias propias del quehacer científico e ingenieril, como la identificación de problemas o preguntas investigables, la formulación de hipótesis, el diseño y desarrollo de experiencias para comprobar las hipótesis planteadas, el análisis de los datos recolectados y la discusión de los mismos en relación con el problema inicial.

El ejemplo propuesto en este trabajo ha sido implementado desde el 3^er^ al 6º curso de Educación Primaria, con estudiantes cuya edad oscila entre los 8 y 11 años de edad. El modelo recogido en este capítulo ha mostrado ser eficaz en el fomento de actitudes favorables hacia la ciencia y en el desarrollo de las expectativas de éxito en ciencias (Toma y Meneses, 2018a, 2018b). La unidad didáctica descrita se postula más acorde para estudiantes de 5º y 6º curso. No obstante, esta ha podido ser implementada asimismo en cursos inferiores, reduciendo la cantidad de parámetros fisicoquímicos de la calidad del agua y ofreciendo mayor guía y apoyo por parte del profesor.

Referencias bibliográficas

Bybee, R. W. (2011). «Scientific and Engineering practices in K-12 classrooms: understanding a framework for K-12 Science Education», en *Science Teacher*, vol. 78, págs. 34-40.

Harlen, W. (2015). *Working with Big Ideas of Science Education.* Italy, Science Education Programme.

Martin-Hansen, L. (2002). «Defining Inquiry», en *Science Teacher*, vol. 69, págs. 34-37. National Science Teachers Association.

Toma, R. B. y Meneses Villagrá, J. Á. (2018a). *Integrating computer programming into STEM: an EVIT analysis*. Actas de la Conferencia Internacional Girep 2018. Disponible en: https://www.researchgate.net/publication/326356978_Integrating_computer_programming_into_STEM_an_EVT_analysis (consultada el 25/7/2018).

Toma, R. B. y Meneses Villagrá, J. Á. (2018b). *Prueba piloto de un modelo STEM integrado con programación computacional*. Actas del 28 Encuentros de Didáctica de las Ciencias experimentales. Servizo de Publicacións Universidade da Coruña.

Appendix 3: Attitude instruments systematic review

This appendix includes the full-text of the accepted for publication article about the first systematic review conducted in this dissertation.

Toma, R. B. (in press). Revisión sistemática de instrumentos de actitudes hacia la ciencia (2004-2016). *Enseñanza de las Ciencias.*

ENSEÑANZA DE LAS CIENCIAS
Investigaciones didácticas ISSN (impreso): 0212-4521 / ISSN (digital): 2174-6486

Revisión sistemática de instrumentos de actitudes hacia la ciencia (2004-2016)

Systematic review of attitude toward science instruments (2004-2016)

Radu Bogdan Toma
Departamento de Didácticas Específicas. Universidad de Burgos, Burgos, España.
rbtoma@ubu.es

RESUMEN • Se realiza una revisión sistemática de las propiedades psicométricas que poseen 15 instrumentos de actitudes hacia la ciencia publicados entre los años 2004 y 2016, identificados en las bases de datos Web of Science Core Collection y Science Direct siguiendo los criterios PRISMA. En conjunto, los resultados revelan que la calidad de la mayoría de los instrumentos actuales sigue siendo preocupante dada la ausencia casi generalizada de adopción de un marco teórico para el desarrollo de los ítems y debido a un deficiente e inadecuado análisis de sus propiedades psicométricas. Estos hallazgos cuestionan la validez y fiabilidad de los instrumentos analizados y ponen de manifiesto la necesidad adoptar diseños metodológicos más rigurosos para el desarrollo de instrumentos de actitudes hacia la ciencia.

PALABRAS CLAVE: actitudes hacia la ciencia, desarrollo y validación, instrumentos, propiedades psicométricas, revisión sistemática.

ABSTRACT • A systematic review of the psychometric properties of 15 attitudes towards science instruments published between 2004 and 2016, identified in the Web of Science Core Collection and Science Direct databases following the PRISMA criteria, is reported. Overall, the quality of most examined instruments continues to be worrying due to an almost generalized absence of adoption of a theoretical framework for item development and due to a scarce and inadequate analysis of their psychometric properties. These findings calls into question the validity and reliability of the instruments examined and highlight the need to adopt more rigorous methodological designs for the development of attitude toward science measurement instruments.

KEYWORDS: attitudes towards science, development and validation, instruments, psychometric properties, systematic review.

INTRODUCCIÓN

En los últimos años, nuevos enfoques educativos para reducir el desinterés hacia la ciencia han cobrado un fuerte impulso y, actualmente, se están promoviendo en gran medida iniciativas enfocadas al fomento de los conocimientos y de las competencias científicas necesarias que permitan a los estudiantes una participación activa y responsable en la sociedad (EC, 2015). Así, uno de los principales objetivos es el desarrollo de actitudes favorables hacia la ciencia (Osborne, Simon, y Collins, 2003), con propuestas como *Horizonte 2020,* promovida por la Comisión Europea, que subrayan la necesidad de hacer atractivas las ciencias para todos los jóvenes (EC, 2016). Por lo tanto, se están impulsando medidas educativas para aumentar la exposición de los estudiantes a cuestiones científicas, especialmente desde grados elementales del sistema educativo (NRC, 2011; EC, 2016), siendo numerosas las políticas educativas consistentes con estas recomendaciones (Bybee, 2013; NRC, 2012, 2013). Este énfasis en desarrollar prácticas educativas dirigidas a despertar el interés y el desarrollo de actitudes favorables hacia la ciencia ha sido acompañado por el desarrollo de nuevos instrumentos de autoinforme para la evaluación de estas reformas.

Dados los resultados reportados por Blalock, Lichtenstein, Owen, Pruski, Marshall, y Toepperwein (2008) en relación con la falta de evidencias psicométricas de los instrumentos de actitudes hacia la ciencia publicados hasta el año 2005, resulta necesario evaluar si los instrumentos diseñados en los últimos años poseen propiedades psicométricas más robustas que los publicados con anterioridad. Aunque algunas investigaciones han abordado esta cuestión recientemente, aún no se ha realizado un análisis exhaustivo de las propiedades psicométricas de los instrumentos desarrollados y publicados después de la revisión de Blalock et al. (2008). Por ejemplo, Aydeniz y Kotowski (2014) informaron sobre algunas limitaciones conceptuales de apenas cinco instrumentos de los comúnmente utilizados para medir las actitudes hacia la ciencia y Potvin y Hasni (2014) proporcionaron información descriptiva sobre algunos de los instrumentos sin examinar la calidad de estos en base a las propiedades psicométricas.

Por esta razón, el propósito de este estudio consiste en actualizar el trabajo de Blalock et al. (2008) identificando y analizando los instrumentos de actitudes hacia la ciencia publicados en los últimos trece años para determinar en qué medida los nuevos esfuerzos que se están promoviendo para la mejora de la educación científica están siendo complementados con el desarrollo de instrumentos de autoinforme válidos y fiables.

MÉTODO

En la realización de este trabajo se ha seguido el procedimiento descrito por Bennett, Lubben, Hogarth y Campbell (2005) para la realización de revisiones sistemáticas y, la declaración PRISMA (Liberati et al., 2009) para el reporte transparente de los resultados derivados de revisiones sistemáticas.

Las revisiones sistemáticas difieren de las revisiones narrativas en que son menos propensas a sesgos de investigación, por lo que se introduce una mayor objetividad al explicar rigurosa y explícitamente el proceso de selección de los estudios y el método utilizado para su revisión y evaluación. Por lo tanto, las revisiones sistemáticas "(...) incluyen una fase de codificación más detallada en la que se identifican las características clave de los estudios con el fin de proporcionar una visión general del trabajo en el área bajo consideración" (ibíd., p. 390 [traducción propia]). Por este motivo, las revisiones sistemáticas tienden a ser más transparentes y reproducibles que las revisiones tradicionales. En consecuencia, a continuación, se detallan las principales fases de este estudio de acuerdo con el procedimiento introducido por Bennett et al., (2005).

Criterio de elegibilidad

Debido a la multiplicidad de definiciones existentes en la literatura relacionada con las actitudes hacia la ciencia, es necesario proveer primero una definición operativa del constructo objeto de estudio. Si bien en el contexto español existen autores españoles muy prolíficos en esta línea de investigación, como por ejemplo Vázquez y Manassero (1995) que han propuesto una taxonomía de las actitudes hacia la Ciencia, este estudio pretende ser un *continuum* del de Blalock et al. (2008), por lo que se ha adoptado la misma conceptualización de actitudes hacia la ciencia que estos autores que, además, se encuentra en consonancia con la empleada en contextos iberoamericanos. Así, siguiendo a Gardner (1975), se define las actitudes hacia la ciencia como las respuestas emocionales hacia lo científico, es decir, hacia la ciencia en general y hacia las disciplinas, carreras o asignaturas científicas en particular. Por tanto, se ha establecido que serán elegibles aquellos estudios cuya conceptualización de las actitudes hacia la ciencia sea coherente con esta definición. Respuestas emocionales como interés –"Espero con ganas las clases de ciencias" (Wang y Berlin, 2010, p. 2423 [traducción propia])–, apreciación –"La ciencia puede ayudar a hacer del mundo un lugar mejor" (Navarro et al., (2015, p. 1465)– o disfrute –"Creo que aprender ciencias es divertido (Zhang y Campbell, p. 601 [traducción propia])– son ejemplos que reflejan actitudes hacia la ciencia.

Búsqueda bibliográfica

Se identificaron artículos potencialmente relevantes para el objetivo de este estudio en las bases de datos *Web of Science Core Collection* y *Science Direct,* empleando una estrategia de búsqueda consistente con el criterio de elegibilidad. En la base de datos de *Web of Science Core Collection* se utilizó una combinación de los términos «actitud* hacia la* ciencia*» como palabras clave en los títulos e «instrumento* OR escala* OR medida*» como palabras clave en el tema, tanto en castellano como en inglés[1]; se restringió la búsqueda por categoría de investigación (Educación e investigación educativa), tipo de documento (artículos), idioma (inglés y español) y año de publicación (2004-2016). En la base de datos *Science Direct*, estos términos se emplearon como palabras clave en todos los campos y se restringió la búsqueda por categoría de investigación (ciencias sociales), tipo de documento (artículos) y año de publicación (2004-2016). Debido a que el instrumento más reciente revisado por Blalock et al. (2008) databa del año 2003, se estableció el año 2004 como año inicial y, dado que la búsqueda en las bases de datos mencionadas se realizó durante la primera semana de agosto de 2017, se estableció el año 2016 como el último año de publicación de artículos.

Tras seleccionar los artículos relevantes de las bases de datos, se utilizó un enfoque de ascendencia, que consiste en examinar la lista de referencias de los artículos seleccionados en busca de artículos potencialmente relevantes que no han sido identificados empleando la estrategia de búsqueda descrita anteriormente.

Criterios de inclusión y exclusión

Se formularon los siguientes criterios de inclusión y exclusión, atendiendo al criterio de elegibilidad en este estudio: (i) estudios de desarrollo y validación de instrumentos, (ii) centrados en el constructo de actitudes hacia la ciencia, (iii) que emplean la teoría clásica de test (*Classical Test Theory*) para la validación del instrumento y, (iv) cuyos instrumentos sean de naturaleza cuantitativa. Se excluyeron los estudios que emplean instrumentos de actitudes hacia la ciencia pero que, sin embargo, no están específicamente enfocados a su desarrollo y validación, los que no cumplen con el criterio de elegibilidad en el que se define el constructo de actitud hacia la ciencia en este estudio, los que no utilizan la teoría clásica de pruebas para la validación del instrumento (es decir, los que utilizan el análisis Rasch) y, los que utilizan métodos cualitativos de recopilación de datos (por ejemplo, entrevistas o dibujos de los estudiantes).

---

[1]En inglés, se han empleado los siguientes términos: «attitude* to* science*» e «instrument* OR scale* OR measure*».

Rúbrica de evaluación

Se empleó la rúbrica desarrollada por Blalock et al. (2008) para evaluar las propiedades psicométricas de los instrumentos incluidos en este estudio, lo que podría permitir la comparación entre los instrumentos incluidos en este trabajo y los analizados en el estudio de Blalock et al. (2008). No obstante, han sido necesarias ligeras adaptaciones. La rúbrica original constaba de cinco secciones: (i) marco teórico; (ii) fiabilidad; (iii) validez; (iv) dimensionalidad y (v) uso, con una evaluación total posible de 28 puntos. La última sección pretendía evaluar hasta qué punto los instrumentos desarrollados habían sido empleados en estudios posteriores al de su desarrollo y validación. Sin embargo, una revisión inicial de los artículos incluidos en este trabajo reveló que aproximadamente la mitad de ellos fueron publicados entre los años 2014 y 2016. Por lo tanto, esta sección de la rúbrica fue descartada en este estudio para asegurar una igualdad de condiciones y no perjudicar los instrumentos publicados en años posteriores, que cuentan con menos probabilidad de haber sido utilizados en otros estudios. Este cambio redujo la puntuación total posible de 28 a 27 puntos (Tabla 1).

Adicionalmente, dado los avances en las definiciones y tipos de propiedades psicométricas surgidos desde el desarrollo de la rúbrica hace ya más de una década, especialmente en relación con las pruebas de validez de los instrumentos, se ha empleado la taxonomía de Polit (2015) y Polit y Yang (2016) para actualizar la terminología referente a la validez de los instrumentos, sustituyéndose tres términos empleados por Blalock et al., (2008) por los propios de la taxonomía revisada y añadiéndose dos criterios de validez adicionales no contemplados en la rúbrica original. Estos cambios, así como la explicación de cada prueba de fiabilidad y validez, se recogen en la Tabla 2.

Fiabilidad en el análisis de los instrumentos

Para asegurar que cada instrumento fuera evaluado de manera correcta y consistente, se efectuó una medición formal de la fiabilidad intra-evaluador en la aplicación de la rúbrica. El autor de este estudio aplicó dos veces la rúbrica de evaluación a cada instrumento con un período de diferencia de 6-8 semanas entre la primera y la segunda aplicación de la rúbrica y calculó la correlación entre los dos conjuntos de datos para examinar si existen discrepancias entre las dos aplicaciones.

El índice de correlación de Pearson reportó una relación significativa ($p < 0,001$) entre los dos conjuntos de evaluaciones de cada sección de la rúbrica. Los datos de la primera y segunda aplicación de la rúbrica estuvieron perfectamente correlacionados ($r = 1$) en las secciones «Marco teórico», «Test re-test», «Error estándar de medición» y «Dimensionalidad» y, fuertemente correlacionados en las secciones «Consistencia interna» ($r = 0,98$) y «Validez» ($r = 0,93$). Los instrumentos en los que no se ha identificado una correlación perfecta entre la primera y la segunda evaluación fueron examinados por tercera vez para resolver las discrepancias existentes. Las discrepancias encontradas consistieron en asignar una consistencia interna ligeramente mayor a una dimensión de un instrumento y en no identificar una validez de contenido en un instrumento. Cabe destacar que estas discrepancias no afectaron la puntuación total asignada a los instrumentos, por lo que los resultados de este estudio poseen un alto nivel de fiabilidad.

Tabla 1. Rúbrica de evaluación

Aspecto	Puntuación			
Marco teórico (0-3)				
Antecedentes teóricos para el desarrollo del instrumento	0 (No)	3 (Sí)		
Fiabilidad (0-9)				
Consistencia interna	0 (no reportado o $\alpha < 0.60$)	1 (α = 0.61-0.70)	2 (α = 0.71-0.80)	3 ($\alpha > 0.80$)
Test-retest	0 (no reportado o $r < 0.60$)	1 (r = 0.61-0.70)	2 (r = 0.71-0.80)	3 ($r > 0.80$)
Error estándar de medición	0 (No)	3 (Sí)		
Validez (0-9)				
Contenido, constructo, convergente, concurrente, discriminante, discriminativa, predictiva	0 (no reportado)	3 (1-2 evidencias)	6 (3-4 evidencias)	9 (> 4 evidencias)
Dimensionalidad (0-6)				
En instrumentos unidimensionales, se ha empleado la puntuación global. En instrumentos multidimensionales, se ha empleado la puntuación de cada dimensión.	0 (No)	3 (Sí)		
En instrumentos sometidos a análisis factorial, ¿las subescalas reflejan los factores adecuados?	0 (No)	3 (Sí)		
	Puntuación total 0-27			

Tabla 2. Taxonomía revisada y explicación de cada propiedad psicométrica

Blalock et al. (2008)	Taxonomía revisada[a]	Explicación[a]
Fiabilidad		
Consistencia interna	Consistencia interna	Refiere al grado en que los ítems miden el mismo constructo y si los ítems seleccionados están interrelacionados entre sí, utilizando α de Cronbach
Test-retest	Test-retest	Evalúa la estabilidad y reproducibilidad de los resultados, partiendo del supuesto de que el rasgo objeto de estudio no ha cambiado en los sujetos
Error estándar de medición	Error estándar de medición	Examina en qué medida los resultados del instrumento están libres de error de medición
Validez		
Contenido	Contenido	Refiere a la idoneidad y relevancia de los ítems para representar la naturaleza y dimensionalidad del constructo objeto de estudio
Congruente	Concurrente	Comprueba si los resultados son coherentes en comparación con otro instrumento de referencia que mide el mismo constructo
Discriminante	Discriminante	Prueba las hipótesis de que el instrumento mide el constructo objeto de estudio y no un constructo diferente al previsto
Grupos de contraste	Discriminativa	Examina el grado en que el instrumento discrimina entre grupos que teoréticamente difieren con respecto al constructo objeto de estudio
Análisis factorial	Constructo	Valora si el instrumento capta la dimensionalidad hipotética del constructo objeto de estudio, utilizando análisis factorial
	Convergente	Evalúa la correlación entre las puntuaciones del constructo objeto de estudio y las de un constructo de convergencia conceptual
	Predictiva	Comprueba si el instrumento predice aquellos constructos que teoréticamente debería predecir

[a]La taxonomía revisada y su explicación se basa en la taxonomía propuesta por Polit (2015) y Polit y Yang (2016).

RESULTADOS

Identificación y selección de estudios

En la Figura 1 se recoge el proceso de identificación y selección de los estudios atendiendo a los criterios PRISMA. En la primera fase, llamada *identificación*, 104 y 54 referencias fueron recuperadas de las bases de datos *Web of Science* y *Science Direct*, respectivamente. En la segunda fase, denominada *cribado,* se aplicaron los filtros de búsqueda, reteniéndose un total de 65 artículos. En la tercera fase, *elegibilidad*, se excluyeron un total de 54 artículos tras la lectura completa de los mismos y la aplicación de los criterios de inclusión. Finalmente, 14 artículos fueron retenidos para su análisis en profundidad en la fase de *inclusión,* once de los cuales fueron derivados de las bases de datos consultadas y tres mediante el enfoque de ascendencia.

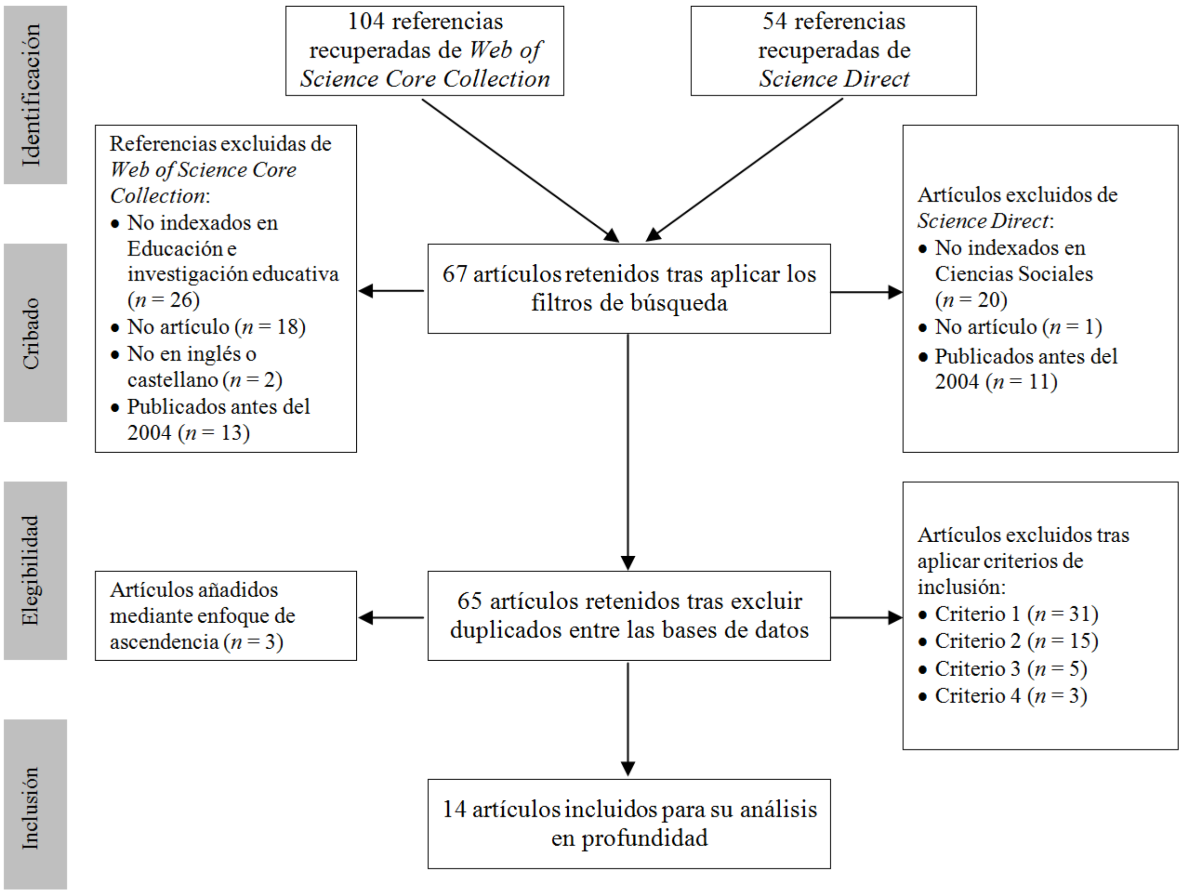


Figura 1. Identificación y selección de estudios según PRISMA

Descripción general de los instrumentos

La Tabla 3 recoge la síntesis y evaluación de los instrumentos incluidos en este estudio. Los 14 artículos seleccionados ofrecían información sobre el desarrollo y la validación de un total de 15 instrumentos. Cinco de estos instrumentos fueron validados empleando muestras de la etapa primaria (hasta 10-11 años, 5º curso EPO), diez de la etapa media (hasta 13-14 años, 6º EPO - 2º ESO), tres en la etapa secundaria (hasta 17-18 años, 3º ESO - 2º BACH), dos para estudiantes universitarios y uno para profesorado[2]. Un total de 12 instrumentos fueron desarrollados con muestras de habla inglesa. De los restantes, el instrumento de Abd-

[2] Algunos instrumentos son aptos para varias etapas educativas, tal y como se indica en la Tabla 3.

El-Khalick, Summers, Said, Wang, y Culbertson (2015) fue validado con muestras de habla árabe e inglesa y el de Navarro, Förster, González, y González-Pose (2016) con muestras de habla española. No se pudo identificar el idioma objeto de estudio del instrumento desarrollado con estudiantes taiwaneses por Wang y Berlin (2010) pues los autores no indicaron en qué idioma fue administrado.

En relación con la estructura de los instrumentos, tan solo Wang y Berlin (2010) desarrollaron uno de naturaleza unidimensional, siendo los restantes multidimensionales. Por otro lado, doce instrumentos emplearon ítems de tipo Likert con un formato de cinco opciones de respuesta, uno utilizó ítems de diferencial semántico y otro una combinación de ambos tipos. La longitud de los instrumentos varió considerablemente, siendo el más corto el desarrollado por Kennedy, Quinn y Taylor (2016), compuesto por apenas diez ítems y, el más largo la adaptación al español del TOSRA de Fraser (1981), realizada por Navarro et al. (2016), que consta de un total de 70 ítems.

Cabe destacar que algunos instrumentos ampliamente utilizados en el contexto español han sido descartados atendiendo al primer criterio de inclusión. Así, por ejemplo, los instrumentos PANA (de Pro y Pérez, 2014) y el ROSE, empleado principalmente por Vázquez y Manassero (2009a), Fernández-Cézar, Pinto-Solano y Muñoz-Hernández (2018) y Marbá-Tallada y Márquez (2010), son empleados en estudios cuyo objetivo principal no es el desarrollo y la validación del propio instrumento. Por otro lado, aunque Vázquez y Manassero (2008, 2009b) han analizado la estructura factorial del instrumento ROSE, estos artículos tampoco están enfocados a la validación de este instrumento.

Calidad psicométrica de los instrumentos

En conjunto, la puntuación final de los instrumentos osciló entre 3 y 17 puntos (Figura 2). Más de la mitad de los instrumentos (9 de 15, 60%) obtuvieron una puntuación inferior a la mitad de la puntuación total de la rúbrica. El instrumento desarrollado por Villafañe y Lewis (2016) obtuvo la puntuación más alta (17 de un total de 27 puntos posibles) y el instrumento de Hillman et al. (2016) fue el que obtuvo menor puntuación, con apenas 3 de los potenciales 27 puntos.

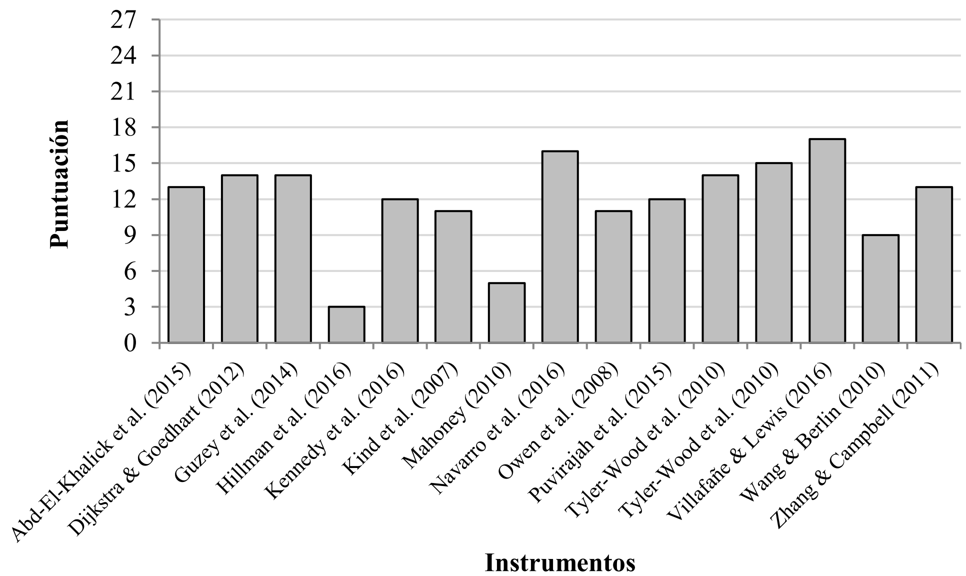


Figura 2. Puntuación total de los instrumentos analizados

Atendiendo a los resultados según las secciones de la rúbrica empleada, se observó que 11 de los 15 instrumentos analizados fueron desarrollados sin un marco teórico de actitudes hacia la ciencia (Tabla 3). De los estudios que sí habían empleado un marco teórico, Abd-El-Khalick et al. (2015) desarrolló un instrumento basado en la Teoría de la Conducta Planeada de Ajzen (1991); Navarro et al. (2016) y Villafañe y Lewis (2016) emplearon la clasificación de actitudes planteada por Klopfer (1971) y posteriormente popularizada por Fraser (1981) y su instrumento TOSRA; finalmente, Zhang y Campbell (2011) emplearon la conceptualización tripartita según la cual las actitudes se clasifican en componentes afectivos, cognitivos y conductuales (Breckler, 1984; Eagly y Chaiken, 1998).

En relación con la fiabilidad de los instrumentos, ningún estudio reportó evidencias psicométricas de fiabilidad temporal (test-retest) ni información acerca del error estándar de medición de los instrumentos desarrollados. Todos los estudios menos el de Abd-El-Khalick et al. (2015), que empleó índices de fiabilidad a partir de análisis factorial confirmatorio, reportaron el índice de α de Cronbach como única evidencia de fiabilidad. No obstante, en la figura 3 se observa que cuatro instrumentos poseen una consistencia interna inferior a la mínima aceptada para estudios exploratorios, según Nunnally (1978), y menos de una cuarta parte de los instrumentos poseen niveles de fiabilidad robustos.

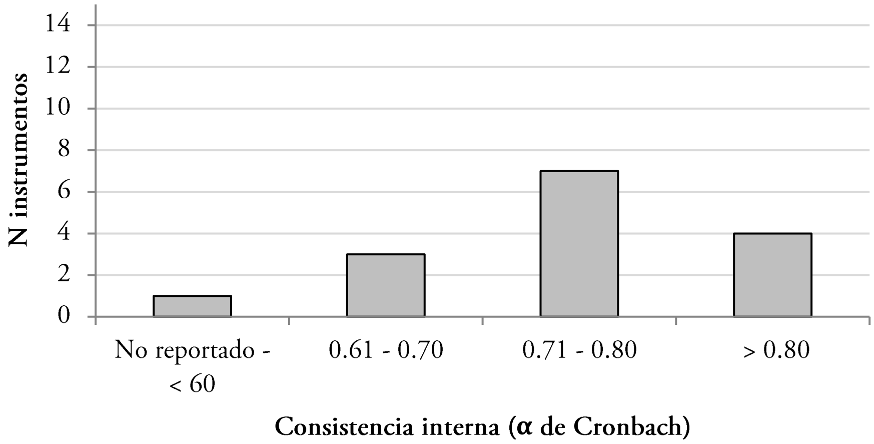


Figura 3. Consistencia interna de los instrumentos analizados

En cuanto a la validez de los instrumentos, más de la mitad (9 de 15, 60%) apenas reportaron una o dos evidencias psicométricas y los seis instrumentos restantes analizados proporcionaron tres o cuatro evidencias de validez. Cabe destacar que ningún instrumento fue sometido a más de cuatro pruebas de validez psicométrica. De las posibles pruebas psicométricas disponibles para analizar la validez de un instrumento, la validez de constructo fue la más empleada, seguida por la validez de contenido (Figura 4). Ningún instrumento fue sometido a pruebas psicométricas de validez concurrente y apenas un instrumento proporcionó información sobre la validez discriminante y predictiva.

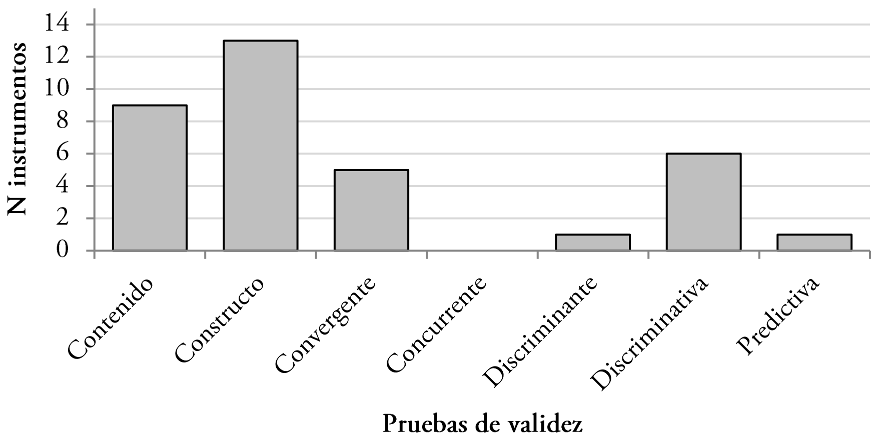


Figura 4. Frecuencia de pruebas de validez reportadas en los instrumentos analizados

Tabla 3. Resumen y evaluación de los instrumentos analizados

Autores	Muestra[a]	Ítems	Tipo de ítems	Marco teórico (0-3)	Fiabilidad (0-9)[b]	Validez (0-9)	Dimensionalidad (0-6)	Puntuación total (0-27)
Abd-El-Khalick, Summers, Said, Wang y Culbertson (2015)	A B C	32	Likert	Sí (3)	.61-.87[c] (1)	Contenido Constructo (3)	Sí (3) Sí (3)	13
Dijkstra y Goedhart (2012)	B	38	Likert	No (0)	α = .71-.87 (2)	Contenido Constructo Convergente Discriminativa (6)	Sí (3) Sí (3)	14
Guzey, Harwell y Moore (2014)	A	28	Likert	No (0)	α = .77-.87 (2)	Contenido Constructo Discriminativa (6)	Sí (3) Sí (3)	14
Hillman, Zeeman, Tilburg y List (2016)	A B C	40	Likert	No (0)	α = .43-.91 (0)	Contenido (3)	-	3
Kennedy, Quinn y Taylor (2016)	B	10	Likert Diferencial semántico	No (0)	α = .82-.98 (3)	Contenido Constructo (3)	Sí (3) Sí (3)	12
Kind, Jones y Barmby (2007)	B	45	Likert	No (0)	α = .72-.94 (2)	Constructo Convergente (3)	Sí (3) Sí (3)	11
Mahoney (2010)	C	23	Likert	No (0)	α = .76-.96 (2)	Contenido Discriminativa (3)	-	5
Navarro, Förster, González y González-Pose (2016)	B	70	Likert	Sí (3)	α = .63-.91 (1)	Constructo Convergente Discriminante (6)	Sí (3) Sí (3)	16

Tabla 3. Resumen y evaluación de los instrumentos analizados (continuación)

Autores	Muestra[a]	Ítems	Tipo de ítems	Marco teórico (0-3)	Fiabilidad (0-9)[b]	Validez (0-9)	Dimensionalidad (0-6)	Puntuación total (0-27)
Owen et al. (2008)	B	22	Likert	No (0)	α = .75-78 (2)	Constructo Discriminativa (3)	Sí (3) Sí (3)	11
Puvirajah, Verma, Li y Martin-Hansen (2015)	B	22	Likert	No (0)	α = .83-.90 (3)	Constructo (3)	Sí (3) Sí (3)	12
Tyler-Wood, Knezek y Christensen (2010)	B	12	Likert	No (0)	α = .78-.94 (2)	Contenido Constructo Convergente (6)	Sí (3) Sí (3)	14
Tyler-Wood, Knezek y Christensen (2010)	B D E	25	Diferencial semántico	No (0)	α = .84-.93 (3)	Contenido Constructo Discriminativa (6)	Sí (3) Sí (3)	15
Villafañe y Lewis (2016)	D	24	Likert	Sí (3)	α = .77-.87 (2)	Constructo Discriminativa Predictiva (6)	Sí (3) Sí (3)	17
Wang y Berlin (2010)	A	30	Likert	No (0)	α = .93 (3)	Contenido Constructo (3)	Sí (3) No (0)	9
Zhang y Campbell (2011)	A	28	Likert	Sí (3)	α = .65-.88 (1)	Constructo Convergente (3)	Sí (3) Sí (3)	13

[a]Los cursos equivalentes a España son: A (hasta 5º E.P.O); B (6º E.P.O hasta 2º E.S.O); C (3º E.S.O hasta 2º BACH); D (universitarios); E (profesorado).

[b]La puntuación ha sido otorgada atendiendo al valor más bajo de fiabilidad.

[c]Los autores han empleado estadísticos de fiabilidad a partir de análisis factorial confirmatorio en lugar de α de Cronbach.

DISCUSIÓN

Los resultados de esta revisión indican que las críticas realizadas por Blalock et al. (2008) hace una década siguen estando vigentes para los estudios actuales enfocados al desarrollo y a la validación de instrumentos de actitudes hacia la ciencia. Además, lo que resulta aún más preocupante es descubrir que las deficiencias en los instrumentos analizados por Munby (1983) hace casi cuarenta años han perpetuado en esta línea de investigación hasta la actualidad. Así, aunque el uso de un marco teórico resulta imprescindible para el desarrollo de instrumentos, los estudios revisados se caracterizan por la ausencia de marco teórico para la conceptualización de qué constituyen las actitudes hacia la ciencia. Por lo tanto, parece ser que los autores de estos instrumentos se han limitado meramente a crear o adaptar ítems a partir de cuestionarios existentes (en la mayoría de los casos también con una pobre conceptualización teórica de las actitudes) y, a emplear dimensiones actitudinales de forma aleatoria y sin una base teórica coherente. Ello lleva a que se incluya bajo el mismo *paraguas* constructos tan diversos -y que cuentan con marcos teóricos propios diferenciados- como creencias y valores (Hilman et al. 2016), autoeficacia y motivaciones (Puvirajah et al. 2015) o habilidad percibida, valores y compromiso (Mahoney, 2010). Asimismo, la ausencia de un marco teórico conduce a la inclusión de constructos cuya relevancia en el estudio de las actitudes hacia la ciencia es en el mejor de los casos dudosa. Por ejemplo, Kind et al. (2007) han incluido un constructo que mide las actitudes hacia la escuela en general y un constructo que mide el «interés combinado» hacia la ciencia; Dijkstra y Goedhart (2012) han mezclado constructos tan dispares como actitudes hacia la ciencia escolar, los científicos, la necesidad de abordar el cambio climático y, comportamientos pro-ambientales sin ofrecer justificación alguna, por lo que la inclusión de estos aspectos en un mismo instrumento resulta cuanto menos cuestionable.

El segundo aspecto que resulta preocupante es la escasez de estudios que han sometido los instrumentos desarrollados a pruebas psicométricas de validez concurrente y discriminante para determinar en qué medida el instrumento propuesto aborda el constructo que se pretende medir. Como consecuencia del desuso de estas pruebas psicométrica, no es de extrañar la baja fiabilidad general observada en los instrumentos analizados, claro indicio de que los ítems no miden el mismo constructo latente. Por otro lado, la falta de información sobre el error estándar de medición de los instrumentos plantea la cuestión de en qué medida los resultados obtenidos mediante el uso de estos instrumentos son el reflejo de la actitud de los individuos objeto de estudio o es el efecto de un instrumento poco fiable. Asimismo, la ausencia de evidencias de fiabilidad temporal (test-retest) hace que estos instrumentos no sean apropiados para estudios longitudinales o estudios enfocados a la evaluación de intervenciones educativas (diseños pretest-posttest). Dado que la estabilidad y reproducibilidad de los resultados no está garantizada, los investigadores que evalúan intervenciones educativas emplearían estos instrumentos *a ciegas* y sin posibilidad alguna de conocer con certeza en qué medida la intervención educativa ha sido eficaz.

El tercer aspecto preocupante es la proliferación del uso de la validez aparente (en inglés, *face validity*) como una evidencia de validez psicométrica, o el mal uso de algunas pruebas psicométricas. Por ejemplo, Kennedy et al. (2016) y Wang y Berlin (2010), entre otros, han empleado la prueba de validez aparente pese a que es bien sabido que se trata de una práctica poco apropiada (DeVellis, 2003) y que, por consiguiente, los instrumentos deben ser sometidos a una validez de contenido mediante el uso de un panel compuesto por expertos en el constructo objeto de estudio. En cuanto al mal uso de algunas pruebas, por ejemplo, Tytler-Wood et al. (2010) han realizado un análisis factorial con un tamaño muestral inferior a 100 casos. Otro ejemplo de prácticas pocos recomendadas está relacionado con la gran mayoría de los estudios que han realizado análisis factorial exploratorio y que han empleado el método *Little Jiffy* (Kaiser, 1970), práctica que consiste en elegir un método de extracción de componentes principales, una rotación ortogonal Varimax y determinar el número de factores a extraer basado en el criterio Kaiser > 1. Este procedimiento ha sido duramente criticado en la literatura por la falta de robustez en los resultados que proporciona y por extraer más dimensiones de las que realmente subyacen al constructo objeto de estudio, por lo que se trata de un método ampliamente desaconsejado para constructos en los que se espera una alta correlación entre las dimensiones (Gaskin y Happell, 2014; Lloret-Segura, Ferreres-Traves, Hernández-Baeza, y Tomás-Marco, 2014).

Por último, todos los instrumentos analizados en este estudio han sido desarrollados empleando ítems que carecen de una base empírica, por lo que han sido formulados siguiendo el criterio el autor del instrumento o a partir de instrumentos existentes. Si bien se trata de una práctica extendida en el desarrollo y la validación de instrumentos en general, Ryan y Aikenhead (1992) advirtieron que estos ítems suelen reflejar las ideologías de los investigadores en lugar de medir con precisión las actitudes reales de los individuos objeto de estudio. Por ello, estos autores propusieron desarrollar instrumentos basados en las respuestas de los estudiantes durante las entrevistas, dotando así a los ítems de una naturaleza empírica, siendo esta recomendación ampliamente ignorada en los instrumentos analizados en este estudio.

IMPLICACIONES

De los resultados de este estudio se derivan varias implicaciones. Debido al panorama desolador de la calidad psicométrica de los instrumentos disponibles para la medición de las actitudes hacia la ciencia, cabría preguntarse por el grado de confianza que se puede depositar en los resultados derivados del uso de estos instrumentos cuya validez y fiabilidad están cuestionadas. Ello requiere, en primera instancia, una reconceptualización del constructo de actitud hacia la ciencia y, posteriormente, el desarrollo de instrumentos que hayan sido sometidos a un riguroso proceso de diseño y validación. Solo de este modo se podría disponer de medidas de autoinforme con propiedades psicométricas más robustas que los instrumentos disponibles actualmente, lo que ayudaría a obtener resultados válidos y fiables que ayuden a respaldar (o en su caso refutar) los supuestos establecidos y consensuados en esta línea de investigación.

Por otro lado, dado la escasez de instrumentos identificados en esta revisión, y en la de Blalock et al. (2008), que hayan sido desarrollados en el contexto español, futuros estudios deberían desarrollar nuevos instrumentos o adaptar y validar instrumentos existentes en otros idiomas para su uso con estudiantes de habla hispana, similar al trabajo de Navarro et al. (2016), revisado en este estudio. De los instrumentos evaluados en este trabajo, quizás los mejores candidatos para este propósito sean el instrumento de Abd-El-Khalick et al. (2015) y el de Zhang y Campbell (2011), que si bien no han sido de los que obtuvieron una mayor puntuación, son instrumentos que han sido desarrollados empleando dos marcos teóricos de amplia trayectoria en el estudio de las actitudes en el ámbito de la psicología social, la Teoría de la Conducta Planeada de Ajzen (1991) en el primero de los casos y la conceptualización tripartita de las actitudes en el segundo de los casos.

En este estudio, se han identificado algunos instrumentos que han sido empleados con asiduidad en el contexto español (p. ej. ROSE o PANA). No obstante, estos instrumentos no han sido retenidos para su revisión debido a que han sido utilizados en estudios cuyo objetivo principal no era el análisis de las propiedades psicométricas de estos instrumentos, incumpliendo de este modo el primer criterio de inclusión establecido en este trabajo. Por tanto, parecería necesario que en futuros trabajos se revise la producción iberoamericana para determinar en qué medida el conocimiento desarrollado sobre las actitudes en España y países latinoamericanos ha sido desarrollado en base a instrumentos con una calidad robusta en términos de validez y fiabilidad.

Por último, es necesario desarrollar y consolidar un cuerpo robusto de instrumentos que reporten resultados válidos y fiables que permitan hacer comparaciones entre las actitudes hacia la ciencia identificadas a nivel internacional y local, así como entre la efectividad de las diferentes intervenciones educativas enfocadas a la promoción de actitudes favorables. El uso de un procedimiento e instrumentos similares para la recolección de datos permitiría una mejor comparación de los resultados obtenidos. Por lo tanto, parecería apropiado desarrollar y validar instrumentos para este fin a otros contextos, niveles educativos e idiomas.

LIMITACIONES Y SIGNIFICANCIA DE ESTE ESTUDIO

Los resultados presentados en esta revisión deben interpretarse teniendo en cuenta que sólo un investigador participó en el proceso de revisión de todos los instrumentos. Esta limitación se ha tratado de minimizar aplicando la rúbrica de evaluación dos veces a cada uno de los instrumentos y calculando la fiabilidad intra-evaluador para obtener resultados fiables. Aunque la principal deficiencia de este método reside en que el evaluador está sujeto a los mismos errores y sesgos, el período de dos meses entre la primera y la segunda evaluación del mismo instrumento ha reducido potencialmente esta deficiencia. Además, el empleo de una rúbrica para la evaluación de los instrumentos ha minimizado el sesgo y los errores que podrían ser introducidos por el investigador, pues el análisis consistió en identificar aquellas pruebas psicométricas que han sido empleadas en cada estudio para validar un instrumento de actitudes hacia la ciencia.

Aunque los resultados de este trabajo deben ser interpretados considerando esta limitación, esta revisión proporciona una evaluación en profundidad de las propiedades psicométricas de los instrumentos de actitudes hacia la ciencia publicados en los últimos doce años, lo cual puede ser útil por diferentes razones para la investigación en el campo de la didáctica de las ciencias. En primer lugar, en este trabajo se resaltan aquellas prácticas que necesitan ser mejoradas en el desarrollo y la validación de instrumentos. En segundo lugar, este estudio proporciona una guía valiosa que permite a los investigadores examinar las ventajas y desventajas de cada instrumento, resaltando aquellos con mejores y peores propiedades psicométricas. Esto facilita a los autores interesados en el estudio de las actitudes hacia la ciencia la selección de instrumentos de medida de acuerdo con sus necesidades de investigación. En tercer lugar, en esta revisión se ofrece una rúbrica actualizada de los estándares psicométricos modernos, que puede ser utilizada para evaluar los instrumentos de actitudes hacia la ciencia que no han sido incluidos en este trabajo, como, por ejemplo, aquellos empleados en el contexto de la investigación en didáctica de las ciencias desarrollada en España y en Latinoamérica. Asimismo, esta rúbrica podría ser empleada para evaluar la fiabilidad y validez de otros instrumentos de naturaleza cuantitativa que están enfocados a medir otros constructos de relevancia educativa como, por ejemplo, la Naturaleza de la Ciencia, las motivaciones o la auto-eficacia. Por último, este trabajo podría fomentar futuras investigaciones en la línea de las actitudes hacia la ciencia al plantear un debate sobre la necesidad de revisar aquellos supuestos que han sido consensuados a partir de los resultados derivados del uso de instrumentos de limitada validez y fiabilidad.

CONCLUSIONES

Este estudio trata de actualizar el trabajo de Blalock et al. (2008) evaluando las propiedades psicométricas que han sido reportadas en los instrumentos de actitud hacia la ciencia publicados entre los años 2004 y 2016. En conjunto, se puede afirmar que la calidad general de los instrumentos sigue siendo limitada y que se siguen perpetuando las prácticas de desarrollo y validación identificadas como inadecuadas e insuficientes en publicaciones pasadas. Además, la mayoría de los instrumentos de actitudes hacia la ciencia analizados presentan una falta de evidencias psicométricas y una ausencia de marco teórico. La toma de decisiones sobre cómo fomentar actitudes favorables hacia la ciencia derivada de los resultados obtenidos de instrumentos que muestran una falta de validez y fiabilidad y una pobre conceptualización teórica resulta muy arriesgada, lo que subraya la necesidad de adoptar diseños metodológicos más robustos en futuros estudios de validación de instrumentos para la evaluación del relevante constructo de actitudes hacia la ciencia, tratando de superar los aspectos negativos que reducen su validez y fiabilidad, así como la confianza que se puede depositar en los resultados derivados de su uso.

REFERENCIAS

Las referencias marcadas con un asterisco (*) indican estudios incluidos en la revisión sistemática.

*ABD-EL-KHALICK, F., SUMMERS, R., SAID, Z., WANG, S. y CULBERTSON, M. (2015). Development and large-scale validation of an instrument to assess arabic-speaking students' Attitudes toward Science. *International Journal of Science Education*, *37*(16), 2637–2663. https://doi.org/10.1080/09500693.2015.1098789

AJZEN, I. (1991). The theory of planned behavior. *Organizational Behavior and Human Decision Processes*, *50*(2), 179–211. https://doi.org/10.1016/0749-5978(91)90020-T

AYDENIZ, M. y KOTOWSKI, M. R. (2014). Conceptual and methodological issues in the measurement of attitudes towards science. *Electronic Journal of Science Education Electronic Journal of Science Education*, *18*(3), 1–24.

BENNETT, J., LUBBEN, F., HOGARTH, S. y CAMPBELL, B. (2005). Systematic reviews of research in science education: Rigour or rigidity? *International Journal of Science Education*, *27*(4), 387–406. https://doi.org/10.1080/0950069042000323719

BLALOCK, C. L., LICHTENSTEIN, M. J., OWEN, S., PRUSKI, L., MARSHALL, C. y TOEPPERWEIN, M. (2008). In pursuit of validity: A comprehensive review of science attitude instruments 1935–2005. *International Journal of Science Education*, *30*(7), 961–977. https://doi.org/10.1080/09500690701344578

BRECKLER, S. J. (1984). Empirical validation of affect, beha ior, and cognition as distinct components of attitude. J*ournal of Personality and Social Psychology*, 47, 1191- 1205.

BYBEE, R. W. (2013). *The case for STEM education: Challenges and opportunities*. Arlington: NSTA press.

DE PRO BUENO, A. y PÉREZ MANZANO, A. (2014). Actitudes de los alumnos de Primaria y Secundaria ante la visión dicotómica de la ciencia. *Enseñanza de Las Ciencias*, *32*(3), 111–132. https://doi.org/10.5565/rev/ensciencias.1015

DEVELLIS, R. F. (2003). *Scale development: Theory and applications.* Thousand Oaks, Calif: Sage Publications.

*DIJKSTRA, E. M., y GOEDHART, M. J. (2012). Development and validation of the ACSI: Measuring students' science attitudes, pro-environmental behaviour, climate change attitudes and knowledge. *Environmental Education Research*, *18*(6), 733–749. https://doi.org/10.1080/13504622.2012.662213

EAGLY, A., y CHAIKEN, S. (1998). Attitude structure. En D. Gilbert, S. Fiske, y G. Lindsay (Eds.), *Handbook of social psychology* (pp. 269-322). New York: McGraw-Hill.

EC (European Commision). (2015). *Science education for responsible citizenship*. Luxembourg.

EC (European Commision). (2016). *Horizon 2020. Monitoring report 2014*. Luxembourg.

FERNÁNDEZ-CÉZAR, R., PINTO-SOLANO, N. y MUÑOZ-HERNÁNDEZ, M. (2018). ¿Mejoran los proyectos de divulgación con experimentación la actitud hacia las clases de ciencias? *Revista de Educacion*, *381*(Julio-Septiembre), 285–307. https://doi.org/10.4438/1988-592X-RE-2017-381-389

FRASER, B. J. (1981). *Test of science-related attitudes*. Melbourne: Australian Council for Educational Research.

GASKIN, C. J. y HAPPELL, B. (2014). On exploratory factor analysis: A review of recent evidence, an assessment of current practice, and recommendations for future use. *International Journal of Nursing Studies*, *51*, 511–521. https://doi.org/10.1016/j.ijnurstu.2013.10.005

*GUZEY, S. S., HARWELL, M. y MOORE, T. (2014). Development of an instrument to assess attitudes toward science, technology, engineering, and mathematics (STEM). *School Science and Mathematics*, *114*(6), 271–279. https://doi.org/10.1111/ssm.12077

*HILLMAN, S. J., ZEEMAN, S. I., TILBURG, C. E. y LIST, H. E. (2016). My attitudes toward science (MATS): The development of a multidimensional instrument measuring students' science attitudes. *Learning Environments Research*, *19*, 203–219. https://doi.org/10.1007/s10984-016-9205-x

KAISER, H. (1970). A second generation Little Jiffy. *Psychometrika*, *35*, 401–415.

*KENNEDY, J., QUINN, F., y TAYLOR, N. (2016). The school science attitude survey: A new instrument for measuring attitudes towards school science. *International Journal of Research & Method in Education*, *39*(4), 422–445. https://doi.org/10.1080/1743727X.2016.1160046

*KIND, P., JONES, K. y BARMBY, P. (2007). Developing attitudes towards science measures. *International Journal of Science Education*, *29*(7), 871–893. https://doi.org/10.1080/09500690600909091

KLOPFER, L. E. (1971). Evaluation of learning in science. En B. S. Bloom, J. T. Kastings y G. F. Madaus (Eds.), *Handbook on summative and formative evaluation of student learning* (pp. 559–642). New York: McGraw-Hill.

LLORET-SEGURA, S., FERRERES-TRAVES, A., HERNÁNDEZ-BAEZA, A. y TOMÁS-MARCO, I. (2014). El análisis factorial exploratorio de los ítems: Una guía práctica, revisada y actualizada. *Anales de Psicología*, *30*(3), 1151–1169. https://doi.org/10.6018/analesps.30.3.199361

LIBERATI, A., ALTMAN, D. G., TETZLAFF, J., MULROW, C., GØTZSCHE, P. C., IOANNIDIS J. P. A., CLARKE, M., DEVEREAUX, P. J., KLEIJNEN, J. y MOHER, D. (2009). The PRISMA statement for reporting systematic reviews and meta-analyses of studies that Evaluate health care interventions: Explanation and elaboration. *Plos Medicine*, *6*(7), e1000100. https://doi.org/10.1371/journal.pmed.1000100

*MAHONEY, M. P. (2010). Students' attitudes toward STEM: Development of an instrument for high school STEM-based programs. *Journal of Technology Studies*, *36*(1), 24–34.

MARBÀ-TALLADA, A. y MÁRQUEZ, C. (2010). ¿Qué opinan los estudiantes de las clases de ciencias? Un estudio transversal de sexto de primaria a cuarto de ESO. *Enseñanza de Las Ciencias*, *28*(1), 19–30.

MUNBY, H. (1983). *An investigation into the measurement of attitudes in science education*. Columbus, OH: SMEAC Information Reference Center, The Ohio State University.

*NAVARRO, M., FÖRSTER, C., GONZÁLEZ, C. y GONZÁLEZ-POSE, P. (2016). Attitudes toward science: Measurement and psychometric properties of the test of science-related attitudes for its use in Spanish-speaking classrooms. *International Journal of Science Education*, *38*(9), 1459–1482. https://doi.org/10.1080/09500693.2016.1195521

NRC (National Research Council). (2011). *Successful K-12 STEM education: Identifying effective approaches in science, technology, engineering, and mathematics. National Academies Press*. Washington, DC: The National Academies Press.

NRC (National Research Council). (2012). *A framework for K-12 science education: Practices, crosscutting concepts, and core ideas*. Washington, DC: National Academy Press.

NRC (National Research Council). (2013). *Next generation science standards: For stated, by states*. Washington, DC: The National Academies Press.

NUNNALLY, J. C. (1978). *Psychometric theory*. New York: McGraw-Hill.

OSBORNE, J., SIMON, S. y COLLINS, S. (2003). Attitudes towards science: A review of the literature and its implications. *International Journal of Science Education*, *25*(9), 1049–1079. https://doi.org/10.1080/0950069032000032199

*OWEN, S. V., TOEPPERWEIN, M. A., MARSHALL, C. E., LICHTENSTEIN, M. J., BLALOCK, C. L., LIU, Y., PRUSKI, L. A. y GRIMES, K. (2008). Finding pearls: Psychometric reevaluation of the Simpson-Troost attitude questionnaire (STAQ). *Science Education*, *92*(6), 1076–1095. https://doi.org/10.1002/sce.20296

POLIT, D. F. (2015). Assessing measurement in health: Beyond reliability and validity. *International Journal of Nursing Studies*, *52*(11), 1746–1753. https://doi.org/10.1016/j.ijnurstu.2015.07.002

POLIT, D. F. y YANG, F. (2016). *Meaurement and the measurement of change: A primer for health professionals*. Philadelphia: Lippincott Williams & Wilkins.

POTVIN, P., y HASNI, A. (2014). Interest, motivation and attitude towards science and technology at K-12 levels: A systematic review of 12 years of educational research. *Studies in Science Education*, *50*(August 2016), 85–129. https://doi.org/10.1080/03057267.2014.881626

*PUVIRAJAH, A., VERMA, G., LI, H. y MARTIN-HANSEN, L. (2015). Influence of a science-focused after-school program on underrepresented high-school students' science attitudes and trajectory: A survey validation study. *International Journal of Science Education Part B-Communication and Public Engagement*, *5*(3), 250–270. https://doi.org/10.1080/21548455.2014.930210

RYAN, A. G. y AIKENHEAD, G. S. (1992). Students' preconceptions about the epistemology of science. *Science Education*, *76*, 559–580.

*TYLER-WOOD, T., KNEZEK, G. y CHRISTENSEN, R. (2010). Instruments for assessing interest in STEM content and careers. *Journal of Technology and Teacher Education*, *18*(2), 341–363.

VÁZQUEZ, A. y MANASSERO, M. A. (1995). Actitudes relacionadas con la Ciencia: una revisión conceptual. *Enseñanza de Las Ciencias, 13*(3), 337-346.

VÁZQUEZ, A. y MANASSERO, M. A. (2008). La vocación científica y tecnológica de las chicas en secundaria y la educación diferenciada. *Bordón. Revista de Pedagogía, 60*(3), 149–163.

VÁZQUEZ, A. y MANASSERO, M. A. (2009a). La relevancia de la educación científica: actitudes y valores de los estudiantes relacionados con la Ciencia y la Tecnología. *Enseñanza de Las Ciencias, 27*(1), 33–48.

VÁZQUEZ, A. y MANASSERO, M. A. (2009b). Factores actitudinales determinantes de la vocación científica y tecnológica en secundaria. *Cultura y Educacion, 21*(3), 319–330. https://doi.org/10.1174/113564009789052280

*VILLAFAÑE, S. M. y LEWIS, J. E. (2016). Exploring a measure of science attitude for different groups of students enrolled in introductory college chemistry. *Chemistry Education Research and Practice*, *17*, 731–742. https://doi.org/10.1039/c5rp00185d

*WANG, T. y BERLIN, D. (2010). Construction and validation of an instrument to measure Taiwanese elementary students' attitudes toward their science class. *International Journal of Science Education*, *32*(18), 2413–2428. https://doi.org/10.1080/09500690903431561

*ZHANG, D. y CAMPBELL, T. (2011). The psychometric evaluation of a three-dimension elementary science attitude survey. *Journal of Science Teacher Education*, *22*, 595–612. https://doi.org/10.1007/s10972-010-9202-3

Systematic review of attitude toward science instruments (2004-2016)

Radu Bogdan Toma
Departamento de Didácticas Específicas
Universidad de Burgos, Burgos, España.
rbtoma@ubu.es

The purpose of this study was to review the psychometric properties of attitude toward science instruments published in the last thirteen years in order to determine whether educational proposals that are being promoted internationally for the improvement of science education are complemented with the development of valid and reliable self-report instruments. This analysis, that builds on and update Blalock et al.'s (2008) work, may help researchers to choose adequate instruments according to their study needs.

To this end, a systematic review following the procedure described by Bennett, Lubben, Hogarth and Campbell (2005) and the PRISMA statement (Liberati et al., 2009) was performed. Potentially relevant articles were retrieved from the *Web of Science Core Collection*, *Science Direct* databases, and using a snowball technique, which consists of examining the reference list of selected articles for relevant studies not identified through the databases. Searches were restricted by research category (Education and educational research), document type (articles), language (English and Spanish) and year of publication (2004-2016). Articles where deemed relevant according to the following inclusion criteria: (I) instrument development and validation studies; (II) focused on the attitudes towards science construct; (III) developed through classical test theory; and (IV) quantitative in nature.

The psychometric properties of each instrument was assessed using a scoring rubric consisting of four parts and with a score ranging from 0 to 27: (I) Theoretical framework -to what extent the instrument was developed according to existing attitudinal theories-; (II) Reliability -whether internal consistency, test-retest and standard measurement error results are reported-; (III) Validity -whether content, construct, convergent, concurrent, discriminant, discriminative and predictive validity results are reported-; and (IV) Dimensionality - whether data was analyzed according to the unidimensionality or multidimensionality of the instrument-.

A total of 14 studies providing information on the development and validation of 15 instruments were retained for in-depth analysis. Overall, 9 (60%) instruments scored less than half of the total rubric score (< 13). A total of 11 (73%) instruments were developed without a theoretical framework. In terms of reliability, no study reported psychometric evidence for test-retest or information about the standard error of measurement, and only four (27%) of the instruments reported robust levels of reliability. As for the validity, up to 9 (60%) instruments reported only one or two psychometric evidences (mostly content or construct validity), and no instrument was subjected to more than four psychometric validity tests.

These results indicate that Blalock et al.'s (2008) and Munby's (1983) criticisms are still valid for current attitudes towards science instruments. These results calls into question the confidence that can be placed in the results derived from studies using instruments whose validity and reliability are at stake and calls for a reconceptualization of the attitude toward science construct and the adoption of rigorous validation procedures to develop valid and reliable measurement instruments that would help to support (or refute) the assumptions and consensus reached in this line of research.

Appendix 4: Beyond attitude instruments systematic review

This appendix includes the full-text of the in-review article about the second systematic review conducted in this dissertation.

Toma, R. B., Lederman, N. G., & Meneses Villagrá, J. A. (under peer review). Still on the wrong path: A systematic review and evaluation of attitude test development practices in science education.

Still on the Wrong Path: A Systematic Review and Evaluation of Attitude Test Development Practices in Science Education

Toma, Radu Bogdan[1]

Lederman, Norman. G.[2].

Meneses-Villagrá, Jesús Ángel[1]

[1]University of Burgos

[2]Illinois Institute of Technology

Funding: This study has received funding from the 2017-2021 edition of the University of Burgos PhD research fellowship and by the Spanish Ministry of Economy, Industry and Competitiveness (MINECO) through the project EDU2017-89405-R. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Conflict of Interest: The authors declare that they have no conflict of interest.

Still on the Wrong Path: A Systematic Review and Evaluation of Attitude Test Development Practices in Science Education

In recent years, there is a growing body of research on developing attitudes assessments in science education. In this study, we propose an updated framework for the development and validation of quantitative measurement instruments and we use it to examine whether recent attitude toward science assessments are robust in terms of theoretical conceptualization, item generation practices and psychometric properties. Following a systematic review procedure, a total of 62 attitude assessments published between 2004 and 2018 years were retrieved from the Web of Science and Scopus databases and in-depth reviewed. Overall, the results show that rigorous and psychometrically valid and reliable attitude assessments are still lacking in science education research. Most reviewed assessments lacked theoretical foundation, were mostly develop upon existing instruments with conceptually poor definition of the attitude construct and were limited in terms of validity and reliability psychometric evidence. Additionally, several methodological and transparency issues were identified, such as poorly informed decisions for item development, misapplication of psychometric tests, and a lack of clarity or missing information in the reporting of the results. Best practices and those that need improvement are discussed with a focus on offering guidelines for future research.

Keywords: validity, reliability, psychometric properties, development, validation

Introduction

Test development and validation process is complex and requires the examination and scrutinise of several aspects related to item generation, theoretical analysis and psychometric evaluation. Without proper understanding of these aspects and an adequate training, test developers may not follow informed decision when developing assessments, often replicating weak or inconsistent procedures reported in other test validation studies of poor methodological quality. Reviews of published attitudes toward science assessments have shown that researchers fail to conduct and report critical psychometric analyses (Aydeniz & Kotowski, 2014; Blalock et al., 2008; Munby, 1983; Potvin & Hasni, 2014). The results of studies on students' attitudes in science education and on the evaluation of scientific-driven educational intervention strongly depend on the quality of the assessments used. Marshall et al. (2000) empirically showed that studies results can differ when the assessments used are unreliable and not valid. Therefore, if measurement instruments with a lack of theoretical framework and poor psychometric properties are being used, the quality and trustfulness of the results in science education research are at stake and remain unclear. As Munby (1983) rightly stated, "Unless proper care is taken to determine the performance characteristics of a test used in research, (…) the results themselves cannot be considered trustworthy" (p. 162).

Despite the proliferation of attitude assessments in the science education research field in recent years, an analysis of the quality of assessments published after Blalock et al. (2008) review is missing. For example, Aydeniz & Kotowski (2014) reported some conceptual and methodological issues in five commonly used assessments for measuring attitudes toward science, but their study was not systematic, nor exhaustive, and no inclusion criteria for the reviewed assessments were given. Potvin & Hasni (2014) provided information about only some attitudes assessments used in the literature, however, their analysis was descriptive and did not examined the assessments quality or their psychometric properties.

Consequently, the primary goal of this study is to examine current practices in the development and validation of attitudes assessments in science education. To this end, traditional and recent recommendations with respect to decisions involved during the test validation process are examined and a guiding framework that includes these established informed practices is advanced. Subsequently, a systematic review is performed. Using the proposed framework as a quality appraisal, assessments published between 2004 and 2018 are described and assessed in terms of item generation practices, and evidence of theoretical and psychometric properties analysis. Those aspects reflecting best practices and those needing improvements are highlighted and implications for science education research are discussed, with a special focus on offering recommendations to assist scholars in the field of science education research to develop attitudes toward science assessments that are consistent with recent test development standards.

A case for attitudes toward science research

Science education has a dual purpose and role. On the one hand, the education of individuals about science, its processes and procedures, and its appreciation is a central focus. On the other hand, it deals with the necessary training and education of the future generations of scientists that eventually will ensure the economic and technological development of countries (Millar & Osborne, 1998; Tytler & Osborne, 2012). Therefore, for the achievement of a democratic decision-making society, the development of scientifically literate citizenship is the desired outcome of science education (Laugksch, 2000; Millar, 1996). Thus, the promotion of interest in science and the development of positive attitudes towards science is one of the main components included within the scientific literacy agenda (OECD, 2006).

Likewise, the importance of attitudes towards science is reflected in the literature reporting a strong relationship between enjoyment of science with expressed intentions to further engage in science learning experiences (Ainley & Ainley, 2011a; 2011b), and in the positive relationship between interest in science and performance and achievement (Bybee & McCrae, 2011; Freedman, 1998; Newell, Tharp, Vogt, Moreno & Zientek (2015).

Given that students' attitudes toward science may affect their science performance, their desire to study science in the future, and the choice of science as a career (Koballa, 1988), attitudes have perennially been a very significant line of research in science education, especially in recent years given the steady decline of individuals interested in studying science in secondary and tertiary education (Archer et al., 2010; Kennedy, Lyons, & Quinn, 2014; Potvin & Hasni, 2014). This situation may affect economic stability and global security both because of the risk associated with not meeting the growing need for a STEM workforce, and also because scientific knowledge and interest in science have been identified as factors that impact people support of government-funded research (Besley, 2016; Miller, 1983; Motta, 2019).

Defining Attitudes

Defining attitudes is not an easy task, as they are not directly observable or measurable. Despite the existing of multiple different conceptualizations, there is a general agreement of conceiving attitudes as a multifaceted rather than unidimensional construct, which encompass cognitive, affective and behavioural aspects (Khine, 2015; Simpson, Koballa, Oliver, Crawley, 1994). Shaw and Wright (1968) considered that beliefs about a particular object (cognitive component) leads to affective reactions towards that particular object (affective component) and that this may affect the behaviour towards that object (behavioural). Therefore, for these authors, there is a three-stage process that goes from the cognitive (beliefs), then to affective (affective reactions), and then to behavioural (predisposition to act in one way or another). Oppenheim (1992) associated the cognitive aspect to beliefs and images about a certain object and the affective component, much more stable and profound than the cognitive one, to the values related to that object. Therefore, attitudes are conceived as individual's evaluations (Ajzen, 1991), and that these favourable or unfavourable evaluations are formed as individuals develop beliefs about the object under study (Fishbein & Ajzen, 1975). In words of Eagly & Chaiken (1995), attitudes can be defined as “(…) a psychological tendency that is expressed by evaluating a particular entity with some degree of favor or disfavor" (p. 414). Thus, attitudes encompass those evaluative aspects of a cognitive, affective and behavioral nature referring to beliefs, thoughts, feelings and emotions towards a given object (e.g. science) or a behavior related with the object (e.g. studying science).

Defining and Measuring Attitudes in Science Education

Although in recent decades studies focused on diagnosing students' attitudes toward science have proliferated considerably, the concept under study remains unclear. Indeed, Osborne, Simon, & Collins (2003) stated that attitudes toward science are “(...) somewhat nebulous, often poorly articulated and not well understood” (p. 1049), a problem that has been warned since more than three decades (Munby, 1983). In this regard, Klopfer (1971) offered insight about the concept under study by categorizing the attitude toward science construct as a set of affective behaviors toward science as an enterprise, scientists, scientific inquiry, scientific careers and towards science-related activities in general. Similarly, Schibeci (1983) concluded that attitude toward science deals with several attitude objects, such as science as a discipline, the study of science, science lessons, among many others. Gardner (1975) provided further clarity by addressing the differences between “scientific attitudes”, conceived as those elements inherent to scientific thinking and research, and “attitudes toward science”, conceptualized as the sociological, psychological and affective conceptions and beliefs about science. In short, attitudes comprise various aspects, such as science in general, the subject of school science, scientific careers, science teachers or the teaching of science, with each one of these aspects being characterized by many more detailed facets (Kind, Jones, & Barmby, 2007).

Consequently, past studies focusing on attitudes toward science have incorporated many constructs such as anxiety toward science, self-esteem in science, motivation toward science, enjoyment of science (Tytler & Osborne, 2012), students affective feelings and cognitive judgments of science (e.g. Zhang & Campbell, 2011), unfavorable outlook of science (e.g. Abd-El-Khalick, Summers, Said, Wang, & Culbertson, 2015) and perception of scientists and value of science to society (e.g. Hillman, Zeeman, Tilburg, & List, 2016), among many others. Recently, few developed assessments of attitudes towards science based on social psychological theories like the Theory of Planned Behavior (TPB, Ajzen, 1991) or the Expectancy-value theory (EVT, Eccles & Wigfield, 1995; Eccles, Wigfield, Harold, & Blumenfeld, 1993; Wigfield & Eccles, 1992, 2000, 2002) to further explore student’s motivation, choice and persistence in studying science-related subjects and careers and others behavioral beliefs about the consequences of engaging in science (Abd-El-Khalick et al., 2015; Said, Summers, Abd-El-Khalick, & Wang, 2016; Summers & Abd-El-Khalick, 2018), thus broadening the definition of attitudes towards science even more. Therefore, while attitudes are still the main focus of research, new models for analyzing and predicting behavior -like Ajzen's (1991) TPB- are receiving an increasing interest in later years to understand individuals career enrollment pattern (Tytler & Osborne, 2012), even though an consensual conceptualization of what constitutes attitudes toward science still does not exist. Thus, the access to valid and reliable instruments focused on measuring these attitudes and their interaction with other educational factors and decision-making in relation to career selection is of great importance.

Attitude Conceptualization Adopted in this Study

Because of the multifaced conceptualization of attitudes towards science, there is a need to first provide an operational definition of the concept under study before searching and selecting tests. A reasonable approach would be to assess the psychometric properties of those assessments that are developed according to existing theories or conceptualizations related to well established and investigated constructs such as self-efficacy or motivation. However, this approach would not be practical since attitude research in science education tend to lump together these and many others constructs under the attitude toward science umbrella without providing a working definition of what constitutes attitudes or what is being measured. Consequently, given the multifaceted conceptualization of attitudes toward science, any proposed definition of the trait under study should be as comprehensive and holistic as possible to be operative and include those assessments measuring the many facets that are swarmed together as attitude toward science constructs. Therefore, in this study a broad conception of attitudes toward science is embodied, and attitudes toward science are conceived as *positive and negative cognitive and affective evaluative judgements of science, science related disciplines, science related studies or careers, its agents (e.g. scientists; science teachers), its processes, and its products stemming from individuals underlying views, opinions, images and motivations related to knowing, enjoying or engaging in scientific-related issues.* Following this conceptualization, tests measuring attitudes, views, opinions, images and motivations related to the scientific enterprise (i.e. science, scientists, scientific careers and school subjects, science learning, science teaching, etc.) are considered to be measuring attitude toward science.

Test Development and Validation paradigms

The test development and validation process require complex and systematic procedures of both theoretically and methodologically consistency. In the literature there are numerous classical and contemporaneous approaches that guided the development of many self-administration assessments in the social sciences (e.g. Arias, Lloreda, & Lloreda, 2014; Clark & Watson, 1995; DeVellis, 2017; Keszei, Novak, & Streiner, 2010; Morgado, Meireles, Neves, Amaral, & Ferreira, 2017; Nunnally & Bernstein, 1994). For example, Churchill's pioneering paradigm (1979) proposed an 8-step procedure in which the construct under study is first defined, items are generated, data are collected, items and number of dimensions are purified through exploratory factor analysis (EFA), and data are collected again to calculate reliability, validity and other statistics of interest such as means and standard deviations for each item. Benson & Clark (1982) established a four-phase procedure consisting of 13 total steps. In the first phase, test developers formulate the statement of purpose of the proposed instrument and identify potentially relevant items by literature review and open-ended questionnaires to the target group. In the second phase, an initial pool of items is developed, and content validity is established. In the third phase, a pilot quantitative evaluation is performed by assessing item-related validity indices and reliability, which lead to a revised instrument that is administered in the fourth and last phase in which further validity and reliability is assessed. Anderson & Gerbing (1988) proposed a renewed paradigm in which the unidimensionality of the instrument is examined using confirmatory factorial analysis (CFA) after having performed exploratory factor analysis (EFA) and having established the reliability of the scale. Hinkin's (1995) review of scale development practices established three phases: in the first phase, items are generated; then, the scale is administered, and its structural validity and reliability are studied; and finally, in the third phase, the criterion validity of the instrument is established by examining the existence of a nomological network with other variables of interest. More recently, DeVellis (2017) proposed an eight-step procedure, which involves very similar task as classical recommendations.

While the development and validation of assessments does not follow a universal protocol, the American Educational Research Association, American Psychological Association, and the National Council on Measurement in Education (AERA et al., 2014) established the *Standards for Educational and Psychological Testing* to guide educational measurement efforts. Consistent with these and classic recommendations, the development and validation of an assessment can be conceived as a three-phase process. Figure 1 shows the process proposed in this study for developing and validating a test. This framework combines recommendations related to the test development process with the most recent psychometric properties taxonomies.

[Insert Figure 1 here]

During the first phase, named here as "item generation", test developers should adopt a theoretical framework and develop or select items accordingly using deductive, inductive, or both methods. When a deductive strategy is used, items are selected and sometimes adapted from literature reviews or existing assessments. When using inductive methods, new items are generated by the test developers, sometimes based on empirical evidence stemming from interviews or focus groups with the target population under study or panel of experts. In addition, during this initial stage researchers also address other parameters related to the initial pool of items, such as clear and simple item redaction, response format, variability of responses, or administration instructions (DeVellis, 2017).

In the second stage, called here "theoretical analysis", the content validity should be examined by analyzing the relevance and representativeness of the items in relation to the theoretical construct. Ideally, content validity is established by involving both a panel of experts and a panel of the target population. Panel of experts are asked to assess the relevance of the items and if they appropriately represent the construct under study, and the target population (i.e. potential test takers) are used to examine the understanding, difficulty and interpretation of the initial pool of items.

In the third and last stage, named here "psychometric evaluation", the newly developed assessment is administered to a large sample and the psychometric properties are examined to establish the reliability and validity of the assessment. To this end, test developers must subject the assessment to a series of rigorous analyses to provide evidence that the proposed assessment is free of measurement error and that it measures the construct it is supposed to be measuring (Scholtes, Terwee, & Poolman, 2011). Different psychometric indices are used to assess whether a test is of high quality based on reliability and validity measurement properties.

As measurement concepts have evolved, and new ideas related to psychometric properties have been developed, different terminology and definitions of measurement properties are mounting, making the reporting of validity and reliability evidences a difficult and confusing task. In order to clarify and standardize the use of the different terms related to psychometric properties, Mokkink et al. (2010) developed the COnsensus-based Standards for the selection of health Measurement Tests (COSMIN) taxonomy through an international Delphi study. The COSMIN initiative reached consensus on which measurement properties are relevant, what are the terminology and definition of these measurement properties, and what are the design, requirements and preferred statistical methods for assessing these properties (Mokkink, Terwee, Knol, et al., 2010; Mokkink, Terwee, Patrick, et al., 2010). More recently, building on the COSMIN study, Polit (2015) and Polit & Yang (2016) proposed an updated taxonomy of psychometric properties. In the next subsection, this validity and reliability taxonomy that we adopted in our review is introduced.

Validity

In relation to the validity domain, the Polit-Yang taxonomy adopts the COSMIN conceptualizations which define validity as the degree to which a test "(…) measures the construct(s) it purports to measure" (Mokkink et al. 2010, p.743). There are three components related to the validity domain: (a) content and face validity, (b) criterion validity, and (c) construct validity.

Content and face validity examines if the content of the test adequately measures the construct of interest. More specifically, face validity examines what the test appears to be measuring based on subjective evaluations related to whether the test looks as thought to be measuring the target construct. Similarly, content validity evaluates the "adequacy of content coverage and relevance for multi-item measures of a construct" (Polit, 2015, p.1750), but following a much more rigorous and systematic procedure. Although face validity is less advisable than content validity, it is usually first used with a group of experts to find critical flaws that can be identified and purified during initial stages of the test development (i.e. too explicit items; items not related with the target construct).

Criterion validity measures the extent to which the score of the proposed assessment reflects equivalent results as a gold-standard assessment measuring the same construct. In other words, an assessment has criterion validity when it is highly correlated with a previously validated assessment measuring the same construct and when it can predict something that it should theoretically predict (Trochim & Donnelly, 2006).

Construct validity examines if the assessment scores "(…) supports the inference that the construct has been appropriately represented" (Polit, 2015, p. 1750). There are three subtypes of construct validity properties, mainly (a) structural validity, (b) hypothesis testing validity, and (c) cross-cultural validity. Structural validity examines if the assessment captures the theoretically expected dimensionality using exploratory factor analysis (EFA) and confirmatory factor analysis (CFA). Hypothesis testing validity examines whether scores on the focal measure are related to constructs with which conceptual convergence is expected (convergent validity), are not related to constructs with which conceptual convergence is not expected (discriminant validity), and whether the measures can discriminate between groups that are known to differ (discriminative validity). Finally, cross-cultural validity is used with translated assessments, which involves translation and back-translation and equivalent testing's between the original and the translated version so that the meaning of the assessment is maintained and are linguistically and culturally appropriate to new context (Borsa, Damásio, & Bandeira, 2012; Cha, Kim, & Erlen, 2007; Haccoun, 1983; Robichaud-Ekstrand, Haccoun, & Millette, 1994).

Reliability

In the Polit-Yang taxonomy, reliability is defined as whether "(…) scores for people who have not changed are the same for repeated measurements, under several situations" (Polit & Yang, 2016, p. 25). This domain includes three types of reliability: (a) temporal reliability, which measures temporal stability or test-retest reliability of the assessment when trait have not changed (i.e. if scores are stable), (b) internal consistency, concerning whether items on a test are part of the same underlying construct, measured through Cronbach alpha, item-total correlation and split-half reliability, and (c) measurement error, which relates to the error in a participant score that is not related to true changes in the trait under study.

Method

Systematic Reviews

Petticrew (2001) categorized literature reviews as "narrative reviews" or "systematic reviews". A systematic approach differs from a narrative review in the use of a structured, transparent, and readily reproducible approach to finding and reviewing of existing literature. While narrative reviews synthesize existing literature through the authors' perspective, systematic reviews adopt a more objective stance that allows readers to make their own judgments about the quality and significance of the evidence presented (Bearman et al., 2012). Thus, systematic reviews differ from narrative reviews in that they are less prone to research bias by introducing more objectivity through detailing rigorously and explicitly the process for selecting the studies and the method used to review and evaluate them. Therefore, systematic reviews tend to be more comprehensive and reproducible than traditional revision(Andrews, 2005). In this sense, systematic reviews use a specific, transparent, and carefully defined approach to literature search, selection, and synthesis (Bearman et al., 2012; Judith Bennett, Dunlop, Knox, Reiss, & Torrance Jenkins, 2018). Systematic reviews are characterized by key aspects such as (a) the establishment of predetermined objectives, (b) predefined eligibility criteria for the inclusion and exclusion of studies, (c) explicit and reproducible search and synthesis methodology, (d) systemic and comprehensive search with the aim of including all studies that meet the inclusion criteria and thus reduce the risk of bias, and (e) systematic presentation and synthesis of the characteristics and findings of the included studies (Andrews, 2005; Gough, 2007; Higgins, Churchill, Chandler, & Cumpston, 2011).

Systematic reviews have gained popularity and acceptance in recent years, and several initiatives promoting the use of systematic reviews have been established, such as the Campbell Collaboration, focused on the review of randomised controlled trial studies in education, criminology and other social sciences, or the Evidence for Policy and Practice Initiative Centre (EPPI-Centre) for the systematic review of existing literature in education (Bennett, Lubben, Hogarth, & Campbell, 2005). Advocates of systematic reviews emphasize its use as a source to inform policy decisions and changes in the area of interest (Bennett, Lubben, & Hogarth, 2007). Synthesized findings derived from systematic reviews could reveal existing strengths and deficiencies in current research and provide evidence and directions to inform practices for future research (Evans & Benefield, 2001). Consistent with the objectives of systematic reviews, this study aims at enabling researchers to quickly become acquainted with the existing literature in the area of science education research with respect to the development and validation of attitude assessments, to examine the reliability and validity of existing assessments by identifying the specific strengths and weaknesses of each one of them, and to report on the implications that these results have for future assessment development studies. In the next sections, the main features of systematic reviews are reported for the present study, following the PRISMA statement for transparent reporting of systematic reviews and meta-analyses (Liberati et al. 2009).

Review Aims

The aims of this systematic review are threefold:

O1. Identify attitude assessments developed and validated in science education research, determining the evolution of the production of attitude assessments, their characteristics and structure, and the context in which they have been developed and validated.

O2. Examine the extent to which attitude assessments have been submitted to psychometric properties tests for providing validity and reliability evidence.

O3. Identify any conceptual and methodological issues related to the development and validation of attitude assessments in science education that can undermine their validity and reliability.

Inclusion and Exclusion Criteria

Studies were deemed relevant according to the following inclusion criteria:

(a) Test validation study. The first criterion was set to encompass only studies which reports on the development and validation of assessments. Therefore, studies using attitudes assessments but not specifically focused on the development and validation of the assessment were excluded, regardless of whether validity and/or reliability evidence were provided.

(b) Assessments measuring attitudes according to the attitude toward science definition adopted in this study. The second criterion was established to contextualize the search and selection of the articles. To this end, the comprehensive definition reported in previous sections was established and used for the selection and inclusion of the studies.

(c) Assessments must be quantitative. The third and last criterion is consistent to the framework for test development and validation adopted. Since many of the indices used for assessing validity and reliability are quantitative in nature or require quantitative data (e.g. EFA or Cronbach α), it was decided to exclude qualitative assessments based on open-ended questions or interviews protocols.

Search Strategy

Potentially relevant articles were identified in the Web of Science Core Collection (WOS) and Scopus database. The selection of these two databases is consistent with recommendations for developing systematic review advising the use of more than one database in order to reduce risk of study selection bias. One of the main characteristics of systematic reviews is the first broad search that is subsequently analysed in order to select articles considered as relevant according to the aims of the study and inclusion criteria. Therefore, study selection bias can be introduced by including only articles indexed in one database or published in specific journals, which would affect the quality and the results arising from a systematic analysis of the existing literature.

The complete search strategy is shown in Table 1. Consistent with the working definition of attitudes toward science adopted in this systematic review, the terms used in the search included *attitude*, *view*, *opinion*, *image*, and *motivation*. Consistent with the Standards for Educational and Psychological Testing (AERA et al., 2014) use of terms to refer to psychological tests, the following terms were used: *scale*, *instrument*, *measure*, *survey*, *questionnaire*, and *tool*. Finally, up to five different terms related to the development and validations of assessments were used. Additionally, a snowball technique was used by screening the reference list of eligible studies to include potentially relevant articles that were not found using the search strategy.

[Insert Table 1 here]

Search was restricted by the following filters: (i) articles, (ii) written in English or Spanish, (iii) indexed in the Education & Educational Research (in WOS) and Social Science categories (in Scopus), and (iv) published since 2004 to June 2018. The time period established builds on and updates Blalock et al.'s (2008) systematic review of science education assessments that included studies published up to 2004 year. Literature searches were performed on June 7th, 2018.

Data Abstraction and Quality Appraisal

The included assessments were summarized using a data abstraction sheet covering information about the reference, year of publication, language, target population, number of items and constructs, and name of the constructs included.

The psychometric quality of the attitude assessments included was assessed using a checklist based on the framework for test development and validation described in previous sections (Supplemental File 1). It consists of 17 aspects that determine strong methodological quality standards based on evidences related to theoretical framework adopted (one aspect), item development strategy used (two aspects), reliability (five aspects) and validity (ten aspects). For translated assessments, it includes additional validity evidence (i.e. cross-cultural validation). Indeed, each assessment was examined to determine if it (i) was developed based on an existing theoretical framework, (ii) if items were developed following a deductive or inductive approach, (iii) if content validity was assessed based on expert panels and target population of study, and (iv) if psychometric evaluation reported evidence for validity and reliability. A descriptive analysis of the psychometric properties included in the studies was undertaken to identify the psychometric properties that are most and least used during the development and validation process, and the strengths and weaknesses of each attitude assessment based on their reported psychometric evidence. Additionally, based on the Standards for Educational and Psychological Testing (AERA et al., 2014), studies included in this systematic review were in-depth analyzed to detect several aspects that can potentially undermine the quality of the published attitude assessments were examined (e.g. administration instructions or lack of it; misuse of psychometric tests; ambiguity in reporting results; use of psychometric terms).

In line with the literature on test validation paradigms and the revised taxonomy, an instrument will be deemed of high quality if developed following the standards and procedures summarized, all of which are considered essential by the *Standards for Educational and Psychological Testing* (AERA et al. 2014). Therefore, high quality assessment are expected to be (i) rooted in a theoretical framework, (ii) tested against content validity using both a panel of experts and a target population, especially when items where developed without using empirical-inductive methods, (iii) show evidence of low measurement error and high internal consistency, and temporal stability if intended to be used in longitudinal studies or intervention studies with pre-posttests designs, (iv) show evidence of at least one criterion validity, (v) show evidence of construct validity based on structural validity type, convergent and discriminant validity in multidimensional assessments, and (vi) in translated instrument, be adapted using cross-cultural procedures.

Results

Search Outcomes and Literature Identification

Using the search methods described above, 1327 records were retrieved, of which 1055 were excluded after applying search filters. Of the remaining 272 articles, 174 were excluded after reading title and abstracts for not meeting inclusion criteria. After removing duplicated records and searching for full articles copies, a total of 54 articles were deemed relevant for review from the Web of Science Core Collection and Scopus databases. After full-text reading and examination of the reference list of eligible studies, 7 more articles were added for in-depth analysis, thus comprising a total of 61 relevant studies reporting on the development and validation of 62 attitude assessments (Figure 2).

[Insert Figure 2 here]

General Characteristics

Source, age of publication, and language. The included studies were published in a total of 32 journals. The *International Journal of Science Education* and the *Journal of Research in Science Teaching* were the ones with more test validation studies in the time period used, with 12 and 6 publications per journal, respectively. Four assessments were published in the *Journal of Chemical Education* and three in the *Physical Review* journal. The rest of the attitude assessments were published in a great variety of journals (Supplemental File 2, Table S1).

Most assessments included in this review were published in 2016, followed by 2013 and 2011. The year with less publication of test validation studies is 2005 and 2006 with only 1 study each year, followed by 2007 and 2008 with only two studies per year (Supplemental File 2, Figure S1).

When considering the studies included in this review and that attitude studies reviewed by Blalock et al. (2008), several trends can be identified with peaks in the frequency of publication (Supplemental File 2, Figure S2). Since the 1940s there has been an increasing interest in the publication of attitude assessments in science education, reaching a peak in the number of publications in the 1970s, with a total of 27 studies per decade. Subsequently, from the 1970s to the 1990s, a decreasing interest in the development of assessments has been identified, with only 12 tests published in the 1980s and only 2 in the 1990s. Finally, since the 2000s there seems to be a new increase in the number of published attitudes tests, a trend that continues to rise until nowadays, with a total of 53 assessments published in the 2010th decade corresponding years. In other words, in the last 8 years, there have been published more attitude assessment validation studies than in the last 30 years combined, which shows that the development of attitude assessments in science education is on the rise more than ever.

The included assessments were developed and validated in 13 languages, which show a global interest in the measurement of attitudes in science education (Supplemental File 2, Table S2). Most of them were, nonetheless, developed for English-speaking individuals (38 out of 62, 61.3%). Up to 5 studies (8.1%) did not specify in which language the assessment was administered. Four of these studies used TIMSS data (i.e. Liou, 2014; Marsh et al., 2013; Sabah, Hammouri, & Akour, 2013; Tee & Subramaniam, 2018). The fifth study not specifying the language was developed in Taiwan (i.e. Chiang & Liu, 2014), however, test developers did not stated if the assessment was administered in Taiwanese of English language.

Aims of the assessments. An attempt was made to categorize assessments according to the aim and attitude construct for which they were developed and validated. However, this task has proved to be extremely difficult given the broad conceptualization of attitude existing in the science education literature. Therefore, based on the conceptualization of attitude of each test developer, five distinct categories were created (Supplemental File 2, Table S3). Thus, included attitude assessments measured a range of different aspects, including (a) attitude toward a specific discipline, like science or chemistry, (b) attitudes toward a specific scientific course (or the teaching of it), (c) motivation to science or science learning, (d) beliefs about specific disciplines, like biology or physics, and (e) others, like interest, emotions, engagement, self-efficacy or social influence, among many others. The categorization of assessments was further complicated by the existing of others sub-categories in each one of the main categories mentioned or by the fact that many assessments were found to be measuring more than one construct that pertain to different categories.

Attitude assessments have been identified at all educational levels (Supplemental File 2, Table S4). Most assessments were developed for being used with middle school students (22 out of 62, 35.5%). There are few assessments for in-service teachers and elementary school levels. One assessment was validated using adult population in general (i.e. Morgan, Collins, Sparks, & Welch, 2018). Some assessments were validated for multiple grades. For example, Abd-El-Khalick et al.(2015), Hillman et al. (2016), Lamb, Annetta, Meldrum, & Vallett (2012), Summers & Abd-El-Khalick (2018), and Unfried, Faber, Stanhope, & Wiebe (2015) assessments can be used in elementary, middle and high school..

Structure. The attitude assessments reviewed varied greatly in terms of length, with the number of factors ranging from 1 to 9 and the number of items ranging from a minimum of 4 to a maximum of 70 (Supplemental File 2, Table S5). Some assessments were unidimensional (i.e. Caputo, 2017; Fortus & Vedder-Weiss, 2014; Hiller & Kitsantas, 2016; Sjaastad, 2013; Tosun & Genç, 2015; Wang, Fredricks, Ye, Hofkens, & Linn, 2016; You, 2016), however, most assessments were multidimensional and thus were composed of several distinct factors. 39 out of 62 assessments included (63%) had more than 20 items. The shortest assessment was unidimensional and included only 4 items to measure attitude toward the learning of science subject (i.e. Caputo, 2017). The largest assessment was an adaptation of Frasers's TOSRA (1978, 1981) into the Spanish language, thus comprising a total of 7 factors and 70 items (i.e. Navarro, Förster, González, & González-Pose, 2016).

Conceptual and Methodological Quality

Theoretical framework and item generation. Up to 30 of the 62 reviewed assessments (48.4%) were developed without a clear conceptualization of the attitude construct and thus were not guided by any theoretical framework (Supplemental File 2, Table S6). Of the remaining 32 assessments referring to theory, some of them used strong theoretical foundation for their development, such as Ajzen (1991) Theory of Planned Behaviour (i.e. Abd-El-Khalick, Summers, Said, Wang, & Culbertson, 2015; Summers & Abd-El-Khalick, 2018), or the Eccles et al., (1983) Expectancy-Value theory (EVT) of achievement motivation (i.e. Andrews, Runyon, & Aikens, 2017; Oh, Jia, Lorentson, & LaBanca, 2013). However, most assessments used theory in a problematic way, like mixing theory constructs with other constructs (i.e. Andrews et al., 2017 used EVT theory but the expectancies for success construct was replaced by interest), or linking two theoretically distinct factors into the same construct (Unfried et al., 2015, mixed in one factor interest, self-efficacy and outcome expectancies items). Test developers not using a specific theoretical foundation often justified the development of the assessment by establishing links between previous literature findings and using the same constructs reported in already existing assessments (e.g. Ergün & Balçın, 2017; Hillman, Zeeman, Tilburg, & List, 2016; Kennedy, Quinn, & Taylor, 2016; Kind, Jones, & Barmby, 2007).

Almost all assessments (*n* = 56, 90.3%) were developed using deductive methods for item generation, mainly using items from already existing assessments (85.5%). Twenty of the 62 assessments (32.3%) used inductive methods to develop new items, with very few of them (*n* = 8, 12.9%) derived from empirical approaches such as interviews or open-ended questionnaires with the target population (e.g. Cheung, 2009; Glynn, Brickman, Armstrong, & Taasoobshirazi, 2011; Morgan, Collins, Sparks, & Welch, 2018; Wang & Berlin, 2010).

Psychometric quality. In general, the revised assessments have been submitted to low level of validation work. Although all assessments were presented as reliable and valid for the examination of attitudes, for the most part this was only supported by the reporting of few reliability and validity psychometric properties. An in-depth analysis of the psychometric assessments of each assessment is reported in Table 2.

[Insert Table 2 here]

Therefore, no study provided the maximum psychometric types of evidence, neither for reliability, nor for validity. 36 of the 62 assessments (58.1%) reported three or less out of nine validity evidence, with five assessments reporting only one validity evidence. Of the 62 assessments, only two (3.2%) presented a total of seven validity evidence (i.e. Caputo, 2017; Velayutham, Aldridge, & Fraser, 2011) and only four (6.5%) reported up to three reliability evidence (i.e. Adams, Wieman, Perkins, & Barbera, 2008; Dermitzaki, Stavroussi, Vavougios, & Kotsis, 2013; Tosun & Genç, 2015; Wilcox & Lewandowski, 2016). As for reliability, 41 of the 62 assessments (66.1%) reported only one out of five existing reliability evidences. Frequency statistics on psychometric properties evidences reported in the reviewed assessments are summarized in Table 3 and Table 4.

[Insert Table 3 here]

[Insert Table 4 here]

Validity evidences. Overall, test developers failed to provide consistent evidence for some types of validity psychometric properties. Thus, in up to 23 of the 62 assessments (41.9), evidence for content and face validity was inexistent, and for only 23 assessments (37.7%) the content validity was examined using interviews with the target population of study. As for criterion related validity, concurrent or predictive tests were missing in 49 of the 62 assessments (79%). The concurrent validity was examined for only three assessments (4.8%) and the predictive validity was assessed in only 10 out of the 62 analysed assessments (16.1%).

In relation to construct validity, all but two attitude assessments were submitted at least to one structural test, mainly EFA (*n* = 43, 69.4%) and CFA (*n* = 32, 51.6%). Eleven of the 62 assessments (17.7%) used Item Response Theory instead of Classical Test Theory to examine the structural validity of the proposed assessment (e.g. Fortus & Vedder-Weiss, 2014; Lamb, Annetta, Meldrum, & Vallett, 2012; Sha, Schunn, & Bathgate, 2015; Tee & Subramaniam, 2018). Information regarding hypothesis-testing validity was rather limited. 39 of the 62 assessments (62.9%) were not tested for convergent validity and the discriminant validity was not assessed in 49 assessments (79%). The ability of the proposed assessments to correctly discriminate between groups it should be able to discriminate (i.e. discriminative validity) was not reported for 32 of the 62 assessments (51.6%). As for translated assessments, most of them performed back-translation procedures for translating the items (n = 11, 78.6%), although only one was validated against the original one through equivalence testing (i.e. Bayar & Karamustafaoğlu, 2015). Descriptive statistics on validity properties are summarized in Table 5.

[Insert Table 5 here]

Reliability evidences. Reliability properties were reported for all assessments, however, most test developers failed to examine some fundamental types of reliability. Consequently, temporal stability or test-retest reliability was reported in only 7 out of the 62 assessments (11.3%, e.g. Bayar & Karamustafaoğlu, 2015; Xu & Lewis, 2011), however some of the method used for computing test-retest reliability were problematic and will be discussed in further detail in next sections. In addition, for internal consistency reliability, the Cronbach alpha was the test most used (n = 53, 85.5%), however, eleven assessments reported results below the minimum acceptable cut-off or alpha > .70 for exploratory research (e.g. Douglas, Yale, Bennett, Haugan, & Bryan, 2014; Hillman et al., 2016; Maier, Greenfield, & Bulotsky-Shearer, 2013; Marsh et al., 2013). Only two out of the 62 reviewed assessments (3.2%) were subjected to split-half reliability and the item-total correlation reliability was reported for only 10 (16.1%) attitude assessments. Some assessments used Item response Theory (IRT) tests for examining internal consistency instead of Classical test Theory (CTT) tests (e.g. Sabah, Hammouri, & Akour, 2013; Sjaastad, 2013; You, 2016). The internal consistency of one assessment was not reported (i.e. Semsar et al., 2011). Finally, the measurement error was reported in only 6 out of 62 assessments (9.7%, e.g. Adams et al., 2008; Liou, 2014; Tosun & Genç, 2015), so the extent to which the assessments adequately captures individuals attitudes remains somewhat unknown for the majority of the analysed assessments. Descriptive statistics on reliability properties are summarized in Table 6.

[Insert Table 6 here]

Other Issues

This review has also identified some aspects that limit the validity and reliability of the included assessments, but which have not, however, been warned by existing literature reviews.

Administration bias. Information regarding aspects that can be considered essential for the adequate use of the assessments (e.g. user's guide, training needs, time completion, and method of administration or administration instructions) was virtually non-existent. In some cases, test developers described the assessments as easy to administer and no time-consuming (e.g. Kennedy et al., 2016). However, since self-administration assessments rise concerns about several sources of bias, this information becomes necessary rather than trivial to ensure consistency in the administration of the assessments (for example, what should the examiner do when misunderstandings about items or response options arise?). The Standards for Educational and Psychological Testing (AERA et al., 2014) clearly stated it so: “The test developer should specify in clear language (…) the contexts in which test scores are to be employed, and the processes by which the test is to be administered and scored” (p. 23).

Social desirability, the tendency of individuals to report a positive image of themselves by offering responses that are compliant to socially acceptable standards (Johnson, Fendrich, & Mackesy-Amiti, 2012), is a serious threat to assessments validity (Huang, Liao, & Chang, 1998) and the information gathered (Adams et al., 2005; Van De Mortel, 2008), especially when examining sensitive topics (King & Bruner, 2000). A broad bias in teachers responses to the *Attitudes toward Science teaching test* (Korur, Vargas, & Serrano, 2016; van Aalderen-Smeets & Walma van der Molen, 2013) or the *Reform-Oriented Science Teaching practices test* (You, 2016) can be expected if individual privacy and confidentiality is not ensured, something that is very difficult to guarantee with paper-pencil assessments. The same may occur with assessments measuring students’ attitudes towards school science (e.g. Caputo, 2017; Cheung, 2009; Kennedy et al., 2016;). Social desirability responses might be expected from students if class teacher is involved in the assessment administration, maybe due to concerns of responses affecting academic results. The items “My science teachers are very good” (Abd-El-Khalick et al., 2015, p. 2653), “My science teachers make science interesting” (Lamb et al., 2012, p. 650), or “I participate in science courses so that the teacher pays attention to me” (Tuan et al., 2005, p. 654) are examples of socially sensitive items that may arise the described concerns. In these cases, assessment respondents may adapt their answer in a culturally approved-manner in order to obtain social approval and avoid criticism.

Misapplication of psychometric tests. Several misapplications of some psychometric tests were identified. For example, Semsar, Knight, Birol, & Smith (2011) examined test-retest reliability using two sub-samples from the same population cohort instead of responses from the same participants at two points in time: “(…) calculating a test-retest coefficient of stability on student responses from two equivalent populations (…) correlating all pre-instruction responses between students in the Fall 2007 and Fall 2008” (p. 273). This procedure is suitable for examining test stability, however it provides none information on the *temporal* (i.e. test-retest) stability. On the other hand, many test developers reported the Cronbach alpha for the entire attitude assessment even if it was an multidimensional assessment (e.g. Guzey et al., 2014; Korur, Vargas, & Serrano, 2016; Tuan, Chin, & Shieh, 2005). Given that alpha assumes unidimensionality of the assessment, using this test for an entire assessments with multiple factors violates this assumption (DeVellis, 2017; Tavakol & Dennick, 2011). Instead, each sub-dimension should be treated as a separate assessment, and therefore, the Cronbach alpha should be calculated for each of the constructs, without the need to report internal consistency results for the full assessment.

Likewise, the procedure used for concurrent validity presents some problematic aspects that need to be raised. For example, Mahoney (2010) examined the concurrent validity by correlating the results of the proposed assessment with a semantic differential test that was not previously validated in the literature and for which none psychometric properties was provided. Navarro, Förster, González, & González-Pose (2016) provided only concurrent validity evidence for two of the total seven constructs proposed, did not use all the items belonging to the gold standard assessment used, and the assessment used as gold standard lacked validity and reliability evidence, thus being wrongly considered a gold standard. Another misapplication was identified in Randler et al. (2011) study, which used a gold standard assessment measuring different constructs that the ones included in the proposed attitude assessment.

Lastly, poorly informed decision related to conducting structural validity tests like EFA were common in the tests reviewed, mainly in relation to the method used to decide factor extraction and retention. For example, Kind et al., (2007) decided to retain a 7-factor structure, despite the scree-test clearly revealing a structure composed of less factors. Puvirajah et al., (2015) retained a 3-factor structure, although scree-test also suggested otherwise. Some test developers retained factors with less than three items (e.g. Abd-El-Khalick et al., 2015; Glynn, Taasoobshirazi, & Brickman, 2009), something that is not suggested (Hair, Black, Babin, Anderson, & Tatham, 2010). Other test developers performed EFA with less than the smallest sample size of 5 individuals per item cut-off with a minimum of 100 cases (e.g. Oh et al., 2013; Romine et al., 2014; Tyler-Wood, Knezek, &Christensen, 2010) which can lead to misleading factor-structure (Henson & Roberts, 2006; Karami, 2015; Kline, 2014) and therefore the construct validity of these assessments is threatened. Due to all these misapplications, future studies addressing the (mis)use of structural validity tests in science education attitudes assessments are worth to be investigated.

Clarity and missing data. Test developers reported the validation results in no structured, clear and transparent format, making it difficult for readers to easily identify and judge the psychometric quality of the assessments. For example, psychometric properties were wrongly referred by different names, some psychometric test were performed without reporting enough information on the procedure followed or the results obtained, or some tests were used without explaining what type of validity or reliability it is measuring (e.g. performing interviews with students without referring to content validity). For example, Wilcox & Lewandowski (2016) erroneously argued that "Concurrent validity examines the extent to which E-CLASS scores are consistent with certain expected results" (p. 7), which is rather a definition of discriminative or predictive validity. As for missing or limited information about psychometric tests, some test developers used the scree-test for determining the numbers of factors to be extracted, however, the figure was not added in the text (e.g. Glynn et al., 2009). Other information not reported by nearly no study is related to the results of EFA analysis. When reporting the factor matrix, test developers only reported the loadings for the expected factor instead of all loadings (e.g. Cheung, 2009; Dermitzaki et al., 2013; Guzey et al., 2014; Kind et al., 2007; Navarro et al., 2016; Schumm & Bogner, 2016; Tuan et al., 2005; Unfried, Faber, Stanhope, & Wiebe, 2015). Some test developers argued that "For the sake of clarity factor loadings lower than .40 are omitted" (e.g. Bayar & Karamustafaoğlu, 2015, p.45); others simply reported loadings of the intended factors, omitting the rest of the loadings or cross-loadings without giving any reason. As raised in the EFA literature, EFA results are very much dependent to the procedure adopted by test developers, so readers should be able to evaluate researchers decisions when conducting EFA (Conway & Huffcutt, 2003; Fabrigar, Wegener, Maccallum, & Strahan, 1999; Ford, MacCullum, & Tait, 1986).

This requires that test developers using this procedure to be explicit about their practices and support their results by reporting extensive information. The lack of clarity identified in this review makes this task difficult, thus considerably limiting the scrutiny of the research community about the decisions made when developing and validating attitudes towards science assessments.

Synthesis and Discussion

Results of this review show that methodologically strong and psychometrically sound attitude assessments in science education are scarce. Among the reviewed assessments, only few were found to be somewhat consistent with recent recommendations for the development and validation of tests, thus most being of deficient quality and falling far short according to current standards. Nonetheless, the ones developed by Velayutham et al. (2011) and Cheung (2009) where found to have the best overall quality. As for Velayutham et al. (2011), authors developed a theoretically based motivation and self-regulation assessment for English-speaking students enrolled in grades 8 to 10. Psychometric evidence was found for just one reliability property (i.e. Cronbach alpha), but several tests were used to examine the validity of the attitude assessment, including content analysis with both a panel of experts and target population, construct validity (i.e. EFA, convergent, discriminant and discriminative) and criterion validity (i.e. predictive). As for Cheung (2009), the test developer developed an assessment to measure Chinese secondary school student's attitudes toward chemistry lessons. The test was rooted in the latent process viewpoint theoretical framework (Oskamp & Schultz, 2005) measuring cognitive, affective and behavioural processes. Several strategies were used to develop the items, including existing assessments, literature review and interviews with target population, and content validity was examined through an expert panel composed of secondary school teachers and target sample. Both Cronbach alpha and item-total correlation tests were used to demonstrate internal consistency reliability and validity was examined though construct validity (i.e. CFA, convergent and discriminative). Additional assessments with acceptable overall quality are the ones developed by Andrews et al. (2017), van Aalderen-Smeets & Walma van der Molen (2013), Wang, Fredricks, Ye, Hofkens, & Linn (2016) and Oh et al. (2013). All of them were developed using a theoretical-framework, examined content validity by interviewing individuals from the target population, and reported reliability evidence (i.e. Cronbach alpha), structural (e.g. EFA/CFA) and hypothesis tests for construct validity, (e.g. convergent, discriminant, discriminative).

Nonetheless, results of this review show that rigorous examination of the psychometric properties of attitude assessments is still lacking in science education research. In addition, some problematic issues that have been raised and criticized previously (e.g. Aydeniz & Kotowski, 2014; Blalock et al., 2008; Munby, 1983; Potvin & Hasni, 2014; Ramsden, 1998) are still relevant and continue to be reflected in assessments validated in recent years. Consequently, significant limitations were found in every published assessment, included the ones described above, either due to the lack of theoretical framework, lack of pilot testing of the items with the target population (i.e. content validity), limited or absence of psychometric analysis, or a combination of more than one of these limitations.

Lack of Conceptualization of the Attitude Construct

While some assessments were rooted in a theory, no dominant framework informing the development and validation process was identified, so test developers used a wide range of theories to develop their assessments. This limited use and testing of a common attitude development framework may affect generalization and comparison of the results obtained on attitudes of students worldwide, from those with different backgrounds or those stemming from educational interventions. Furthermore, for the most part, assessments were validated without a reference to theory. While many test developers justified the inclusion of certain attitudes constructs based on previous studies on attitudes toward science, there were few test developers that explicitly conceptualized the proposed assessments from an attitude development theory. A very common misguided practice used by test developers was to conduct a literature review and include those constructs that had been previously used in existing assessments and that are thought to be measuring attitudes (e.g. Kennedy et al., 2016; Lamb et al., 2012; Morgan et al., 2018). Some test developers tried to link an adapted assessment to a theory, even if the original assessment was developed without any theory (e.g. Heredia & Lewis, 2012). These practices, combined with the extensive use of existing items or making slight changes in already validated assessments, leads to the development of conceptually poor assessments extremely similar to the existing ones, as adverted in previous literature reviews (i.e. Blalock et al., 2008). For example, inconsistent construct structure (e.g. Guzey, Harwell, & Moore, 2014), factors loadings not reflecting the intended attitude construct (e.g. Caputo, 2017), many cross loadings and redundant factors measuring the same construct (e.g. Dijkstra & Goedhart, 2012), or assessments showing different factor structure in subsequent studies [i.e. the CLASS assessment developed by Adams et al. (2006)] illustrates these concerns.

Although some of these assessments may be presented as reliable and valid, the confident use of them is highly debatable.For example, the CLASS assessment (Adams et al., 2006) was described as a "(…) new test designed to measure student belief about physics and about learning physics" (p. 1), however, an examination of the included items and constructs derived reveals that in fact, the CLASS assessments measures everything but beliefs. The relation between beliefs and the CLASS constructs like conceptual connections, personal interest, and problem solving or applied conceptual understanding is uncertain at best.

Other assessments, like the one proposed by Kind et al. (2007) were found to be measuring attitudes toward science and science class and also attitudes toward school in general. These test developers criticized the lack of clarity and definition of the attitude toward science construct in existing literature, and defined attitudes toward science measure as "(…) a way of mapping students' cognitive and emotional opinions about various aspects of science" (p. 873). Therefore, the inclusion of items assessing attitudes toward school in general is simply unreasonable and is completely inconsistent with the conception adopted, thus, compromising the construct validity of the proposed assessment.

This lack of rational decisions was present in many other assessments. For example, Hillman et al. (2016) mixed attitudes toward science class with beliefs and desire to become a scientist. Bauer (2008) assessment measures interest and utility as one main construct, even if it is well known that utility is a separate construct of multidimensional nature (Wigfield & Cambria, 2014). Dijkstra & Goedhart (2012) developed an assessment aiming at measuring attitudes towards climate change and science which is composed by such different factors as attitudes towards school science, societal implication of science, scientists, career in science, pro-environmental behavior, and urgency for climate change. Test developers stated that "All the attitude and behavior areas listed were chosen as areas that could possibly be affected by an initiative such as Carbo Schools", which is clearly not a valid justification, nor an informed decision based on any theoretical framework for attitude development. Unfried, Faber, Stanhope, & Wiebe (2015) developed an attitude toward STEM assessment; however, these authors conceived engineering and technology as the same construct. In addition, a close examination of the constructs reveals several conceptual issues. For example, the attitude toward math construct is made up mainly by self-efficacy items, such as "Math is hard for me"; "I am good at math", or "I can get good grades in math" (p. 630); however, the attitude toward science construct mixes items about the intentions to enroll in science [i.e. "I might choose a career in science" (p. 630)], self-efficacy items [i.e. "I can understand most subjects easily, but science is hard for me to understand" (p. 630)] and beliefs items like "When I am older, knowing science will help me earn money" (p. 630). The same issues are identified in the attitude toward engineering/technology construct. Test developers also included an "(…) construct measuring students attitudes toward 21st century skills." (p. 625), which in fact does not measure *attitude towards* but just skills or at best self-efficacy as can be seen in the following items: "I can lead other to reach a goal", "I can make my own goals for learning", and "I can use time wisely when working on my own" (p. 631). If the assessment were to be measuring *attitudes towards,* thus a favorable or unfavorable disposition towards these skills, items should be redacted such as "I *like* leading others to reach a goal" or "Using time wisely when working on my own *is very important for me*". In addition, the inclusion of a construct measuring general skills within an assessment aimed at measuring attitude toward STEM remains unclear, and clearly exemplifies the existing conceptual issue in attitude research in science education.

Poor Item Selection and Development

Another concern is related to the procedures followed in developing or selecting items. Studies using empirically developed items were minority in this review, with most test developers showing a clear preference for existing items or newly constructed items without empirical reference or pilot test. This practice has two main limitations. On the one hand, assessments using non-empirically derived items may typically reflect the researcher ideologies instead of accurately measure individuals attitudes [e.g. I study physics to learn knowledge that will be useful in my outside of school (Douglas et al., 2014)], compromising the validity of the assessment (Aikenhead & Ryan, 1992). On the other hand, if items are not derived from responses of the target population, it is possible that conceptually difficult, irrelevant and confusing aspects may to be included. For example, the relevance of the following items for K-12 students may be questionable: "The topics taught in my science class are important in the real world" (Lamb et al., 2012, p. 650), "Knowing science will help me earn a living" (Puvirajah, Verma, Li, & Martin-Hansen, 2015, p. 257), or "The primary purpose for learning science is to try to solve real world problems" (Zhang & Campbell, 2011, p. 605). These items appear to be measuring knowledge about the utility of science for the society instead of the test-taker, but not a positive or negative inclination towards this discipline.

Lack of Psychometric Types of Evidence

The psychometric analysis illustrates several problems with the available assessments, mainly related to the lack of evidence of their measurement properties. Overall, validity and reliability psychometric information was found to be lacking or too limited, and a declining pattern in assessments developed in recent years for examining content, criterion and hypothesis-testing construct validity was identified. Given the overwhelming amount of existing information available in the literature related to the development of attitude assessments, these results are not encouraging and are of concern.

Furthermore, with some rare exceptions, like the Colorado Learning Attitudes about Science Survey (CLASS, Adams et al., 2006) and the Science Motivation Questionnaire (SMQ II, Glynn et al., 2011), assessments have not been tested beyond the validation study, therefore, lacking external validation. From thus that were used in subsequent studies, the CLASS (Adams et al., 2006) reported inconsistent results in relation to its factor structure. Regarding the SMQ II, Schumm& Bogner (2016) replicated the original factor structure; however, they reported that two items had insufficient model fit and two subtests will require further examination due to low convergent validity and cross loadings. You, Kim, Black, & Min (2018) examined the reliability and validity of the SMQ II assessment using IRT (i.e. Rasch-Andrich Rating test model), suggesting also the deletion of two items.

The lack of studies reporting on the convergent and discriminant validity of the proposed assessments is also worrying. Considering that attitude is a multidimensional construct, these results are problematic and the construct validity of many assessments can therefore be questioned. This concern is even more problematic in assessments showing cross loadings between items (e.g. Adams et al., 2008; Dermitzaki et al., 2013; Romine, Sadler, Presley, & Klosterman, 2014). The importance of examining this hypothesis-testing validity properties in multidimensional tests is reflected by Owen et al. (2007) psychometric re-evaluation of the Women in Science Test (WiSS). They reported that CFA analysis did not support the original three-factor model, with too high correlations between the original subtests (0.89 and above), which suggest that the structure of the original assessment measured almost the same factor and therefore its validity is dubious. Therefore, the lack of convergent and discriminant tests between sub-constructs calls into question the ability of most reviewed assessments to validly measure individuals' attitudes.

Likewise, temporal stability was not tested for almost all assessments reviewed, thus the extent to which the scores are consistent and accurate in repeated administrations of the assessment is a matter of doubt (AERA et al., 2014). Since assessment in science education is usually used to track the formation and evolution of a desired construct, and to examine interventions impact on the construct under study, the use of assessments lacking temporal stability evidence will lead to results which reliability is unknown at best or non-existent at worst. Without the confidence that the assessment used is reliable and that the results are both representative and stable over time, it will be difficult to determine whether the differences or lack of it identified after the intervention are due to the (in)effectiveness of the educational intervention or because of inconsistencies issues of the assessment. Therefore, the use of assessments whose temporal reliability is unknown may camouflage the real effects of an intervention, which could have serious consequences on the implications and conclusions drawn. The same problem holds for longitudinal studies aimed at studying how attitudes develop throughout the education system, in which wrong conclusions stemming from the use of assessments that were not tested against temporal stability may affect the advancement in this line of research.

Impacts of Purpose on Rigor

Finally, beside the lack of evidence of most types of validity and reliability properties, it is of concern that authors continue to develop new assessment for their own work or projects instead of using existing ones, as already warned by Blalock et al. (2008). For example, Dijkstra & Goedhart (2012) developed a test suited for the Carboschools project, aimed at measuring science attitudes, pro-environmental behaviours, climate change attitudes and knowledge. However, half of the constructs (i.e. attitude towards school science, societal implication of science, attitude toward scientist, and attitude toward a career in science) were created by mixing items from different existing attitude test that also measured these very same attitude objects/constructs, instead of just using the original tests that in fact were more robust in terms of evidence of validity and reliability properties. Likewise, instead of using existing published attitude tests, Tytler-Wood, Knezek and Christensen (2010) developed a Likert-type and semantic differential assessments for the specific evaluation of the impact of the Middle Schoolers Out to Save the World project, with insufficient psychometric evaluation. Puvirajah, Verma & Martin-Hansen (2015) designed a test for the measurement of the impact of a Science-focused after-school program on underrepresented high-school students' attitudes towards science that was also not justified given the many existing attitude tests measuring the same constructs as theirs. Kind, Jones and Barmby (2007) instrument was developed for the evaluation of the Lab in a Lorry initiative and, astonishingly and wrongly, authors used this informal science education program as the rationale behind the included constructs in the test: "All the attitude areas listed, with the exceptions of attitude towards school, were chosen as areas that could possibly be affected by an initiative such as Lab in a Lorry." (p. 878).

These non-exhaustive examples illustrate inadequate and misinformed methodological practices that ignore the conceptual and methodological advances in the literature related to the development and validation of measurement instrument (e.g. AERA, APA & NCME, 2014; DeVellis, 2017). Given that in only four studies it was explicitly stated that the validation of the instrument was part of a larger project, it was not possible to establish a pattern of differences in the quality of the assessments presented in studies that developed their own test for a specific project in comparison with investigations that just focused on the development and validation of a test. However, a closer examination of Table 2 reveals that the studies discussed before reported significantly less psychometric properties, and that therefore fall way shorter in meeting the quality criteria for valid and reliable assessments than those developed in studies where the main focus was the instrument validation itself, instead of the evaluation of a specific project.

Implications

Several implications arise from this review. First, a more comprehensive and clearly described conceptualization and concretion of the latent construct under study (i.e. attitude toward science) is required to reach consensus on which constructs should indeed be considered within the "attitude toward science" umbrella and which should not. This would make the assessment of attitudes toward science more operational, preventing test developers from developing new instruments for each specific project, and form using a random number of non-informed dimensions and items for assessing attitudes. Without this reconceptualization, the vagaries of the various conceptualizations identified in this review raise the question about whether the construct has any benefit to attitudes in science education research. Until then, it is advisable to abandon any attempt to further develop assessments based exclusively on existing literature and also to restrict the use of assessments that are not rooted in any theoretical framework or informed attitude conceptualization.

Second, there is a need to consolidate a robust body of assessments reporting valid and reliable results that would allow comparisons between the attitudes toward science of students from different contexts and background and the results of different educational interventions carried out worldwide. Although educational intervention results are context-dependent, using a similar procedure and assessments for data collection would allow for better comparison of the results obtained. However, there is a need to first provide test-retest reliability evidence before using assessments for repeated measurement purposes. Thus, if the assessments were not tested in repeated administration and no temporal stability evidence was provided, then their use in quasi-experimental time-series design or intervention studies with pre-post test data collection phases should be restricted.

Future studies should further validate the above mentioned existing assessments that have appeared to have acceptable quality. In addition, since assessments for elementary education are scarce, the studies aiming to develop new measures for this grades are strongly advised to use a comprehensive attitude development-theoretical framework like the Expectancy-Value (Eccles et al., 1983; Eccles, Wigfield, Harold, & Blumenfeld, 1993; Eccles, 2005; Eccles et al., 1983; Eccles & Wigfield, 1995; Wigfield et al., 2015; Wigfield & Eccles, 1992, 2000) or the Social Cognitive Career Theory (Lent & Brown, 1994; Lent, Brown, & Hackett, 2002) which have been shown in the literature to be effective. Also, when translating and validating existing assessments in other languages, test developers should pay attention to the translation process and use cross-cultural techniques and equivalence tests between the original and the translated version of the assessment.

Third, test developers aiming at developing new attitude toward science measures should include empirically developed items and submit the assessment to content and face validity with the target population before large-scale validation. In addition, test developers should subject their newly developed assessment to more empirical work to evaluate more psychometric properties than those included in this review. Thus, following the comprehensive framework for test development outlined in this work, test developers should examine at least one psychometric property for temporal stability reliability (especially when the test is intended for measuring interventions outcomes or for time-series design longitudinal studies), internal consistency reliability, criterion validity and construct validity through EFA or CFA. In addition, test developers should examine convergent and discriminant validity between the proposed assessment and existing ones or other theoretically related construct (e.g. attitudes and intentions). For criterion validity, the lack of quality of the existing assessments makes it difficult to choose a *gold standard* test for examining concurrent validity. However, an imperative criterion is to choose assessments that were validated to measure the same constructs and sub-domains as the newly proposed assessment. If there is no assessment that can be considered gold standard, which could easily be the case of attitudes assessments in science education, test hypothesis between the proposed assessment and measures of a construct with conceptual convergence should be tested (Polit, 2015).

It is understandable that test development and validation is very time consuming and sometimes not worth to be included in a single publication; however, this is not what it is requested here. Rather, it is intended to emphasize the need for authors to abandon the dreadful pattern of developing their own test according to their own specific project instead of using existing ones. Therefore, since test development and validation is not an easy task, it is warned of the importance for an assessment to undergone several psychometric studies before being recommended as a valid tool for measuring attitudes toward science, and even more psychometric analysis before advising its use in interventions evaluation studies.

Fourth and last, further studies dedicated to fill existing attitude assessments psychometric gaps or develop new ones should make an effort to report the results in a clear, transparent and exhaustive way. Explicit reference to the psychometric properties analyse should be mandatory, as this will help readers to examine the strengths and limitations of the test. The use of the measurement taxonomy proposed by Polit (2015), outlined in this work, would help to structure the reporting in this field. Nonetheless, given the extensive evidence of the suboptimal reporting of many of the studies included in this review, the development of a reporting guideline for studies evaluating psychometric properties of attitude toward science assessments may be required. In addition, clarity, completeness and missing data about the developed assessments are aspects worth addressing in future studies. As different modes of test administration has been proved to introduce biasing effects (Bowling, 2005), test administration method and instructions are a must. Although in the field of science education studies about the effect of different modes of administration are missing, in other fields such as health care-related research, many studies reported that there is an impact of method of administration in self-administered assessments (e.g. Chan, Orlando, Ghosh-Dastidar, Duan, & Sherbourne, 2004; McHorney, Kosinski, & Ware, 1994) and that bias can be reduced with adequate training (Cella et al., 2015).

Strengths and Limitations

The results presented in this review must be interpreted considering the following strengths and limitations. This review provides an in-depth evaluation of studies reporting on the development and validation of attitude toward science assessments using an exhaustive data extraction procedure based on recent recommended standards. The broad search strategy and the retrieval across *WoS* and *Scopus* databases captured 61 validation studies published in the last fourteen years, proposing a total of 62 assessments. The study search strategy used was comprehensive; however, non-English and non-Spanish assessments, and those not published as articles were likely missed. In addition, as there is no clear conceptualization of what constitutes to measure “attitudes” in science education research, and because assessments for assessing attitudes are being indexed using a wide range of key words, the retrieval of some articles was particularly challenging and therefore a manual search was necessary.

Another limitation may be related to the approach taken to examine the quality of the reviewed attitude assessments. The approach used in this review posits that for an assessment to be considered valid and reliable, it should have been submitted to as many psychometric tests as possible and through multiple studies. Nonetheless, it must be acknowledged that this review considered mostly whether the proposed assessments were subjected to several psychometric tests to assure validity and reliability. While this approach is consistent with the Standards for Educational and Psychological Testing (AERA et al. 2014) stating that “(…) unless a standard is deemed clearly irrelevant, inappropriate, or technically infeasible for a particular use, all standards should be met, making all of them essentially ‘primary’ for the context.” (p. 5), it may be that *the more psychometric tests, the better* approach used has limitations and, consequently, some assessments analysed may be of higher quality than it is stated in this study, even if they have undergone fewer psychometric tests than expected. The opposite is also likely to occur and find assessments that despite having been subjected to many psychometric tests, still have internal inconsistencies (like the CLASS assessment). For this reason, future studies that explicitly and thoroughly assess the quality of existing attitude assessment based on the psychometric results would be a desired complement of this study.

Although the results of this systematic review should be taken within these limitations, this study provides an in-depth examination of the development and validation of attitudes assessments in science education research, with a specific interest in practices related to the reporting of the psychometric properties of the proposed assessments between 2004 and 2018. The results of this review can be useful as (a) it provides a framework for test development and validation studies based on up to date recommendations that can be used for further refining existing assessments or guide the development of new ones, (b) identifies practices within the development of attitude assessments in science education that need to be improved, and (c) highlights the strengths and limitations of 62 attitudes toward science assessments so that science education researchers and practitioners can make an informed choice when selecting assessments for their studies.

Conclusions

This study summarizes the range of psychometric properties recommended in recent specialized literature for developing and validating self-administering measurement assessments and examines the extent to which attitude assessments published between 2004 and 2018 in science education research were validated following informed decision in relation to item generation, theoretical analysis, and psychometric evaluation. Using a systematic review procedure, this study provides an in-depth evaluation of 62 attitude assessment developed in science education. The results demonstrate that although a rather large body of attitude assessment were developed and published in the science education field in the last 14 years, the conceptual and methodological quality of most of them are below modern standards for educational and psychological testing. Most assessments lacked theoretical foundations and were almost based on existing assessments with conceptually poor definition of the attitude construct. In addition, many of the attitude assessments were limited in terms of validity and reliability psychometric evidence, and several limitations related to data reporting and misuse of some psychometric tests were also identified.

While no assessment was virtually free of limitations, some performed better than others on the quality criteria used in this review and reached acceptable-quality standards, and so its use can be recommended. However, further information on assessments' reliability, validity, and applicability are clearly necessary. Future refinement and validation studies of these assessments could lead to a robust body of reliable assessments measuring attitudes in science education. In addition, more research is needed to develop theoretically well-grounded assessments. Particularly, a deeper understanding of the construct under study (i.e. attitudes) is paramount so that future attitude test development and validation studies can propose assessments with stronger theoretical background capable of assessing attitudes toward science with high standards of validity and reliability.

In conclusion, the development and validation of valid and reliable assessments is a critical need for science education research. The findings of the psychometric quality assessment, however, suggest that attitude assessments published in science education in the last two decades are usually lacking theoretically framework and validity and reliability psychometric evidences. A deficiency of valid and reliable assessments utterly constrains the ability to understand and improve individual's attitudes, and to evaluate science education interventions aimed at reducing the science pipeline. It is hoped that this review provides guidelines for researchers and contributes to the improvement of tests development and validation practices in science education.

Acknowledgments

Funding: This study has received funding from the 2017-2021 edition of the University of Burgos PhD research fellowship and by the Spanish Ministry of Economy, Industry and Competitiveness (MINECO) through the project EDU2017-89405-R. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Conflict of Interest: The authors declare that they have no conflict of interest.

References

References marked with an asterisk (*) indicate studies included in the systematic review.

*Abd-El-Khalick, F., Summers, R., Said, Z., Wang, S., & Culbertson, M. (2015). Development and large-scale validation of an instrument to assess Arabic-speaking students' attitudes toward science. *International Journal of Science Education*, *37*(16), 2637–2663. doi:10.1080/09500693.2015.1098789

Adams, S. A., Matthews, C. E., Ebbeling, C. B., Moore, C. G., Joan, E., Fulton, J., & Hebert, J. R. (2005). The effect of social desirability and social approval on self-reports on physical activity. *American Journal of Epidemiology*, *161*(4), 389–398.

*Adams, W. K., Perkins, K. K., Podolefsky, N. S., Dubson, M., Finkelstein, N. D., & Wieman, C. E. (2006). New instrument for measuring student beliefs about physics and learning physics: The Colorado learning attitudes about science survey. *Physical Review Special Topics - Physics Education Research,2*(1), 1–14. doi:10.1103/PhysRevSTPER.2.010101

*Adams, W. K., Wieman, C. E., Perkins, K. K., & Barbera, J. (2008). Modifying and validating the colorado learning attitudes about science survey for use in chemistry. *Journal of Chemical Education, 85*(10), 1435–1439. doi:10.1021/ed085p1435

AERA, APA & NCME. (2014). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association.

Aikenhead, G. S., & Ryan, A. G. (1992). The development of a new instrument: ‘Views on science-technology-society’ (VOSTS). *Science Education*, *76*(5), 477–491.

Ainley, M., Ainley, J. (2011a). A cultural perspective on the structure of student interest in science. *International Journal of Science Education, 33*(1), 51-71. doi:10.1080/09500693.2010.518640

Ainley, M., Ainley, J. (2011b). Student engagement with science in early adolescence: The contribution of enjoyment to students’ continuing interest in learning about science. *Contemporary Educational Psychology, 36*, 4-12. doi:10.1016/j.cedpsych.2010.08.001

Ajzen, I. (1991). The theory of planned behavior. *Organizational Behavior and Human Decision Processes, 50*(2), 179–211.

Andrews, R. (2005). The place of systematic reviews in education research. *British Journal of Educational Studies, 53*(4), 399–416.

Anderson, J. C., & Gerbing, D. W. (1988). Structural equation modeling in practice: A review and recommended two-step approach. *Psychological Bulletin, 103*(3), 411–423. doi:10.1037/0033-2909.103.3.411

*Andrews, S. E., Runyon, C., & Aikens, M. L. (2017). The math-biology values instrument: Development of a tool to measure life science majors’ task values of using math in the context of biology. *CBE Life Sciences Education, 16*(45), 1–12. doi:10.1187/cbe.17-03-0043

Archer, L., Dewitt, J., Osborne, J., Dillon, J., Willis, B., & Wong, B. (2010). “Doing” science versus “being” a scientist: Examining 10/11-year-old schoolchildren’s constructions of science through the lens of identity. *Science Education, 94*(4), 617–639. doi.org/10.1002/sce.20399

Arias, M. R. M., Lloreda, M. J. H., & Lloreda, M. V. H. (2014). *Psicometría*. Alianza Editorial.

Aydeniz, M., & Kotowski, M. R. (2014). Conceptual and methodological issues in the measurement of attitudes towards science. *ElectronicJournal of Science Education, 18*(3), 1–24.

*Bauer, C. F. (2008). Attitude toward chemistry: A semantic differential instrument for assessing curriculum impacts. *Journal of Chemical Education, 85*(10), 1440–1445. doi:10.1021/ed085p1440

*Bayar, A., & Karamustafaoğlu, O. (2015). The Colorado learning attitudes about science survey (CLASS): The study of validity and reliability. *International Journal of Secondary Metabolite*, *2*(1), 40–57.

Bearman, M., Smith, C. D., Carbone, A., Slade, S., Baik, C., Hughes-Warrington, M., & Neumann, D. L. (2012). Systematic review methodology in higher education. *Higher Education Research & Development, 31*(5), 625–640. doi:10.1080/07294360.2012.702735

Bennett, J., Dunlop, L., Knox, K. J., Reiss, M. J., & Torrance Jenkins, R. (2018). Practical independent research projects in science: A synthesis and evaluation of the evidence of impact on high school students. *International Journal of Science Education*, 1–19. doi:10.1080/09500693.2018.1511936

Bennett, J., Lubben, F., & Hogarth, S. (2007). Bringing science to life: A synthesis of the research evidence on the effects of context-based and STS approaches to science teaching. *Science Education, 91*(3), 347–370. doi:10.1002/sce.20186

Bennett, J., Lubben, F., Hogarth, S., & Campbell, B. (2005). Systematic reviews of research in science education: Rigour or rigidity? *International Journal of Science Education*, *27*(3), 387–406. doi:10.1080/0950069042000323719

Benson, J., Clark, F. (1982). A guide for instrument development and validation. *The American journal of occupational therapy*, *36*(12), 789-800. doi:10.5014/ajot.36.12.789

Besley J. C. (2016). The National Science Foundation's science and technology survey and support for science funding, 2006–2014. *Public Understanding of Science 27*. 94–109.

Blalock, C. L., Lichtenstein, M. J., Owen, S., Pruski, L., Marshall, C., & Toepperwein, M. A. (2008). In pursuit of validity: A comprehensive review of science attitude instruments 1935-2005. *International Journal of Science Education, 30*(7), 961–977. doi:10.1080/09500690701344578

*Blanco Blanco, Á., Casas Moreno, Y., & Mafokozi Ndabishibije, J. (2016). Adaptación y propiedades psicométricas de escalas sociocognitivas. Una aplicación en el ámbito vocacional científico-matemático. *REOP - Revista Española de Orientación y Psicopedagogía, 27*(1), 8–28. doi:10.5944/reop.vol.27.num.1.2016.17005

Borsa, J. C., Damásio, B. F., & Bandeira, D. R. (2012). Cross-cultural adaptation and validation of psychological instruments: some considerations. *Paidéia, 22*(53), 423–432.

Bowling, A. (2005). Mode of questionnaire administration can have serious effects on data quality. *Journal of Public Health, 27*(3), 281–291.

Bybee, R. McCrae, B. (2011). Scientific literacy and student attitudes: perspectives from PISA 2006 science. *International Journal of Science Education, 33*(1), 7-26. doi:10.1080/09500693.2010.518644

*Caputo, A. (2017). A brief scale on attitude toward learning of scientific subjects (ATLoSS) for middle school students. *Journal of Educational, Cultural and Psychological Studies, 2017*(16), 57–76.

Cella, D., Hahn, E., Jensen, S., Butt, Z., Nowinski, C., Rothrock, N., & Lohr, K. (2015). *Patient-reported outcomes in performance measurement*. Durham: RTI Press.

Cha, E. S., Kim, K. H., & Erlen, J. A. (2007). Translation of scales in cross-cultural research: Issues and techniques. *Journal of Advanced Nursing, 58*(4), 386–395. doi:10.1111/j.1365-2648.2007.04242.x

Chan, K. S., Orlando, M., Ghosh-Dastidar, B., Duan, N., & Sherbourne, C. D. (2004). The interview mode effect on the center for epidemiological studies depression (CES-D) scale.An item response theory analysis. *Medical Care, 42*(3), 281–289. doi:10.1097/01.mlr.0000115632.78486.1f

*Cheung, D. (2009). Developing a scale to measure students' attitudes toward chemistry lessons. *International Journal of Science Education, 31*(16), 2185–2203. doi:10.1080/09500690802189799

*Chiang, W. W., & Liu, C. J. (2014). Scale of academic emotion in science education: Development and validation. *International Journal of Science Education, 36*(6), 908–928. doi:10.1080/09500693.2013.830233

Clark, L. A., & Watson, D. (1995). Constructing validity: Basic issues in objective scale development. *Psychological Assessment, 7*(3), 309–319. doi:10.1037/1040-3590.7.3.309.

Churchill, G. A., Jr. (1979). A paradigm for developing better measures of marketing constructs. *Journal of Marketing Research*, *16*, 64–73.

Conway, J. M., & Huffcutt, A. I. (2003). A review and evaluation of exploratory factor analysis practices in organizational research. *Organizational Research Methods, 6*(2), 147. doi:10.1177/1094428103251541

*Dermitzaki, I., Stavroussi, P., Vavougios, D., & Kotsis, K. T. (2013). Adaptation of the students' motivation towards science learning (SMTSL) questionnaire in the Greek language. *European Journal of Psychology of Education, 28*(3), 747–766. doi:10.1007/s10212-012-0138-1

DeVellis, R. F. (2017). *Scale development. Theory and applications*. Los Angeles: SAGE.

*Dijkstra, E. M., & Goedhart, M. J. (2012). Development and validation of the ACSI: Measuring students' science attitudes, pro-environmental behaviour, climate change attitudes and knowledge. *Environmental Education Research, 18*(6), 733–749. doi:10.1080/13504622.2012.662213

*Douglas, K. A., Yale, M. S., Bennett, D. E., Haugan, M. P., & Bryan, L. A. (2014). Evaluation of Colorado learning attitudes about science survey. *PhysicalReview Special Topics - Physics Education Research, 10*(2), 1–10. doi:10.1103/PhysRevSTPER.10.020128

Eagly, A. H., & Chaiken, S. (1995). Attitude strength, attitude structure, and resistance to change. In R. E. Petty & J. A. Krosnich (Eds.), *Ogio State University series on attitudes and persuasion, Vol. 4. Attitude strength: Antecedents and consequences* (pp. 413–432). New York: Lawrence Erlbaum Associates, Inc.

Eccles, J., Adler, T., Futterman, R., Goff, S., Kaczala, C., Meece, J., & Midgley, C. (1983). Expectancies, values, and academic behaviors. In J. Spence (Ed.), *Achievement and achievement motivation* (pp. 75–146). San Francisco, CA: W. H. Freeman.

Eccles, J. S. (2005). Subjective task value and the Eccles et al. mode of achievement related choices. In A. J. Elliott & C. S. Dweck (Eds.), *Handbook of competence and motivation* (pp. 105–121). New York: Guilford.

Eccles, J. S., Futterman, R., Goff, S. B., Kaczala, C. M., Meece, J. L., & Midgley, C. (1983). Expectancies, values and academic behaviours. In J. T. Spence (Ed.), *Achievement and achievement motives: Psychological and sociological approaches* (pp. 75–146). San Francisco: W. H. Friedman.

Eccles, J. S., &Wigfield, A. (1995). In the mind of the achiever: The structure of adolescents' academic achievement related-beliefs and self-perceptions. *Personality and Social Psychology Bulletin, 21,* 215–225.

Eccles, J., Wigfield, A., Harold, R., & Blumenfeld, P. (1993). Age and gender differences in children's self and task perceptions during elementary school. *Child Development, 64*(3), 830–847.

*Ergün, A., & Balçın, M. D. (2017). Turkish adaptation of questionnaire on attitudes towards engineers and scientists. *International Electronic Journal of Elementary Education, 10*(1), 103–113. doi:10.26822/iejee.2017131891

Evans, J., & Benefield, P. (2001). Systematic reviews ofeducational research: Does the medical method fit? *British Educational Research Journal, 27*(5), 527–541.

Fabrigar, L. R., Wegener, D. T., Maccallum, R. C., & Strahan, E. J. (1999). Evaluating the use of exploratory factor analysis in psychological research. *Psychological Methods, 4*(3), 272–299.

Fishbein, M., & Ajzen, I. (1975). *Belief, attitude, intention and behavior: An introduction to theory and research.* Reading, MA: Addison Wesley.

Ford, J. K., MacCullum, R. C., & Tait, M. (1986). The application of exploratory factor analysis in applied psychology: A critical review and analysis. *Personnel Psychology, 39*(2), 291–314. doi:10.1111/j.1744-6570.1986.tb00583.x

*Fortus, D., & Vedder-Weiss, D. (2014). Measuring students' continuing motivation for science learning. *Journal of Research in Science Teaching, 51*(4), 497–522. doi:10.1002/tea.21136

Fraser, B. J. (1978). Development of a test of science-related attitudes. *Science Education, 62*(4), 509–515. doi:10.1002/sce.3730620411

Fraser, B. J. (1981). *Test of science related attitudes*. Melbourne: Australian Council for Educational Research.

Freedman, M. P. (1998). Relationship among laboratory instruction, attitude toward science, and achievement in science knowledge. Journal of Research in Science Teaching, 34(4), 343-357. doi:10.1002/(SICI)1098-2736(199704)34:4<343::AID-TEA5>3.0.CO;2-R

Gardner, P. L. (1975). Attitudes to science: A review. *Studies in Science Education*, *2*, 1–41.

*Glynn, S. M., Brickman, P., Armstrong, N., &Taasoobshirazi, G. (2011). Science motivation questionnaire II: Validation with science majors and nonscience majors. *Journal of Research in Science Teaching, 48*(10), 1159–1176. doi:10.1002/tea.20442

*Glynn, S. M., Taasoobshirazi, G., & Brickman, P. (2009). Science motivation questionnaire: Construct validation with nonscience majors. *Journal of Research in Science Teaching, 46*(2), 127–146. doi:10.1002/tea.20267

Gough, D. (2007). Weight of evidece: A framework for appraisal of the quality and relevance of evidence. *Applied and Practice-Based Research, 22*(2), 213–228.

*Guzey, S. S., Harwell, M., & Moore, T. (2014). Development of an instrument to assess attitudes toward science, technology, engineering, and mathematics (STEM). *School Science and Mathematics, 114*(6), 271–279. doi:10.1111/ssm.12077

Haccoun, R. R. (1983). Una nouvelle technique de vérification de l'équivalence de mesures psychologique straduites. *Revue Québécoise de Psychologie, 8*(3), 30–39.

Hair, J. F., Black, W. C., Babin, B. J., Anderson, R. E., & Tatham, R. L. (2010). *Multivariate data analysis (Vol 7.).* Upper Saddle River: Pearson Prentice Hall.

Henson, R. K., & Roberts, J. K. (2006). Use of exploratory factor analysis in published research: Common errors and some comment on improved practice. *Educational and Psychological Measurement, 66*(3), 393–416. doi:10.1177/0013164405282485

*Heredia, K., & Lewis, J. E. (2012). A psychometric evaluation of the Colorado learning attitudes about science survey for use in chemistry. *Journal of Chemical Education, 89*(4), 436–441. doi:10.1021/ed100590t

Higgins, J. P. T., Churchill, R., Chandler, J., & Cumpston, M. S. (2011). *Cochrane handbook for systematic reviews of interventions. Version 5.1.0.* Avaible from: http://handbook-5-1.cochrane.org/

*Hiller, S. E., & Kitsantas, A. (2016). The validation of the citizen science self-efficacy scale (CSSES). *International Journal of Environmental and Science Education, 11*(5), 543–558. doi:10.12973/ijese.2016.405a

*Hillman, S. J., Zeeman, S. I., Tilburg, C. E., & List, H. E. (2016). My attitudes toward science (MATS): The development of a multidimensional instrument measuring students' science attitudes. *Learning Environments Research, 19*, 203–219. doi:10.1007/s10984-016-9205-x

Hinkin, T. R. (1995). A review of scale development practices in the study of organizations. *Journal of Management, 21*(5), 967–988. doi:10.1111/j.1745-4603.1989.tb00456.x

Huang, C., Liao, H., & Chang, S.-H. (1998). Social desirability and the clinical self-report inventory: Methodological reconsideration. *Journal of Clinical Psychology, 54*(4), 517–528.

Johnson, T. P., Fendrich, M., &Mackesy-Amiti, M. E. (2012). An evaluation of the validity of the Crowne-Marlowe need for approval scale. *Quality and Quantity, 46*(6), 1883–1896. doi:10.1007/s11135-011-9563-5

Karami, H. (2015). Exploratory factor analysis as a construct validation tool: (mis)applications in applied linguistics research. *TESOL Journal, 6*(3), 476–498. doi:10.1002/tesj.176

Kennedy, J. P., Lyons, T., & Quinn, F. (2014). The continuing decline of science and mathematics enrolments in Australian high schools. *Teaching Science*, *60*(2), 34–46.

*Kennedy, J., Quinn, F., & Taylor, N. (2016). The school science attitude survey: A new instrument for measuring attitudes towards school science. *International Journal of Research & Method in Education, 39*(4), 422–445. doi:10.1080/1743727X.2016.1160046

Keszei, A. P., Novak, M., & Streiner, D. L. (2010). Introduction to health measurement scales. *Journal of Psychosomatic Research, 68*(4), 319–323. doi:10.1016/j.jpsychores.2010.01.006

Khine, S. M. (2015). *Attitude measurements in science education: Classic and contemporary approaches*. Information Age Publishing, INC Charlotte, NC.

*Kind, P., Jones, K., & Barmby, P. (2007). Developing attitudes towards science measures. *International Journal of Science Education, 29*(7), 871–893. doi:10.1080/09500690600909091

King, M. F., & Bruner, G. C. (2000). Social desirability bias: A neglected aspect of validity testing. *Psychology and Marketing, 17*(2), 79–103.

Kline, P. (2014). *An easy guide to factor analysis*. New York: Routledge.

Klopfer, L. E. (1971). Evaluation of learning in science. In B. S. Bloom, J. T. Hastings, & G. F. Madaus (Eds.), *Handbook of formative and summative evaluation of student learning*. London: McGraw-Hill.

Koballa, T. R. Jr. (1988). The determinants of female junior high school students' intentions to enroll in elective physical science courses in high school: Testing the applicability of the theory of reasoned action. *Journal of Research in Science Teaching*, *25*(6), 479-492.

*Korur, F., Vargas, R. V., & Serrano, N. T. (2016). Attitude toward science teaching of Spanish and Turkish in-service elementary teachers: Multi-group confirmatory factor analysis. *Eurasia Journal of Mathematics, Science and Technology Education, 12*(2), 303–320. doi:10.12973/eurasia.2016.1215a

Laugksch, R. C. (2000). Scientific literacy: A conceptual overview. *Science Education, 84*, 71–94. Doi:10.1002/(SICI)1098-237X(200001)84:1<71::AID-SCE6>3.0.CO;2-C

*Lamb, R. L., Annetta, L., Meldrum, J., & Vallett, D. (2012). Measuring science interest: Rasch validation of the science interest survey. *International Journal of Science and Mathematics Education, 10*, 643–668. doi:10.1007/s10763-011-9314-z

Lent, R. W., & Brown, S. D. (1994). Toward a unifying social cognitive theory of career and academic interest, choice, and performance. *Journal of Vocational Behavior, 45*, 79–122.

Lent, R. W., Brown, S. D., & Hackett, G. (2002). Social cognitive career theory. In D. Brown, L. Brooks, & Associates (Eds.), *Career choice and development* (pp. 255–311). San Francisco, CA: Jossey-Bass.

Liberati, A., Altman, D. G., Tetzlaff, J., Mulrow, C., Gøtzsche, P. C., Ioannidis J. P. A., Clarke, M., Devereaux, P. J., Kleijnen, J., Moher, D. (2009). The PRISMA statement for reporting systematic reviews and meta-analyses of studies that Evaluate health care interventions: Explanation and elaboration. *Plos Medicine*, *6*(7), e1000100. doi:10.1371/journal.pmed.1000100

*Liou, P. Y. (2014). Evaluating measurement properties of attitudinal items related to learning science in Taiwan from TIMSS 2007. *Journal of Baltic Science Education, 13*(6), 856–869.

*Mahoney, M. P. (2010). Students' attitudes toward STEM: Development of an instrument for high school STEM-based programs. *The Journal of Technology Studies, 36*(1), 24–34. doi:10.1007/s11165-013-9387-5

*Maier, M. F., Greenfield, D. B., & Bulotsky-Shearer, R. (2013). Development and validation of a preschool teachers' attitudes and beliefs toward science teaching questionnaire. *Early Childhood Research Quarterly, 28*, 366–378. doi:10.1016/j.ecresq.2012.09.003

*Marsh, H. W., Abduljabbar, A. S., Abu-Hilal, M. M., Morin, A. J. S., Abdelfattah, F., Leung, K. C., … Parker, P. (2013). Factorial, convergent, and discriminant validity of TIMSS math and science motivation measures: A comparison of Arab and Anglo-Saxon countries. *Journal of Educational Psychology, 105*(1), 108–128. doi:10.1037/a0029907

Marshall, M., Lockwood, A., Bradley, C., Adams, C., Joy, C., & Fenton, M. (2000). Unpublished rating scales: A major source of bias in randomized controlled trials of treatments for schizophrenia. *British Journal of Psychiatry, 176*, 249–252. doi:10.1192/bjp.176.3.249

McHorney, C. A., Kosinski, M., & Ware, J. E. J. (1994). Comparisons of the costs and quality of norms for the SF-36 health survey collected by mail versus telephone interview: Results from a national survey. *Medical Care, 32*(6), 551–567.

Millar, R. (1996). Towards a science curriculum for public understanding. *School Science Review, 77*(280), 7–18.

Miller, J. D. (1983). Scientific literacy: A conceptual and empirical review. *Daedalus*, *112*. 29–48.

Millar, R., & Osborne, J. F. (1998). *Beyond 2000: Science education for the future*. London: King's College London.

Mokkink, L. B., Terwee, C. B., Knol, D. L., Stratford, P. W., Alonso, J., Patrick, D. L., … De Vet, H. C. (2010). The COSMIN checklist for evaluating the methodological quality of studies on measurement properties: A clarification of its content. *BMC Medical Research Methodology, 10*(22), 1–8. doi:10.1186/1471-2288-10-22

Mokkink, L. B., Terwee, C. B., Patrick, D. L., Alonso, J., Stratford, P. W., Knol, D. L., … de Vet, H. C. W. (2010). The COSMIN study reached international consensus on taxonomy, terminology, and definitions of measurement properties for health-related patient-reported outcomes. *Journal of Clinical Epidemiology, 63*(7), 737–745. doi:10.1016/j.jclinepi.2010.02.006

Mokkink, L. B., Terwee, C. B., Patrick, D. L., Alonso, J., Stratford, P. W., Knol, D. L., … De Vet, H. C. W. (2010). The COSMIN checklist for assessing the methodological quality of studies on measurement properties of health status measurement instruments: An international delphi study. *Quality of Life Research, 19*(4), 539–549. doi:10.1007/s11136-010-9606-8

Morgado, F. F. R., Meireles, J. F. F., Neves, C. M., Amaral, A. C. S., & Ferreira, M. E. C. (2017). Scale development: Ten main limitations and recommendations to improve future research practices. *Psicologia: Reflexao e Critica, 30*(1), 1–20. doi:10.1186/s41155-016-0057-1

Mota, M. (2019). Explaining science funding attitudes in the United States: The case for science interest. *Public Understanding of Science, 28*(2), 161-176. doi:10.1177/0963662518795397

*Morgan, M., Collins, W., Sparks, G., & Welch, J. (2018). Identifying relevant anti-science perceptions to improve science-based communication: The negative perceptions of science scale. *Social Sciences, 7*(64), 1–18. doi:10.3390/socsci7040064

Munby, H. (1983). Thirty studies involving 'Scientific Attitude Inventory': What confidence can we have in this instrument? *Journal of Research in Science Teaching, 20*, 141–162.

*Navarro, M., Förster, C., González, C., & González-Pose, P. (2016). Attitudes toward science: Measurement and psychometric properties of the test of science-related attitudes for its use in Spanish-speaking classrooms. *International Journal of Science Education, 38*(9), 1459–1482. doi:10.1080/09500693.2016.1195521

Newell, A. D., Tharp, B. Z., Vogt, G. L., Moreno, N. P., Zientek, L. R. (2015). Students' attitudes toward science as predictors of gains on student content knowledge: Benefits of an after-school program. *School science and mathematics, 115*(5), 216-225. doi:10.1111/ssm.12125

Nunnally, J. C., & Bernstein, I. H. (1994). *Psychometric Theory*. New York: McGraw-Hill.

OECD. (2006). *Assessing scientific, reading and mathematical literacy: A framework for PISA 2006*. Paris.

*Oh, Y. J., Jia, Y., Lorentson, M., &LaBanca, F. (2013). Development of the educational and career interest scale in science, technology, and mathematics for high school students. *Journal of Science Education and Technology, 22*, 780–790. doi:10.1007/s10956-012-9430-8

Oppenheim, A. N. (1992). *Questionnaire design, interviewng and attitude measurement*. London: Pinter.

Osborne, J., Simon, S., & Collins, S. (2003). Attitudes towards science: A review of the literature and its implications. *InternationalJournal of Science Education, 25*(9), 1049–1079. doi:10.1080/0950069032000032199

*Owen, S. V., Toepperwein, M. A., Pruski, L. A., Blalock, C. L., Liu, Y., Marshall, C. E., & Lichtenstein, M. J. (2007). Psychometric reevaluation of the women in science scale (WiSS). *Journal of Research in Science Teaching, 44*(10), 1461–1478. doi:10.1002/tea.20187

Petticrew, M. (2001). Systematic reviews from astronomy to zoology: Myths and misconceptions. *British Medical Journal, 322*(7278), 98–101.

Polit, D. F. (2015). Assessing measurement in health: Beyond reliability and validity. *International Journal of Nursing Studies, 52*(11), 1746–1753. doi:10.1016/j.ijnurstu.2015.07.002

Polit, D. F., & Yang, F. (2016). *Measurement and the measurement of Change: A primer for health professionals*. Philadelphia: Lippincott Williams & Wilkins.

Potvin, P., & Hasni, A. (2014). Interest, motivation and attitude towards science and technology at K-12 levels: A systematic review of 12 years of educational research. *Studies in Science Education, 50*(1), 85–129. doi:10.1080/03057267.2014.881626

*Puvirajah, A., Verma, G., Li, H., & Martin-Hansen, L. (2015). Influence of a science-focused after-school program on underrepresented sigh-school students' science attitudes and trajectory: A survey validation study. *International Journal of Science Education, Part B: Communication and Public Engagement, 5*(3), 250–270. doi:10.1080/21548455.2014.930210

*Randler, C., Hummel, E., Gläser-Zikuda, M., Vollmer, C., Bogner, F. X., &Mayring, P. (2011). Reliability and validation of a short scale to measure situational emotions in science education. *International Journal of Environmental & Science Education, 6*(4), 359–370.

Robichaud-Ekstrand, S., Haccoun, R. R., &Millette, D. (1994). Une méthode pour faire valider la traductiond'une questionnaire. *Canadian Journal of Nursing Research, 26*(3), 77–87.

Romine, W. L., Sadler, T. D., & Wulff, E. P. (2017). Conceptualizing student affect for science and technology at the middle school level: Development and implementation of a measure of affect in science and technology (MAST). *Journal of Science Education and Technology, 26*, 534–545. doi:10.1007/s10956-017-9697-x

*Romine, W., Sadler, T. D., Presley, M., & Klosterman, M. L. (2014). Student interest in technology and science (SITS) survey: Development, validation, and use of a new instrument. *International Journal of Science and Mathematics Education, 12*, 261–283. doi:10.1007/s10763-013-9410-3

*Sabah, S., Hammouri, H., &Akour, M. (2013). Validation of a scale of attitudes toward science across countries using Rasch model: Findings from TIMSS. *Journal of Baltic Science Education, 12*(5), 692–703.

*Said, Z., Summers, R., Abd-El-Khalick, F., & Wang, S. (2016). Attitudes toward science among grades 3 through 12 arab students in Qatar: Findings from a cross-sectional national study. *International Journal of Science Education, 38*(4), 621–643. doi:10.1080/09500693.2016.1156184

*Salta, K., &Koulougliotis, D. (2015). Assessing motivation to learn chemistry: Adaptation and validation of science motivation questionnaire II with Greek secondary school students. *Chemistry Education Research and Practice, 16*, 237–250. doi:10.1039/c4rp00196f

Schibeci, R. A. (1983). Selecting appropriate attitudinal objectives for school science. Science Education, 67(5), 595-603. doi.org/10.1002/sce.3730670508

Scholtes, V., Terwee, C., &Poolman, R. (2011). What makes a measurement instrument valid and reliable? *Injury, 42*(3), 236–240.

*Schumm, M. F., & Bogner, F. X. (2016). Measuring adolescent science motivation. *International Journal of Science Education, 38*(3), 434–449. doi:10.1080/09500693.2016.1147659

*Semsar, K., Knight, J. K., Birol, G., & Smith, M. K. (2011). The Colorado learning attitudes about science survey (CLASS) for use in biology. *CBE Life Sciences Education, 10*(3), 268–278. doi:10.1187/cbe.10-10-0133

*Sha, L., Schunn, C., & Bathgate, M. (2015). Measuring choice to participate in optional science learning experiences during early adolescence. *Journal of Research in Science Teaching, 52*(5), 686–709. doi:10.1002/tea.21210

Shaw, M. E., & Wright, J. M. (1968). *Scales for the mesurement of attitude*. New York: McGraw-Hill.

Simpson, R. D., Koballa, T. R., Jr., Oliver, J. S. and Crawley, F. E. (1994). Research on the Affective Dimensions of Science Learning. In D. Gabel (Ed.), *Handbook of Research on Science Teaching and Learning* (pp. 211–234). New York: Macmillan.

*Sjaastad, J. (2013). Measuring the ways significant persons influence attitudes towards science and mathematics. *International Journal of Science Education, 35*(2), 192–212. doi:10.1080/09500693.2012.672775

*Summers, R., & Abd-El-Khalick, F. (2018). Development and validation of an instrument to assess student attitudes toward science across grades 5 through 10. *Journal of Research in Science Teaching, 55*(2), 172–205. doi:10.1002/tea.21416

Tavakol, M., &Dennick, R. (2011). Making sense of Cronbach's alpha. *International Journal of Medical Education, 2*, 53–55. doi:10.5116/ijme.4dfb.8dfd

*Tee, O. P., & Subramaniam, R. (2018). Comparative study of middle school students' attitudes towards science: Rasch analysis of entire TIMSS 2011 attitudinal data for England, Singapore and the U.S.A. as well as psychometric properties of attitudes scale. *International Journal of Science Education, 40*(3), 268–290. doi:10.1080/09500693.2017.1413717

*Tosun, C., & Genç, M. (2015). Adaptation of science attitude scale developed for elementary school students to turkish: Validity and reliability study. *Elementary Education Online, 14*(3), 946–960. doi:10.17051/io.2015.08787

Trochim, W. M., & Donnelly, J. P. (2006). *The research methods knowledge base*. Cincinnati: Atomic Dog Publishing Inc.

*Tuan, H. L., Chin, C. C., & Shieh, S. H. (2005). The development of a questionnaire to measure students' motivation towards science learning. *International Journal of Science Education, 27*(6), 639–654. doi:10.1080/0950069042000323737

*Tyler-Wood, T., Knezek, G., & Christensen, R. (2010). Instruments for assessing interest in STEM content and careers. *Journal of Technology and Teacher Education, 18*(2), 341–363. doi:10.1016/j.tate.2005.03.006

Tytler, R., & Osborne, J. (2012). Attitudes and aspirations towards science. In B. J. Fraser, K. Tobin, & C. J. McRobbie (Eds.), *Second International Handbook of Science Education* (pp. 597–625). Dordrecht: Springer Netherlands. doi:10.1007/978-1-4020-9041-7

*Unfried, A., Faber, M., Stanhope, D. S., & Wiebe, E. (2015). The development and validation of a measure of student attitudes toward science, technology, engineering, and math (S-STEM). *Journal of Psychoeducational Assessment, 33*(7), 622–639. doi:10.1177/0734282915571160

*van Aalderen-Smeets, S., &Walma van der Molen, J. (2013). Measuring primary teachers' attitudes toward teaching science: Development of the dimensions of attitude toward science (DAS) instrument. *International Journal of Science Education, 35*(4), 577–600. doi:10.1080/09500693.2012.755576

Van De Mortel, T. F. (2008). Faking it: Social desirability response bias in self- report research. *Australian Journal of Advanced Nursing, 25*(4), 40–48.

*Velayutham, S., Aldridge, J., & Fraser, B. (2011). Development and validation of an instrument to measure students' motivation and self-regulation in science learning. *International Journal of Science Education, 33*(15), 2159–2179. doi:10.1080/09500693.2010.541529

*Villafañe, S. M., & Lewis, J. E. (2016). Exploring a measure of science attitude for different groups of students enrolled in introductory college chemistry. *Chemistry Education Research and Practice, 17*(4), 731–742. doi:10.1039/c5rp00185d

*Wang, T., & Berlin, D. (2010). Construction and validation of an instrument to measure Taiwanese elementary students' attitudes toward their science class. *International Journal of Science Education, 32*(18), 2413–2428. doi:10.1080/09500690903431561

*Wang, M. Te, Fredricks, J. A., Ye, F., Hofkens, T. L., & Linn, J. S. (2016). The math and science engagement scales: Scale development, validation, and psychometric properties. *Learning and Instruction, 43,* 16–26. doi:10.1016/j.learninstruc.2016.01.008

*Wendt, J. L., &Rockinson-Szapkiw, A. (2018). A psychometric evaluation of the English version of the dimensions of attitudes toward science instrument with a U.S. population of elementary educators. *Teaching and Teacher Education, 70*, 24–33. doi:10.1016/j.tate.2017.11.009

Wigfield, A., & Cambria, J. (2014). Expectancy-value theory: Retrospective and prospective. In T. C. Urdan& S. A. Karabenick (Eds.), *The decade ahead: Theoretical perspectives on motivation and achievement (advances in motivation and achievement, Volume 16 Part A)* (pp. 35–70). Emerald Group Publishing Limited.

Wigfield, A., & Eccles, J. S. (1992). The development of achievement task values: A theoretical analysis. *Developmental Review, 12,* 265–310.

Wigfield, A., & Eccles, J. S. (2000). Expectancy-value theory of motivation. *Contemporary Educational Psychology, 25*, 68–81.

Wigfield, A., & Eccles, J. S. (2002). The development of competence beliefs, expectancies for success, and achievement values from childhood through adolescence. In A. Wigfield& J. S. Eccles (Eds.), *Development of achievement motivation* (pp. 91–120). San Diego: Academic Press.

Wigfield, A., Eccles, J. S., Fredricks, J. A., Simpkins, S., Roeser, R. W., &Schiefele, U. (2015). Development of achievement motivation and engagement. In M. E. Lamb & R. M. Lerner (Eds.), *Socioemotional processes: Vol 3. Handbook of child psychology and developmental science* (pp. 657–700). Hoboken: Wiley.

*Wilcox, B. R., & Lewandowski, H. J. (2016). Students' epistemologies about experimental physics: Validating the colorado learning attitudes about science survey for experimental physics. *Physical Review Physics Education Research, 12*, 1–11. doi:10.1103/PhysRevPhysEducRes.12.010123

*Xu, X., & Lewis, J. E. (2011). Refinement of a chemistry attitude measure for college students. *Journal of Chemical Education, 88*(5), 561–568. doi:10.1021/ed900071q

*Yasar, S., &Anagün, S. S. (2009). Reliability and validity studies of the science and technology course scientific attitude scale. *Journal of Turkish Science Education, 6*(2), 43–54.

*You, H. S. (2016). Rasch validation of a measure of reform-oriented Science teaching practices. *Journal of Science Teacher Education, 27*(4), 373–392. doi:10.1007/s10972-016-9466-3

*You, H. S., Kim, K., Black, K., & Min, K. W. (2018). Assessing science motivation for college students: Validation of the science motivation questionnaire II using the Rasch-Andrich rating scale model. *Eurasia Journal of Mathematics, Science and Technology Education, 14*(4), 1161–1173. doi:10.29333/ejmste/81821

*Zhang, D., & Campbell, T. (2011). The psychometric evaluation of a three-dimension elementary science attitude survey. *Journal of Science Teacher Education, 22*, 595–612. doi:10.1007/s10972-010-9202-3

Table 1.

Search strategy

Web of Science Core Collection database						
Topic attitud* view* opinion* image* motivation*	AND	Title scien*	AND	Title scale* instrument* measur* survey questionnaire* tool*	AND	Topic validat* develop* psychometric* properties evaluate*
Scopus database						
Title, abstract and key words attitud* view* opinion* image* motivation*	AND	Title scien*	AND	Title scale* instrument* measur* survey questionnaire* tool*	AND	Title, abstract and key words validat* develop* psychometric* properties evaluate*
Full search strategy						
attitud* OR view* OR opinion* OR image* OR motivation* AND scien* AND scale* OR instrument* OR measur* OR survey* OR questionnaire* OR tool* AND validat* OR develop* OR psychometric* OR properties OR evaluat*						

Table 2.

Psychometric quality of attitudes instruments in Science education

Reference	Theory	Items	Reliability			Validity				
			Reliability	Internal consistency	Measurement error	Content & face validity	Criterion validity	Construct validity		
								Structural	Hypothesis testing	Cross cultural
Abd-El-Khalick et al.(2015)	Yes	ES, NE	?	CFA based reliability*	?	EP, TP	?	EFA, CFA	?	?
Adams et al.(2006)	?	ES, E	Test-retest	?	Yes	EP, TP	Predictive	EFA	Discriminative	n/a
Adams et al. (2008)	?	ES, NE	Test-retest*	Cronbach α	Yes	EP, TP	?	?	Discriminative	n/a
Andrews et al.(2017)	Yes*	ES, NE	?	Cronbach α	?	EP, TP	?	EFA, CFA	Convergent Discriminant	n/a
Bauer (2008)	?	?	Test-retest	Cronbach α	?	?	?	EFA	Convergent Discriminant Discriminative	n/a
Bayar & Karamustafaoğlu(2015)	?	ES	Test-retest	Item-total *r*	?	TP	?	EFA, CFA	Convergent Discriminative	Yes
Blanco Blanco et al. (2016)	Yes	ES, NE	?	Cronbach α	?	EP	?	EFA	Discriminant*	Yes
Caputo(2017)	Yes	ES	?	Cronbach α	Yes	EP	Predictive	EFA, CFA	Convergent Discriminant Discriminative	n/a
Cheung(2009)	Yes	ES, LR, E	?	Cronbach α Item-total *r*	?	EP, TP	?	CFA	Convergent Discriminative	?
Chiang &Liu(2014)	Yes*	?	?	Cronbach α Item-total *r*	?	?	?	EFA*, CFA	Convergent Discriminant Discriminative	n/a
Dermitzaki et al.(2013)	Yes	ES	?	Cronbach α Item-total *r* Split-half	?	TP	?	CFA	Discriminant	Yes
Dijkstra & Goedhart(2012)	No	ES	?	Cronbach α	?	EP	?	EFA	Convergent Discriminative	?
Douglas et al.(2014)	No	ES	?	Cronbach α	?	?	?	EFA, CFA	?	n/a
Ergün & Balçın (2017)	?	ES	?	Cronbach α Item-total *r*	?	TP	?	EFA, CFA	?	Yes

Table 2.

Psychometric quality of attitudes instruments in Science education (continued)

Reference	Theory	Items	Reliability			Validity				
			Reliability	Internal consistency	Measurement error	Content & face validity	Criterion validity	Construct validity		
								Structural	Hypothesis testing	Cross cultural
Fortus et al. (2014)	Yes	NE	?	Cronbach α	?	EP, TP		EFA, IRT	Discriminative	n/a
Glynn et al.(2011)	Yes	ES, E	?	Cronbach α	?	TP	?	EFA, CFA	Convergent Discriminative	n/a
Glynn et al. (2009)	Yes	ES, E	?	Cronbach α	?	?	?	EFA	Convergent Discriminative	n/a
Guzey et al. (2014)	Yes*	LR	?	Cronbach α	?	EP	?	EFA	Discriminative	n/a
Heredia & Lewis(2012)	Yes*	ES	?	Cronbach α Item-total *r*	?	?	?	CFA	?	n/a.
Hiller & Kitsantas(2016)	Yes	ES	?	Cronbach α	?	EP	Predictive	EFA, CFA	?	n/a.
Hillman et al. (2016)	?	ES	?	Cronbach α	?	EP	?	?	?	n/a.
Kennedy et al.(2016)	?	ES, NE	?	Cronbach α Beta Lambda	?	TP	?	EFA	?	n/a.
Kind et al.(2007)	?	ES	?	Cronbach α	?	?	?	EFA	Convergent*	n/a.
Korur et al. (2016)	Yes	ES	?	Cronbach α	?	?	?	CFA	Convergent Discriminant	Yes
Lamb et al.(2012)	?	ES	?	Cronbach α	?	EP	?	EFA, IRT	?	n/a.
Liou(2014)	Yes*	ES	?	Cronbach α	Yes	?	Predictive	EFA, CFA, IRT	?	n/a.
Mahoney(2010)	?	ES	?	Cronbach α	?	EP, TP	Concurrent*	EFA	Discriminative	n/a.
Maier et al.(2013)	?	ES, LR	?	Cronbach α Item-total *r*	?	TP	?	EFA, CFA	Convergent Discriminative	n/a.
Marsh et al.(2013)	Yes*	NE	?	Cronbach α	?	?	?	CFA	Convergent Discriminant	n/a.
Morgan et al. (2018)	?	E	?	Cronbach α	?	?	Predictive*	EFA, CFA	?	n/a.
Navarro et al. (2016)	Yes	ES	?	Cronbach α	?	?	Concurrent*	EFA, CFA	Discriminant	Yes

Table 2.

Psychometric quality of attitudes instruments in Science education (continued)

Reference	Theory	Items	Reliability			Validity				
								Construct validity		
			Reliability	Internal consistency	Measurement error	Content & face validity	Criterion validity	Structural	Hypothesis testing	Cross cultural
Oh et al. (2013)	Yes	ES, E	?	Cronbach α	?	EP, TP	?	EFA, CFA	Discriminant	n/a.
Owen et al.(2007)	?	ES	?	Cronbach α Split-half	?	?	?	EFA, CFA	Discriminative	n/a.
Puvirajah et al.(2015)	?	ES	?	Cronbach α	?	?	?	EFA	?	n/a.
Randler et al.(2011)	?	ES	?	Cronbach α	?	?	Concurrent*	CFA	Discriminative	n/a.
Romine et al. (2017)	?	ES, NE	?	Cronbach α	?	?	?	EFA IRT	Discriminative	n/a.
Romine et al. (2014)	?	ES	?	Cronbach α	?	EP	?	EFA*, CFA, IRT	Discriminative *	n/a.
Sabah et al.(2013)	?	ES	?	Item reliability indices	?	?	?	EFA, IRT	?	n/a.
Salta &Koulougliotis (2015)	Yes	ES	?	Cronbach α	?	TP		CFA	Discriminative	Yes
Schumm & Bogner (2016)	Yes	ES	?	Cronbach α	?	?	?	EFA	Convergent Discriminative	Yes
Semsar et al.(2011)	?	ES	Test-retest*	?	?	EP, TP	?	EFA	Discriminative	n/a.
Sha et al., (2015)	Yes	NE	?	Amor's θ	?	?	?	IRT	Convergent	n/a.
Sjaastad(2013)	Yes	ES, E	?	Cronbach α Separation indices	?	TP	?	IRT	?	n/a.
Summers & Abd-El-Khalick (2018)	Yes	ES, NE	?	CFA based reliability	?	EP, TP	?	CFA	?	n/a.
Tee & Subramaniam(2018)	?	ES	?	Separation indices	?	?	?	IRT	?	n/a.
Tosun & Genç (2015)	?	ES	?	Cronbach α Item-total *r*	Yes	?	?	EFA, CFA	Discriminative	Yes
Tuan et al. (2005)	?	ES	?	Cronbach α	?	EP	?	EFA	Convergent Discriminant Discriminative	n/a.

Table 2.

Psychometric quality of attitudes instruments in Science education (continued)

Reference	Theory	Items	Reliability			Validity				
			Reliability	Internal consistency	Measurement error	Content & face validity	Criterion validity	Construct validity		
								Structural	Hypothesis testing	Cross cultural
Tyler-Wood et al. (2010)	?	ES	?	Cronbach α	?	EP	?	EFA	Discriminative	n/a.
Tyler-Wood et al.(2010)	?	ES	?	Cronbach α	?	EP	?	EFA	Convergent	n/a.
Unfried et al.(2015)	Yes*	ES	?	Cronbach α	?	EP, TP	?	EFA, CFA	?	n/a.
van Aalderen-Smeets & Walma van der Molen (2013)	Yes	NE	?	Cronbach α	?	TP	Predictive	EFA, CFA	Convergent* Discriminative	n/a.
Velayutham et al.(2011)	Yes	ES, NE	?	Cronbach α	?	EP, TP	Predictive	EFA	Convergent Discriminant Discriminative	n/a.
Villafañe & Lewis(2016)	Yes	ES	?	Cronbach α	?	?	Predictive	CFA	Discriminative	n/a.
Wang et al. (2016)	Yes	E	?	Cronbach α	?	EP, TP	Predictive	CFA	Convergent	n/a.
Wang & Berlin(2010)	?	ES		Cronbach α Item-total *r*		EP, TP	?	EFA	Discriminative *	Yes
Wendt & Rockinson-Szapkiw (2018)	Yes	ES	?	Cronbach α	?	EP	?	CFA	?	Yes
Wilcox & Lewandowski (2016)	?	ES	Test-retest*	Cronbach α Item-total *r*	?	?	?	EFA*	Convergent Discriminative	n/a.
Xu & Lewis(2011)	?	ES	Test-retest	Cronbach α	?	?	Predictive	EFA, CFA	Convergent	n/a.
Yasar &Anagün(2009)	?	ES	?	Cronbach α	?	?	?	EFA	Discriminant	n/a.
You (2016)	Yes	ES	?	Item separation indices	Yes	?	?	IRT	?	n/a.
You et al. (2018)	Yes	ES	?	Cronbach α Item separation indices	?	?	?	IRT	Discriminative	n/a.
Zhang & Campbell(2011)	Yes	ES	?	Cronbach α	?	EP	?	CFA	Convergent	n/a.

Note:
? unknown/not reported, *n/a.* not applicable, *unclear or problematic; *LR* literature review, *ES* existing scales, *E* empirical (interviews/focus group), *NE* not empirical*; EP* Expert panel, *TP* Target population, *EFA* Exploratory Factor Analysis, *CFA* Confirmatory Factor Analysis, *IRT* Item Response Theory

Table 3.

Frequency statistics for reliability evidences

Reliability evidences	*N*	%
1	41	66.1
2	17	27.4
3	4	6.5
4	0	0
5	0	0
Total	62	100.0

Note:
The maximum reliability evidence is 5.

Table 4.

Frequency statistics for validity evidences

Validity evidences	*N*	%
1	5	8.1
2	9	14.5
3	22	35.5
4	10	16.1
5	12	19.4
6	2	3.2
7	2	3.2
8	0	0
9	0	0
Total	62	100.0

Notes:
[a]The maximum validity evidence is 9.
[b]Cross-cultural was not considered, as it does only apply to translated tests.
[c]IRT evidence was not considered since its use involves additional EFA and CFA tests.

Table 5.

Descriptive statistics on validity psychometric evidences reported

Validity evidences	N	%
Content & face		
Expert panel	27	43.5
Target population	23	37.7
Not reported	23	41.9
Criterion		
Concurrent	3	4.8
Predictive	10	16.1
Not reported	49	79.0
Construct		
Structural		
EFA	43	69.4
CFA	32	51.6
IRT	11	17.7
Not reported	2	3.2
Hypothesis testing		
Convergent	23	37.1
Discriminant	13	21.0
Discriminative	30	48.4
Not reported	19	30.6
Cross-cultural[a]	11	78.6
Back-translation	11	78.6
Equivalence tests	1	7.1

Note:
[a]Cross-cultural validity only applicable to 14 tests

Table 6.

Descriptive statistics on reliability psychometric evidences reported

Reliability evidences	N	%
Test-retest	7	11.3
Internal consistency		
Cronbach α	53	85.5
Item-total *r*	10	16.1
Split-half	2	3.2
Other	9	14.5
Measurement error	6	9.7

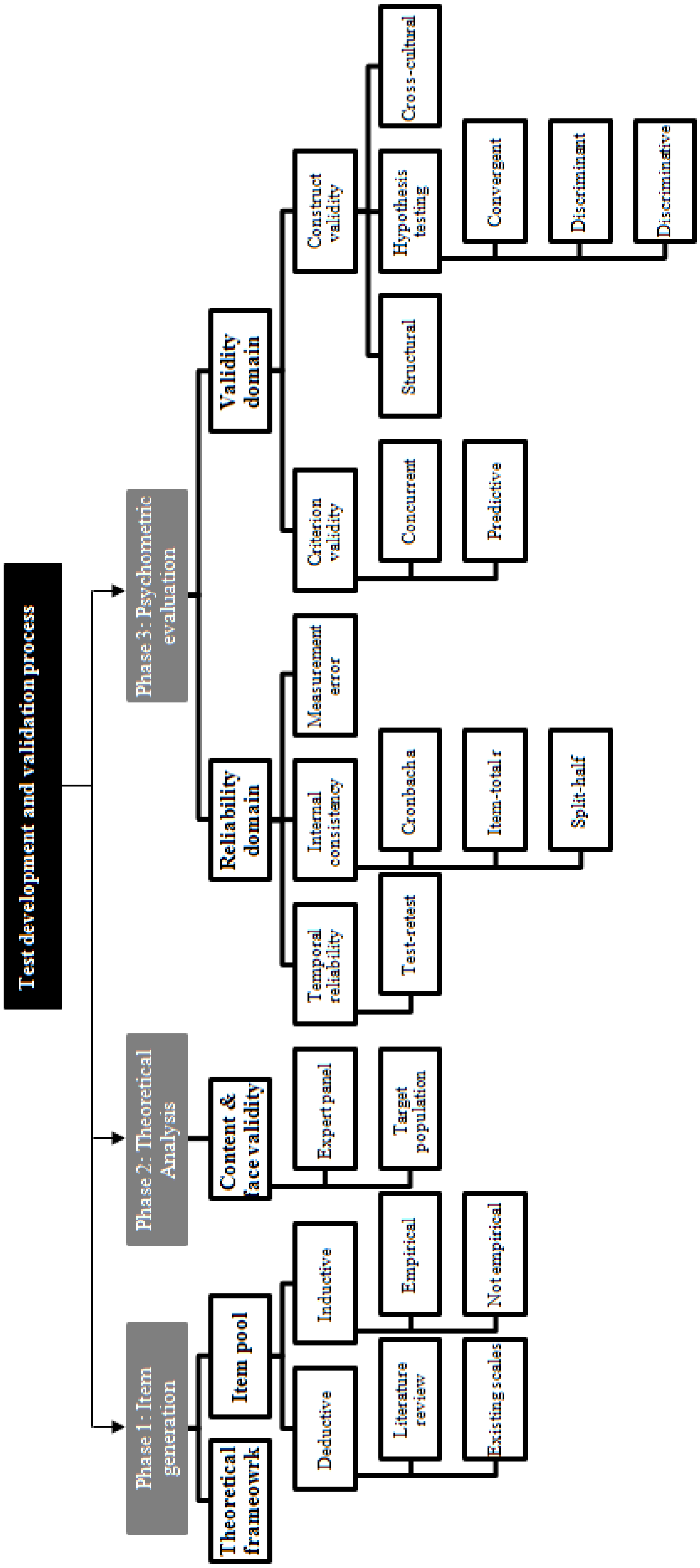


Figure 1. Paradigm for test development and validation

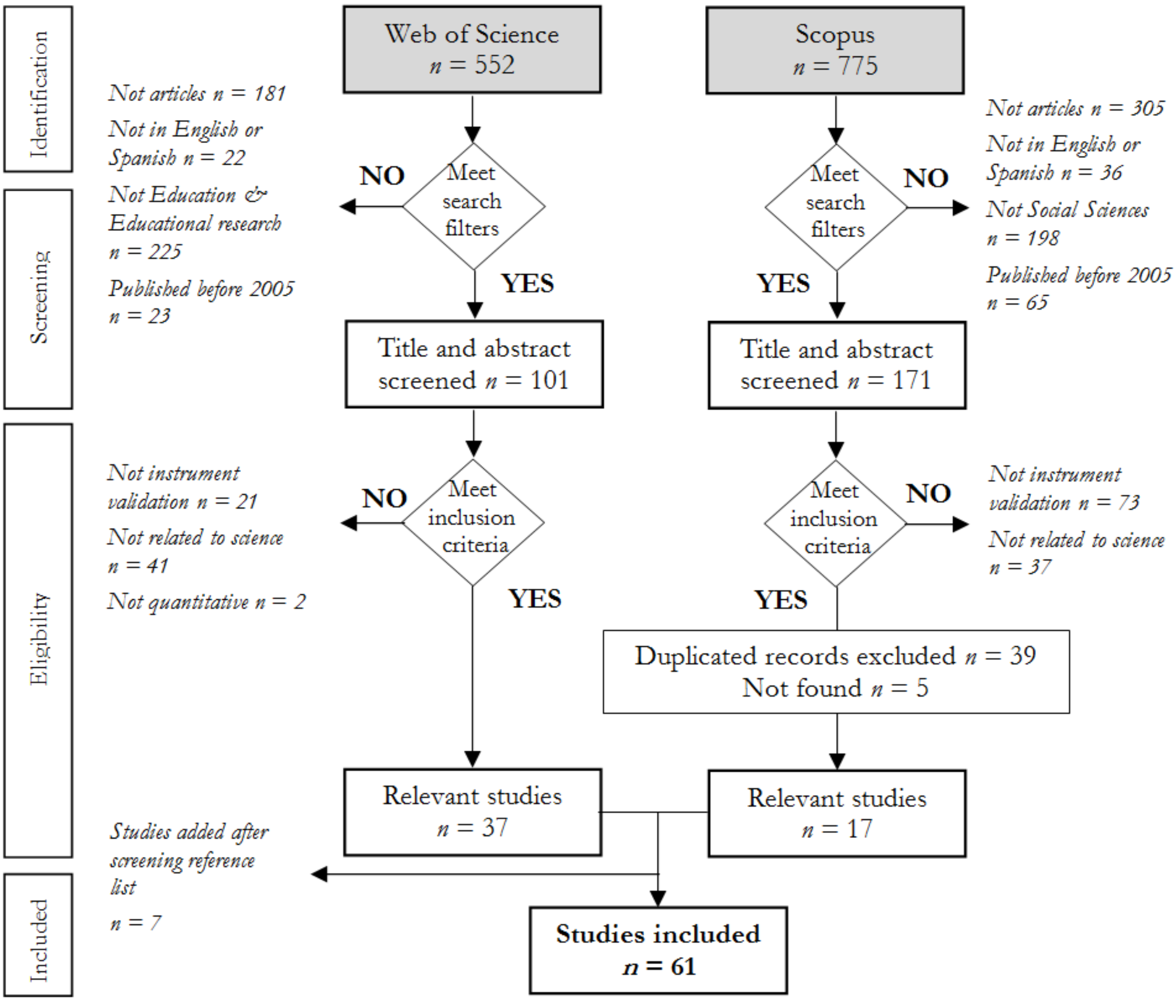


Figure 2. Studies identification and selection flowchart

Supplemental File 1: Quality appraisal checklist

In this online resource, we provide the quality appraisal checklist used to examine evidences of theoretical underpinnings and psychometric properties in attitude assessments published in science education research, as reported in the parent article entitled: "Still on the Wrong Path: A Systematic Review and Evaluation of Attitude Test Development Practices in Science Education".

Phases, steps, and indices	Evaluation			
PHASE 1: ITEM GENERATION				
Theoretical framework				
The items, where developed according to an existing theoretical framework?	*Yes*	*?*	*	*n/a*
Item pool development				
Where deductive – literature review techniques used?	*Yes*	*?*	*	*n/a*
Where deductive – existing scales techniques used?	*Yes*	*?*	*	*n/a*
Where inductive – empirical techniques used?	*Yes*	*?*	*	*n/a*
Where inductive – not empirical techniques used?	*Yes*	*?*	*	*n/a*
PHASE 2: THEORETICAL ANALYSIS				
Content and face validity				
Was content and face validity examined using a panel of experts?	*Yes*	*?*	*	*n/a*
Was content and face validity examined by interviewing the target population?	*Yes*	*?*	*	*n/a*
PHASE 3: PSYCHOMETRIC EVALUATION				
Reliability domain				
Temporal reliability				
Was temporal stability (i.e. test-retest) examined?	*Yes*	*?*	*	*n/a*
Internal consistency				
Was the Cronbach α index examined?	*Yes*	*?*	*	*n/a*
Were item-total correlations examined?	*Yes*	*?*	*	*n/a*
Was the split-half index examined?	*Yes*	*?*	*	*n/a*
Measurement error				
Was the measurement error examined?	*Yes*	*?*	*	*n/a*
Validity domain				
Criterion Validity				
Was concurrent validity examined?	*Yes*	*?*	*	*n/a*
Was predictive validity examined?	*Yes*	*?*	*	*n/a*
Construct validity				
Was structural validity examined through EFA?	*Yes*	*?*	*	*n/a*
Was structural validity examined through CFA?	*Yes*	*?*	*	*n/a*
Was structural validity examined through IRT?	*Yes*	*?*	*	*n/a*

Hypothesis testing				
Was convergent validity examined?	*Yes*	*?*	*	*n/a*
Was discriminant validity examined?	*Yes*	*?*	*	*n/a*
Was discriminative validity examined?	*Yes*	*?*	*	*n/a*
Cross-cultural validity				
When applicable, were assessments adapted at least using forward and back-translations?	*Yes*	*?*	*	*n/a*

Notes:
If aspects were reported, circle *Yes*; if not, circle *?* (unknown); if reported but problematic in terms of misinformed practices or unclear results, circle * (yes, but problematic); if not applicable, circle *n/a* (not applicable).

Supplemental File 2: Tables and Figures

In this online resource, we provide additional Tables and Figures for the parent article entitled: "Still on the Wrong Path: A Systematic Review and Evaluation of Attitude Test Development Practices in Science Education":

List of tables:

List of figures:

Table S1.

Included assessments according to journal of publication

Journals	Publications	
	N	%
CBE Life Sciences Education	2	3.3
Chemistry Education Research and Practice	2	3.3
Early Childhood Research Quarterly	1	1.6
Elementary Education Online	1	1.6
Environmental Education Research	1	1.6
Eurasia Journal of Mathematics, Science & Technology Education 2	2	3.3
European Journal of Psychology of Education	1	1.6
International Electronic Journal of Elementary Education	1	1.6
International Journal of Assessment Tools in Education	1	1.6
International Journal of environmental & Science Education	2	3.3
International Journal of Research & Method in Education	1	1.6
International Journal of Science and Mathematics Education	2	3.3
International Journal of Science Education	12	19.7
International Journal of Science Education, Part B	1	1.6
Journal of Baltic Science Education	2	3.3
Journal of Chemical Education	4	6.6
Journal of Educational Psychology	1	1.6
Journal of Educational, Cultural and Psychological Studies	1	1.6
Journal of Psychoeducational Assessment	1	1.6
Journal of Research in Science Teaching	6	9.7
Journal of Science Education and Technology	2	3.3
Journal of Science Teacher Education	2	3.3
Journal of Technology and Teacher Education	1	1.6
Journal of Turkish Science Education	1	1.6
Learning and Instruction	1	1.6
Learning Environments Research	1	1.6
Physical Review	3	4.9
REOP	1	1.6
School Science and Mathematics	1	1.6
Social Sciences	1	1.6
Teaching and Teacher Education	1	1.6
The Journal of Technology Studies	1	1.6
TOTAL	61	100

Table S2.

Language of included assessments

Language[a]	Publications	
	N	%
Arab	2	3.2
Chinese	1	1.6
Dutch	2	3.2
English	38	61.3
French	1	1.6
German	2	3.2
Greek	2	3.2
Hebrew	1	1.6
Italian	2	3.2
Norwegian	1	1.6
Spanish	4	6.5
Taiwanese	1	1.6
Turkish	5	8.1
Not specified	5	8.1

Note:
[a]Some assessments were validated in more than one language.

Table S3.

Aim of included assessments

Aim of publication	Publications	
	N	%
Attitude toward discipline	16	25.8
Attitude toward specific course	20	32.3
Motivation	9	14.5
Beliefs	8	12.9
Others	9	14.5
Total	62	100.0

Table S4.

Included assessments according to grade level

Grade level[a]	N	%
Elementary	13	21
Middle	22	35.5
High school	18	29
Undergraduates	17	27.4
In service teachers	5	8.1
Adults in general	1	1.6

Note:
[a]Some assessments were designed for multiple grade levels

Table S5.

Structure of the included assessments

Characteristics	Publications	
	N	%
Factors		
Unidimensional	7	11.3
2/3	19	30.6
4/5	22	35.5
6/7	10	16.1
8 or more	4	6.5
Items		
1-10	9	14.5
11-20	14	22.6
21-30	21	33.9
31-40	13	21
41 or more	5	8.1

Table S6.

Descriptive statistics on theoretical framework and item development strategy used

Theory and item criteria	N	%
Theoretical Framework		
Yes	32	51.6
No	30	48.4
Item development method		
Literature review	3	4.8
Existing scales	53	85.5
Empirical	8	12.9
Not empirical	12	19.4
Not reported	2	1.6

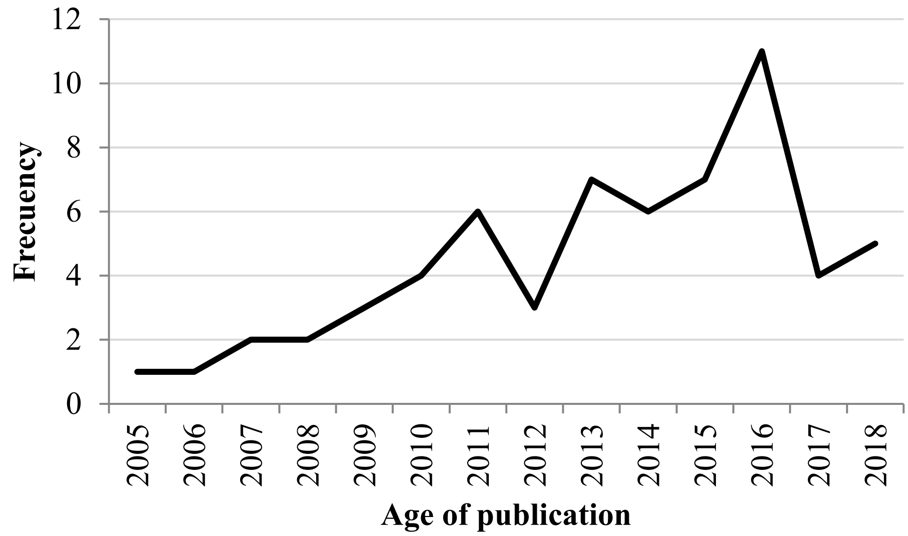


Figure S1. Publication of attitude test development studies included in this review

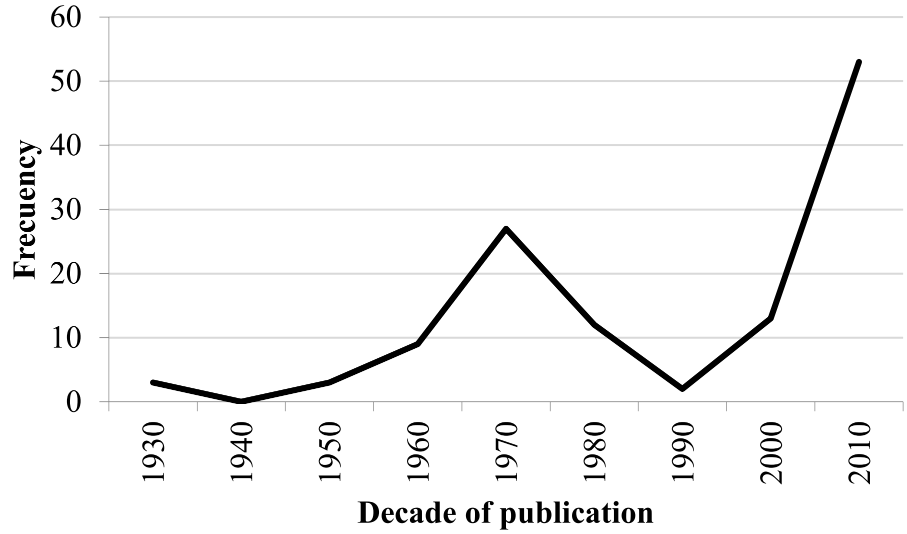


Figure S2. Publication of attitude test development studies per decade. Represented are attitude studies included this and Blalock et al's. (2008) review.

Appendix 5: SUCCESS instrument validation study #1

This appendix includes the full-text of the in-review article about the first psychometric study on the SUCCESS instrument used in this dissertation to assess the effectiveness of the proposed pedagogical framework for STEM-based science education in improving elementary school student's expectancy of success in school science.

Toma, R. B., & Meneses Villagrá, J. A. (under peer review). Development and validation of the SUCCESS instrument: Towards a valid and reliable measure of expectancies of success in school science.

Development and Validation of the SUCCESS Instrument: Towards a Valid and Reliable Measure of Expectancies of Success in School Science

[1,2*]Toma, Radu Bogdan; [1,3]Meneses Villagrá, Jesús Ángel

[1]University of Burgos. Faculty of Education. Department of Specific Didactics.

c/Villadiego, s/n, 09001, Burgos (Spain).

[2]E-mail address: rbtoma@ubu.es; Telephone number: +34647709487;
ORCID: http://orcid.org/0000-0003-4846-7323

[3]E-mail address: meneses@ubu.es ORCID: https://orcid.org/0000-0003-4839-0418

* Corresponding author.

Abstract

Research focused on factors affecting the decline in enrolments in science related university degrees has constantly struggled with poorly conceptualized and developed measurement instruments. This study reviews the theoretical and psychometric properties of the most relevant measurement instruments widely used in the Spanish science education research literature. Afterwards, the development and validation of an instrument for the assessment of Spanish-speaking elementary students' expectancies of success in school science is reported. This new instrument is developed following modern psychometric standards and seeks to overcome the pitfalls identified in the reviewed existing instruments by rooting the items in a robust theoretical framework, using a panel of experts and cognitive interviews with the target sample for content validity and item readability and interpretation, principal component (PCA) and principal axis factoring (PAF) exploratory factor analysis for construct validity, and Cronbach α and item-total r for internal consistency reliability. The results reveal a short and parsimonious unidimensional scale composed of 6 Likert-type items that are consistent with theoretical expectations and that provide evidence for strong validity and reliability psychometric properties. This promising instrument can be used to advance research on factors affecting enrollments in science education careers from early stages of the educational system.

Keywords: expectancies of success, elementary education, psychometric properties, scale, validation

Funding: This study has received funding from the 2017-2021 edition of the University of Burgos PhD research fellowship and by the Spanish Ministry of Economy, Industry and Competitiveness (MINECO) through the project EDU2017-89405-R. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Conflict of Interest: On behalf of all authors, the corresponding author states that there is no conflict of interest.

Development and Validation of the SUCCESS Instrument: Towards a Valid and Reliable Measure of Expectancies of Success in School Science

Research focused on factors affecting the decline in enrolments in science related university degrees has constantly struggled with poorly conceptualized and developed measurement instruments. This study reviews the theoretical and psychometric properties of the most relevant measurement instruments widely used in the Spanish science education research literature. Afterwards, the development and validation of an instrument for the assessment of Spanish-speaking elementary students' expectancies of success in school science is reported. This new instrument is developed following modern psychometric standards and seeks to overcome the pitfalls identified in the reviewed existing instruments by rooting the items in a robust theoretical framework, using a panel of experts and cognitive interviews with the target sample for content validity and item readability and interpretation, principal component (PCA) and principal axis factoring (PAF) exploratory factor analysis for construct validity, and Cronbach α and item-total *r* for internal consistency reliability. The results reveal a short and parsimonious unidimensional scale composed of 6 Likert-type items that are consistent with theoretical expectations and that provide evidence for strong validity and reliability psychometric properties. This promising instrument can be used to advance research on factors affecting enrollments in science education careers from early stages of the educational system.

Keywords: expectancies of success, elementary education, psychometric properties, scale, validation.

Introduction

Nowadays, due to a decreasing pattern in enrolments in STEM (science, technology, engineering and mathematics) related careers, the promotion of interest in these disciplines from an early age is postulated as one of the priority objectives of science curricula worldwide (Archer et al. 2010; Kennedy, Lyons and Quinn 2014; Lyons and Quinn 2010, 2015; Osborne, Simon and Collins 2003; Potvin and Hasni 2014). Studies addressing these issues focused mostly on the development of favorable attitudes toward science from early stages (Osborne et al. 2003). Nevertheless, research on attitudes has faced a lack of conceptual clarity (Gardner 1975, 1995; Pearl 1974) and many methodological issues derived from the use of instruments with uncertain or poor psychometric properties (Blalock et al. 2008; Munby 1983, 1997). Indeed, Gardner (1975) questioned the validity of instruments that were used as unidimensional despite revealing a multidimensional structure. Munby (1983) review of existing instruments concluded that not much confidence can be placed in research on attitudes towards science due to the lack of validity and reliability evidences of the existing instruments. More recently, Blalock et al. (2008) reaffirmed these issues by identifying that there is no consensus established on which aspects should be addressed when investigating attitudes towards science and by revealing that science education measurement instruments published between 1935 and 2005 were mostly developed without a guiding theoretical framework and lacked rigorous validity and reliability psychometric evaluations. Authors (2019), after concluded that psychometrically valid and reliable instruments are scarce in the Spanish-speaking context, especially for the elementary education stage, proposed a single-item attitude towards school science instrument measuring the six more evaluated attitude constructs in science education literature. While this instrument showed great validity and reliability psychometric properties, there are study design (e.g. intervention studies) where multiple-items instruments may be preferred for its greater sensitivity in detecting differences between groups, broader coverage of the construct under study, and less prone to interpretation bias.

This situation requires the development and validation of valid and reliable instruments for the early detection of those factors that influence students' educational choices and preferences, so that appropriate policy measures and educational interventions aimed at promoting scientific vocations can be developed. Consistently, in this study we review the two most used instruments within the Spanish science education research to examine the extent to which these instruments present the conceptual and methodological issues identified internationally by Blalock et al. (2008) and Munby (1983), among many others. Subsequently, we present the development and validation of a theoretically driven, brief, self-reporting, paper and pencil instrument called SUCCESS, designed for rapid administration and easy assessment of elementary students' expectancies of success in the school science subject. This instrument was designed to address and overcome the limitations found in the existing instruments, and thus allow further valid and reliable examination of factors affecting young students interest in school science.

Existing measurement instruments for Spanish-speaking context

In Spain, this decline in enrollments in STEM fields has been largely addressed by examining the development and evolution of students' attitudes toward science, using mainly two instruments, the ROSE and the COCTS, reaching together a high impact in science education research with up to 964 citations (Table 1).

The ROSE is Vázquez and Manassero's adaptation of Schreiner and Sjøberg's (2004) instrument that was used in the international study entitled "Relevance of Science Education (ROSE)" (Schreiner & Sjøberg, 2004), which has subsequently been translated into numerous languages, including Spanish. This instrument is composed of 146 Likert type items with 4 response options, focused on measuring 5 different aspects. Three subscales are directly related to science (i.e. Opinions about science and technology; Attitudes toward science classes, and Environmental challenges) and the remaining two scales are related with the desired future job and with experiences and participation in extracurricular activities. This instrument, or part of it, has been used in the Spanish context in up to 15 relevant studies with samples from elementary to preservice teachers.

The COCTS is Vázquez, Manassero and Acevedo's Spanish adaptation of Aikenhead and Ryan's (1992) VOSTS instrument, originally designed for the measurement of students understanding of epistemological and sociological aspects of science and technology. Its Spanish version is composed of a total 35 multiple-choice items with a total of 222 response options. Each item consists of a caption in which a problem is advanced to the test taker, followed by a range of response options that reveal naïve to informed conceptions. This instrument has been used in up to 11 Spanish studies involving secondary students to in-service teachers.

Absence of theoretical framework for instrument development

The attitude construct lends itself to many different conceptualizations. Eagly and Chaiken (1995) provided a widely accepted definition that conceives attitude as "(...) a psychological tendency that is expressed by evaluating a particular entity with some degree of favor or disfavor" (p. 414). Thus, attitudes encompass those evaluative aspects of a cognitive, affective and behavioral nature referring to beliefs, thoughts, feelings and emotions towards a given objective. More specifically to scientific education, Klopfer (1971) categorized attitudes towards science as a set of affective behaviors toward science, scientists, scientific research, scientific careers, and activities related to science in general. Due to this multiplicity of conceptualizations, it is paramount that attitudinal instruments are developed according to a specific theoretical framework (Summers and Abd-El-Khalick 2018; Pearl 1974; Messick 1989; Shrigley and Koballa 1992).

Nevertheless, neither the ROSE nor the COCTS were developed according to a theoretical framework of attitudes and none of the 26 studies using these instruments (Table 1) in the Spanish contexts have discussed what conceptualization of attitudes toward science was being adopted. This absence of a guiding theoretical framework is reflected in the deficient design of both instruments. For example, the ROSE instrument is composed of several subscales measuring attitudes toward science, technology and school science indistinctively. Moreover, most items refer to both technology and science within the same subscale (i.e. "Science and technology can solve all problems"; Vázquez and Manassero 2004, p. 389). In addition, some subscales have double and triple-barreled items, like "Science and technology make our lives healthier, easier and more comfortable" (Vázquez and Manassero 2004, p. 389). As it is well known in the psychometric literature, items referring to different objects (i.e. science and technology) or with more than one meaning (i.e. easier, comfortable, healthier) introduce ambiguity and are therefore strongly discouraged; thus, a test-taker may consider that technology does make life easier and more comfortable but not healthier, or that science does help us solve problems while technology does not. Hence, this poor design clearly reveal a completely erroneous practice that ignores all the consensual recommendation from the literature related to the development and validation of measurement instrument (e.g. AERA, APA and NCME 2014, Crocker and Algina 1986; DeVellis 2017; Nunnally and Bernstein 1994).

Table 1. Characteristics of the main instruments used in the Spanish context

Instrument and studies	Total citations[a]	Target audience	Items and constructs	Theory[b]	Validity	Reliability	Item sample
ROSE Fernández-Cézar and Pinto-Solano (2017); Fernández-Cézar, Pinto-Solano, and Muñoz-Hernández (2018); Marbà-Tallada and Márquez (2010); Mazas and Bravo Torija (2018); Pérez-Franco, de Pro Bueno, and Pérez Manzano (2018); Vázquez and Manassero (2004, 2005, 2008a, 2008b, 2009a, 2009b, 2009c, 2009d, 2010, 2011).	537	Elementary Secondary Preservice teachers Adults	146 Likert type items: a) Opinions about science and technology b) Attitudes toward science classes c) Environmental challenges d) My future job e) Out of school experiences	No	Construct validity	Item total *r* = .5 - .68 for b subscale* α = .25 - .91	- I would like to become a scientist. - The natural world is sacred and should not be bothered - Science will not be useful for my future job
COCTS Acevedo, Vázquez, Manassero, and Acevedo (2002); Garcia-Ruiz, Maciel, and Vázquez (2014); Manassero and Vázquez (2001a, 2001b); Manassero, Vázquez, and Acevedo (2002, 2004); Vázquez, Manassero, and Talavera (2010); Vázquez, Acevedo, Manassero, and Acevero (2006); Vázquez and Manassero (1998, 1999a, 1999b)	427	Secondary Undergraduates Graduates In-service teachers	35 items and 222 multiple choice options: a) Science and technology b) Influence of the society in science and technology c) Influence of science and technology in society d) Scientists characteristics e) Social construction of the scientific knowledge f) Social construction of technology g) Nature of Science	No	Content validity	No	- Defining what is science is difficult, because science is complex and encompasses many things. But science PRINCIPALLY is: - A country's politics affects its scientists as they are a part of its society (i.e., scientists are not isolated from their society).

[a] Refers to the total citation registered by Google Scholar for each set of studies using the reviewed instrument, as of 22 of April 2019.
[b] Refers to whether the instrument was developed according to an attitude theoretical framework.
*Authors used just 8 items instead of the 15 total items comprising the “Attitude toward science classes” subscale.

The lack of theoretical framework is also evidenced in the COCTS instrument. While Aikenhead and Ryan (1992) designed and refined this instrument for the measurement of individuals conceptions about the nature of science and technology authors using this instrument in the Spanish context conceived it as an instrument for the measurement of attitudes toward science and technology instead of epistemological conceptions. Indeed, Manassero, Vázquez and Acevedo (2004) developed a methodological procedure to derive from Aikenhead and Ryan's (1992) original VOSTS an index that allegedly reflects individuals' attitudes towards science and technology. While creative, this approach may seem problematic since nature of science is a cognitive salient that refers to knowledge about the epistemology of science (Lederman 1992) while attitudes towards science are considered as emotional and affective outcomes (Eagly and Chaiken, 1995; Klopfer, 1971).

This lack of explicit conceptualization of the construct being studied, which has also been noted for decades internationally (e.g. Schibeci 1984, Shrigley and Koballa 1992), makes comparisons between studies problematic and, more critically, calls into question the extent to which the validity and reliability psychometric properties reported can be robust or the results derived from the use of these instruments can be trustworthy (Pearl 1974), given that high indices of reliability of instruments without clear conceptualization of the construct under study do not guarantee that the instrument can be used with confidence (Summers and Abd-El-Khalick 2018).

Methodological issues and lack of validity and reliability evidences

Although these instruments have been translated and widely used in the Spanish context, they have not been subjected to cross-cultural validation procedures involving back-translation (AERA et al. 2014) and rigorous and informed validation studies, nor have their psychometric properties been systematically reported in the literature. Thus, the COCTS instrument has been subjected only to content validity using a panel of experts. None of the 11 studies using it in the Spanish context has provided any evidence of other type of essential validity or reliability indices (Table 1). Similarly, authors using the ROSE instrument have reported only evidence of construct validity, although with misinformed practices and severe methodological issues. More specifically, Vázquez and Manassero (2004) subjected the "Opinions about Science and Technology" subscale to exploratory factor analysis (EFA), revealing a problematic structure composed of four constructs with three items displaying cross loadings between two constructs, and with one construct being measured by only two items. Authors wrongly argued that "(…) these three items have been assigned to the first factor because of its greater loading" [own translation] (Vázquez and Manassero, 2004, p. 392), which, in fact, is not the case when inspecting the EFA results (p. 391) where the 11^{th} item was retained in the first factor despite showing greater loadings on the fourth factor (.503) than in the first one (.482). A better procedure would consist of repeating the factor analysis without the problematic items (Lloret-Segura, Ferreres-Traves, Hernández-Baeza and Tomás-Marco 2014).

Likewise, Vázquez and Manassero (2005) analyzed the factorial structure of the "Attitude toward science classes", revealing four factors which also were problematic due to 5 items with cross loadings that have not been deleted, factors with only two items, and factors with reliability results below the α > .70 cutoff established for exploratory studies (Nunnally and Bernstein 1994). Moreover, Vázquez and Manassero (2009a) provided the factorial structure of the entire ROSE instrument, deriving up to 20 distinct factors, of which 12 showed deceptively low internal consistency results, and even one factor with a negative Cronbach α of -.196 (Vázquez and Manassero, 2009a, p. 323), highlighting clear methodological issues and misapplication of psychometric indices. Surprisingly and ominously, this new factorial structure was completely different from the one previously reported in Vázquez and Manassero (2004, 2005) studies, which again calls into questions about the validity of the results reported and the confidence with which the ROSE instrument can be used. In short, these examples show a clearly inadequate and misinformed methodological practice (Costello & Osborne 2005; Lloret-Segura et al. 2014) and raise important questions about the psychometric properties reported for the ROSE instrument.

In terms of reliability evidence, item-total correlations for the ROSE instrument have been reported in only 2 studies (i.e. Fernández-Cézar and Pinto-Solano 2017; Fernández-Cézar, Pinto-Solano and Muñoz-Hernández 2018) and the Cronbach α in only 5 of the 15 studies, although most of the times with results lower α > .70 and with ranges varying from a minimum of α = .25 to a maximum of α = .91. For example, Vázquez and Manassero (2009a) reported extremely low reliability results for 12 extracted factors, and Vázquez and Manassero (2009d) results of internal consistency were also extremely insufficient in two of the four factors retained.

In addition to the lack of robust psychometric properties, there are other methodological issues related with the use of the ROSE instrument in the Spanish context. Surprisingly, most studies using the ROSE instrument (with few exceptions e.g. Ferández-Cézar and Pinto-Solano 2017) analyzed the results without considering the factorial structure of each subscale. Instead, authors (e.g. Vázquez and Manassero 2008; Marbá-Tallada and Márquez 2010; Pérez-Franco, Pro-Bueno and Pérez-Manzano 2018) analyzed each item independently, making up to a total of 25 individual comparisons, and in some cases even 50 comparisons when the gender variable was examined, without using any *p* value correction, so the likelihood that the results derived from these studies are only a reflection of type I errors rather than significant results is very high.

This brief and non-exhaustive review reveals that two of the most used measurement instruments in the Spanish science education research have even worse limitations as the international instruments reviewed by Munby (1983) and Blalock et al. (2008). Despite this lack of validity evidences and the extremely low reliability of these instruments, they have been used and promoted as robust in up to 37 highly cited studies that informed educational interventions within the line of research concerning attitudes towards science in the Spanish context. It can therefore be concluded that the results derived from the use of the instruments reviewed are at least doubtful, and that the lack and limitations of the psychometric evidence provided make their use totally inadvisable. This situation reveals an urgent need to develop psychometrically robust measurement instruments for the assessment of factors that may influence the interest and vocational development of Spanish students towards scientific disciplines.

Aim and rationale for this study

Given the multifaceted conceptualization of attitudes, the lack of consensus in the science literature about what is measured when attitudes towards science are approached (Osborne et al. 2003; Ramsden 1998), and due to several and perpetual conceptual and methodological issues associated with measuring attitudes toward science, as reflected in previous sections and in extensive international literature reviews (e.g. Aydeniz and Kotowski 2014; Blalock et al. 2008; Munby 1983, 1997; Potvin and Hasni 2014), in this study we stray from attitudes toward science and draw on social psychology theories of achievement motivation to develop a theoretically driven and psychometrically valid and reliable instrument. This decision is also supported by past and recent research indicating that student's self-efficacy and expectancies of success in science and mathematics are low (Sellami, El-Kassem, Al-Qassass and Al-Rakeb 2017; Wigfield and Eccles 1990; Wigfield 2004), factors that may negatively affect the development of intentions to enroll in future science studies. Given that students' self-efficacy influences the type of activities in which students judge themselves to be competent (Bandura, Barbaranelli, Caprara and Pastorelli 2001), and given that students with a high self-concept are more likely to persist in science and engineering careers (Mau 2003), the detection and understanding of expectancies of success in science from the elementary stages of the education system may be more beneficial for counteracting the loss of interest in science at later stages than the "(...) somewhat nebulous, often poorly articulated and not well understood" construct of attitudes toward science (Osborne et al. 2003, p. 1049). Despite existing literature calling for the evaluation of expectancies of success from an early age, the Spanish instruments revised do not tackle this aspect and, therefore, there is a need to address this gap in the literature. Consequently, in the next sections, the development and validation of a valid and reliable instrument rooted in a conceptually valid theory and developed following modern psychometric standards is reported.

Development and validation approach

The process of developing and validating an instrument requires several phases and numerous steps involving informed theoretical and methodological decision. In the literature there are numerous classical (e.g. Churchill 1979; Nunnally and Bernstein 1994) and contemporary (e.g. Boateng, Neilands, Frongillo, Melgar-Quiñonez and Young 2018; DeVellis 2017) approaches that guide this process.

Consistently, the SUCCESS instrument was developed and validated following classical test theory (CTT) principles (Crocker and Algina 1986; DeVellis 2017; Nunnally and Bernstein 1994) and the Standards for Educational and Psychological Testing (AERA et al. 2014). While the development and validation of an instrument does not follow a universal or consensual protocol, we have followed a procedure consisting of three phases and 5 total steps, based on that of Boateng et al. (2018), which groups together the recommendations of the specialized literature. The specific procedure followed in this study is reported in Fig. 1.

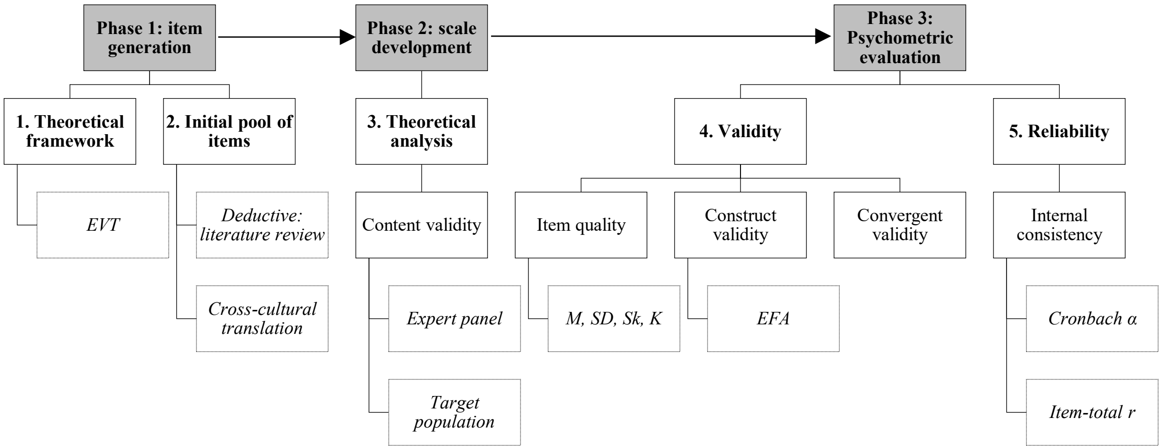


Fig. 1: Roadmap followed for instrument development and validation

In the first phase, the construct under study (i.e. expectancy of success) was defined and specified by drawing on theories of achievement motivation. Next, based on the theoretical underpinnings adopted, a deductive approach based on literature review was used for the development and adaptation from existing literature an initial list of potentially relevant items. In the second phase, one panel of experts composed of university professors and elementary education teachers examined the content validity of the proposed items, leading to a refined pool of items. Next, the readability of the items was examined through cognitive interviews with the target population, which led to the final pool of items. Finally, in the third and last phase, the retained items were administered to a large-scale sample, consisting of Spanish elementary school students, and the psychometric properties in terms of item quality, construct validity, convergent validity and internal consistency reliability were analyzed. For in-depth details of the methods and procedures used in each phase and step, see Online Resource 1.

Phase 1: Item generation

Theoretical framework

The SUCCESS instrument is based on Eccles and colleagues Expectation-Value theory (EVT), which aims to explain the relationship between motivations and academic achievement by linking performance, persistence and choice of a given that or activity with individual's beliefs about his or her expectancies of success and the value he or she places on that task or activity (Eccles et al. 1983; Eccles, Wigfield, Harold and Blumenfeld 1993; Eccles and Wigfield 1995; Wigfield and Cambria 2010; Wigfield and Eccles 1992, 2000, 2002). The first key construct of the theory is the "expectancies of success”, defined by Eccles et al. (1983) as an individual's beliefs about the success he or she will have in a given task or activity, whether in the immediate or long-term future. Conceptually, there are two dimensions that underlie individuals’ expectancies of success: (i) beliefs in one's own ability, which refers to the individual's perception of own competence to be successful in a given activity or task in the present, and (ii) expectations, which refers to the individual's perception of successfully performing that given activity or task in the future. However, empirical studies have shown that beliefs and expectations are highly correlated and cannot be differentiated empirically (Eccles et al. 1993; Wigfield and Eccles 2000). Therefore, expectancies of success include internal beliefs about both the ability to perform a task in the present and in the future. For example, a student enrolled in elementary education may feel highly competent in learning science, which may lead to the development of positive beliefs about his or her ability to succeed in future science courses or subjects. This expectation may foster the selection of science subjects during secondary education or the engagement with extracurricular science activities, and, eventually, the selection of a science related degree. Recently, the EVT model has been used to explain the selection of STEM subjects (Andersen and Ward 2014), so it is postulated as an ideal theoretical framework for examining disinterest in STEM related degrees.

Development of initial pool of items

The potential items (Table 2) were selected from six existing expectancies of success instruments (i.e. Ball, Huang, Cotten and Rikard 2017; Guo, Marsh, Parker, Morin and Dicke 2017; Guo, Marsh, Parker, Morin and Yeung 2015; Kosovich, Hulleman, Barron and Getty 2015; Thomas and Strunk 2017; Xu 2017). Following a cross-cultural adaptation procedure, items were forward and back-translated by two external bilingual university teachers until consensus was reached and the Spanish versions of the items were equivalent to the original language (Borsa, Damásio and Bandeira 2012). Next, items were adapted to refer to school science. For example, the item “If I don’t understand something in *math*, I often think I’ll never understand it” (Xu, 2017, p. 13) was adapted to “If I don't understand something in *school science classes*, I often think I'll never understand it”. Likewise, items referring to a general domain, like “I am able to do my schoolwork well” (Ball et al. 2017, p. 337) were reformulated to tackle the school science domain (i.e. I am able to do my *school science* schoolwork well”.

Phase 2: scale development

Content validity

The panel of experts reached an agreement above 80% on 11 out of the 18 items included in the initial pool (Table 2). The main reason for rejecting items was the length and the difficulty of them that could affect students' comprehension. Thus, long items or items composed by two sentences as “If I don't understand something in school science classes, I often think I will never understand it” or “Even if I do my school science homework, I don't understand anything during science classes" have been assessed as not relevant for elementary students, especially those enrolled in 3rd and 4th grade. In relation to items of very similar meaning, priority has been given to those whose wording was more intelligible. For example, the item “I learn things very quickly in school science” was deemed more relevant than “I'm sure I can understand teacher explanation of school science concepts”; or instead of “School science is not one of my strengths”, the item “I am not a student who does well in school science” was preferred.

Cognitive interviews with target audience

Think-aloud interviews revealed comprehension problems and incongruent interpretations for three of the retained items. Most students were confused about the item "I know I can do well in school science" and while some students interpreted it in terms of being able to learn the school science content, others though that it assessed whether they are successful in passing school science exams school. In addition, the item "I think I can succeed in school science classes" led also to different interpretations among students. While the item was originally designed to measure students expectancies of success in terms of achievement (i.e. passing an exam with good grades), some 3rd and 4th graders understood the items in relation to being successful in "Bringing homework to class (…)", "Knowing what to answer when the teacher asks you" or "Having classmates listen to you [when doing group work]". Finally, the interviews revealed that the item "School science is more difficult for me than for the rest of my classmates" was poorly designed and did not measure expectancies of success. Rather, students argued that they cannot know their peer's achievement and the difficulties they experience in school science, and that they do not know how to objectively respond to the item. For these reasons, these three items were excluded and an initial list of 8 items constituted the final questionnaire that was administered to a large-scale sample and submitted to construct validity (i.e. structural and convergent validity) and reliability (i.e. internal consistency) indices. Table 2 reports the process followed for retaining and excluding items.

Table 2. Theoretical analysis of the items

Initial item pool	Panel of experts		Students
	Agreement	[1]Decision	[2]Decision
If I don't understand something in school science classes, I often think I will never understand it (r)	37.5	Excluded	-
Even if I do my school science homework, I don't understand anything during science classes (r)	50	Excluded	-
I am able to do my science schoolwork well	87.5	Retained	Retained
I can do well in school science exams	62.5	Excluded	-
I can learn the content of school science classes	50	Excluded	-
School science is very hard for me (r)	87.5	Retained	Retained
I can get good grades in school science	100	Retained	Retained
I am not a student who does well in school science (r)	87.5	Retained	Retained
I learn things quickly in school science	100	Retained	Retained
School science is more difficult for me than for the rest of my classmates (r)	75	Retained	Excluded
School Science is not one of my strengths (r)	37.5	Excluded	-
I know I can do well in school science.	87.5	Retained	Excluded
I think I can succeed in school science	87.5	Retained	Excluded
I'm sure I can understand teacher explanation of school science concepts	62.5	Excluded	-
I usually do well in school science	87.5	Excluded	-
Sometimes, I have difficulties understanding school science content (r)	87.5	Retained	Retained
It is very difficult for me to finish school science homework's (r)	87.5	Retained	Retained
I am very good at school science	100	Retained	Retained

[1]Retained items after inter-rater agreement with panel of experts
[2]Retained items after think-aloud interviews with students (in bold)
(r) Items that should be reverse scored

Phase 3: Psychometric evaluation

Sample and procedure

A sample of 442 elementary school students was constituted using a non-probability convenience sampling technique. Students participated in a week-long intensive curriculum enrichment program in the city of Burgos, Spain. The CRIEB school center aims to provide complementary education to students enrolled in schools located in rural areas of the province of Burgos by assisting to a week-long enrichment curriculum program in the urban area of Burgos. The students included in this study can be considered as representatives of students from rural schools in the province of Burgos, but not from urban students in Burgos.

The instrument was administered in paper and pencil format upon arrival of the students at the CRIEB school center in the presence of the first author. In order to avoid socially desired responses, classroom teachers from the visiting schools were neither present nor involved during the administration of the instrument. Written informed consent was obtained from all students' parents or legal custodians, and students were informed of the voluntary and anonymously nature of their participation and that the responses would not affect their school grades. At the moment of the instrument administration, students were enrolled in 3rd to 6th elementary grades. After excluding incomplete questionnaires, the valid sample was constituted by a total of 418 students (47.1% girls) ranging between 8 and 12 years old (M = 10.16, SD = 1.12).

Item quality

The descriptive analysis revealed that each response option has been endorsed in all items, although with a slight tendency in choosing option 4 out of 5. This is reflected in the mean of the items, in which all but one has a value close to 4, with a standard deviation of approximately ±1 for each item. Skewness values ranged from -1.18 to -0.10 and kurtosis from -0.48 to 1.22, results that fit the recommended values in the literature (Table 3).

Table 3. Descriptive statistics for initial pool of items

Items	1	2	3	4	5	*M*	*SD*	*Sk*	*K*
I am very good at school science	7.9	8.6	29.7	35.4	18.4	3.48	1.13	-0.58	-0.22
I am not a student who does well in school science (r)	6.5	7.4	20.6	31.9	33.6	3.79	1.18	-0.82	-0.09
I can get good grades in school science	5.6	5.1	14.7	39.1	37.5	3.98	1.10	-1.18	0.96
School science is very hard for me (r)	6.2	10.3	28.3	29.5	25.7	3.58	1.16	-0.50	-0.48
I learn things quickly in school science	4.4	10	41	27.7	16.8	3.42	1.02	-0.23	-0.23
I can do my science schoolwork well	3.2	2.7	19.2	42.2	32.7	3.99	0.96	-1.04	1.22
It is very difficult for me to finish school science homework	8.6	12.9	19.4	26.8	32.3	3.61	1.29	-0.58	-0.77
Sometimes, I have difficulties understanding school science content	13.3	28.9	22.7	24.5	10.6	2.90	1.22	0.10	-1.02

n = 418
M: mean; *SD*: standard deviation; *SK*: skewness; *K*: kurtosis.

Construct validity

The Kaiser-Meyer-Olkin (KMO) measure of sampling adequacy was .815, indicating a high correlation between the variables and that the result derived from factor analysis will be robust (Lloret-Segura et al. 2014). Bartlett's test of Sphericity reached statistical significance (χ^2 = 550.683, $p < .001$), which suggest that the variables included in the correlation matrix can be subject to factor analysis. The sample size reached a 50:1 distribution, well above the minimum criterion set at 20:1 (20 individuals x 8 items = 160 minimum individuals needed), so the EFA results are expected to be reliable and accurate (Costello and Osborne 2005).

Following Kaiser's criterion (1970), both PCA and PAF extraction methods revealed the existence of two factors with initial eigenvalues above the criteria K > 1 (Factor 1: 3.079; Factor 2: 1.098). However, Cattell's (1966) scree plot indicated that a unidimensional factor structure suits best the data, decision that was further supported by results of Horn's parallel analysis (1965) showing just one factor with eigenvalues exceeding those obtained from a randomly-generated sample of the same size (8 items, 418 subjects, 100 replications), as reported in Fig. 2. Given that Kaiser's criterion is well-known for overestimating the number of factors to be extracted and retained (Costello and Osborne 2005; Kline 2014; Patil, Singh, Mishra and Donavan 2008), and because the existing consensus on considering Horn's parallel analysis as the most reliable indicator for factor extraction decisions (Choi et al. 2001; Hayton et al. 2004), we decided to retain only one factor. This unidimensional factorial structure is also consistent with the theoretical conceptualization of the expectancies of success construct (Eccles and Wigfield 1995; Wigfield and Cambria 2010).

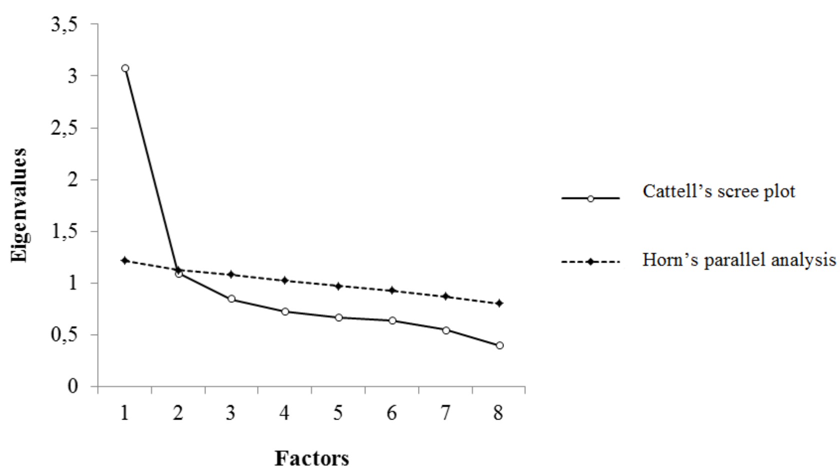


Fig. 2: Comparison between scree plot and parallel analysis results

When extracting a single-factor structure, both PCA and PAF revealed two items (i.e. Sometimes, I have difficulties understanding school science content; It is very difficult for me to finish school science homework's) with communalities well below the criterion >.3, showing that these items share little variance with the remaining ones. After removing these items, PCA and EFA was rerun and a parsimonious unidimensional structure composed of 6 items explaining a total of 45.606% of the variance was finally established (Table 4). All items included in this factor had loadings well above the >.30 criteria, with ranges between .634 and .734 for PCA and .535 and .671 for PAF. Communalities of retained items were above the minimum criteria established for PCA, with ranges between .401 and .538, however, one item (i.e. I am able to do my science schoolwork well) revealed a communality of .286 in the PAF extraction procedure. Given that the loading of this item was high, and since its exclusion reduces the internal consistency of the instrument, it was decided to retain it for the final factorial structure. In addition, since the EFA literature suggest retaining at least 3 or 4 items per factor extracted, with a sample of at least 200 cases (Fabrigar, Wegener, Maccallum y Strahan 1999; Lloret-Segura et al. 2014), the factorial results for the SUCCESS instrument can be considered robust.

Table 4. Results of Principal Component Analysis (PCA) and Principal Axis Factoring (PAF)

Items	[1]PCA		[2]PAF	
	Factor 1	h^2	Factor 1	h^2
School science is very hard for me (r)	.734	.538	.671	.451
I am not a student who does well in school science (r)	.705	.497	.628	.395
I am very good at school science	.672	.451	.581	.338
I can get good grades in school science	.655	.428	.560	.314
I learn things quickly in school science	.648	.421	.556	.309
I am able to do my science schoolwork well	.634	.401	.535	.286
Eigenvalues	2.736			
% of variance explained	45.606			

[1]Principal Component Analysis extraction method with Promax rotation
[2]Principal Axis Factoring extraction method with Promax rotation
h^2: Items communalities
(r): Items that should be reverse scored

Convergent validity

Spearman rho test revealed that the retained items from the SUCCESS instrument are moderately and statistically significant correlated (Table 5). The range of correlations varies between a minimum of $r = .303$ and a maximum of $r = .477$, results that confirm the convergent validity of the extracted factor.

Table 5. Spearman rho correlation matrix

Items	1	2	3	4	5	6
1. I am very good at school science	-	-.402	.374	-.382	.348	.325
2. I am not a student who does well in school science (r)		-	-.469	.477	-.323	-.372
3. I can get good grades in school science			-	-.353	.325	.413
4. School science is very hard for me (r)				-	-.477	-.316
5. I learn things quickly in school science					-	.303
6. I am able to do my science schoolwork well						-

All correlations are significant at $p < .01$

Internal consistency reliability

The set of retained items showed satisfactory levels of internal consistency (Table 6). Cronbach's α was .761, above the minimum criterion of .70 needed for exploratory studies (Nunnally and Bernstein 1994), and although apparently this value is not very high, it is consistent with recommendations related to the use of self-report scales with few items and for samples composed of young individuals (Newman and McNeil 1998). In addition, all items included in the extracted factor had item-total correlations higher than the minimum criterion of .30, with a range between .462 and .565, which provides additional evidence of acceptable internal consistency.

Table 6. Internal consistency results

Items	Item total *r*	Cronbach α
School science is very hard for me (r)	.565	
I am not a student who does well in school science (r)	.533	
I am very good at school science	.498	.761
I can get good grades in school science	.482	
I learn things quickly in school science	.474	
I am able to do my science schoolwork well	.462	

Discussion

This study provides a measure of expectancies of success which is, to the best of author's knowledge, the first effort to develop and validate, following robust psychometric criteria, an instrument specifically focused on the school science domain and for its use with Spanish-speaking elementary students. Through a multi-stage approach and multiple psychometric assessments, this study provides a conceptually robust and psychometrically valid and reliable measurement instrument that can be used to address from early stages of the educational system the decreasing pattern in STEM degrees enrollments. Overall, the SUCCESS instrument was found to have excellent indications of content and construct validity, promising reliability indices, and to be enough parsimonious for students enrolled as low as 3rd grade of elementary education.

The development and validation of the SUCCESS instrument addresses validity and reliability concerns identified in both existing Spanish attitude instruments and international literature (Aydeniz and Kotowski 2014; Blalock et al. 2008). Theoretical and psychometric consideration during the development and validation of the SUCCESS instruments included: (i) the use of a solid theoretical model, (ii) the adaptation of existing items in the literature through an intercultural adaptation procedure, (iii) the use of an iterative process of item analysis with both a panel of experts and the target sample for rejecting irrelevant or difficult items, (iv) the use of a large sample of students and multiple robust psychometric tests to establish the structural and convergent validity, and the quality and reliability of the retained items, and (v) the adoption of robust scale development and validation recommendations that are in line with modern psychometric standards.

When compared with the ROSE or COCTS instrument, the SUCCESS scale clearly displays better psychometric properties and therefore it can be used with more confidence. Likewise, if the SUCCESS scale is assessed with the rubric used by Blalock et al. (2008), it is obtained that a final psychometric property score of 17 points (Theory = 3; Reliability = 2; Validity = 6; Dimensionality = 6), which is above the median score of 9.5 points reported by Blalock et al. (2008, p. 968) review. Indeed, the SUCCESS instrument score, and therefore its psychometric properties, are higher and of better quality than all but two (97%) of the 66 measurement instruments published between 1935 and 2005 in science education research and analyzed by Blalock et al. (2008), being surpassed only by German's (1988) Attitude Toward Science in School Assessment and by Noll's (1935) Scientific Attitude instrument.

Several implications for science education research are derived from this study. First, because of its short length and rapid administration, the SUCCESS instrument can be used to examine the relationship between students' expectancies of success and other relevant educational outcomes (e.g., values or motivation related to studying science), especially in longitudinal studies where several data collection phases are needed (Campbell, Stanley and Cage 1963). The results derived can contribute to a deeper understanding of the factors that may improve or hinder students' expectancies of success and to an understanding of students' declining interest towards STEM degrees.

Second, future studies can validate the SUCCESS scale for students enrolled in middle and secondary school, thus allowing cross-sectional studies that will provide information on the development and progression of students' expectancies of success in science. This should be helpful in identifying at what point in the education system students' expectancies of success in science starts to decrease. In this way, studies could be conducted to analyze what factors at that educational stage are causing this decline, and what educational measures should be taken to improve this situation.

Finally, additional studies are required to further assess SUCCESS instrument reliability, especially test-retest reliability if the proposed instrument is planned to be used in intervention studies with pre-posttest designs. Although Nunnaly and Bernstein (1994) advocated against the use of test-retest indices for estimating reliability because of concerns related to carryover effects by respondent's remembering previous responses, this form of reliability may be crucial when studies are focused on improving a salient attribute. In this situation, inferences from the efficacy of the intervention can only be made if the ability of the instrument to measure true change on the construct under study is first established (Polit 2015). Also, in response to the criticisms of Blalock et al. (2008) about the lack of studies that explore in greater depth the psychometric properties of the instruments beyond the original validation study, it would be necessary to study the construct validity of the SUCCESS scale using CFA and Item Response Theory (IRT) validation methods. None of these aspects could have been addressed in this article because IRT validation methods require larger sample size than the one drawn in this study, and because CFA cannot be computed on the same sample used for EFA, being generally advised to randomly split the sample in two subsamples and perform EFA with one subsample and CFA with the other one (Lloret-Segura et al. 2014). Since our sample size was not large enough, we limited our construct validity through the study of the structural validity analysis by computing EFA.

Despite these limitations, this study advanced a promising measure to evaluate the expectancies of success of Spanish students enrolled in elementary education and used a robust protocol that can inform future scale development and validation studies in science education research. Unlike most existing measures used internationally or in the Spanish context, the SUCCESS has been specifically developed following a theoretical framework and has been submitted to rigorous psychometric evaluations, so it could be therefore considered as a valuable instrument for advancing research on factors affecting enrollments in science education degrees from early stages of the educational system.

Compliance with Ethical Standards

Conflict of Interest: On behalf of all authors, the corresponding author states that there is no conflict of interest.

Ethical approval: "All procedures performed in studies involving human participants were in accordance with the ethical standards of the institutional and/or national research committee of the PhD program in Education of the University of Burgos (Spain) and with the 1964 Helsinki declaration and its later amendments or comparable ethical standards."

References

References marked with an asterisk (*) indicate the Spanish studies included in Table 1 that were reviewed in this study.

Authors (2019) *PLoS ONE.*

*Acevedo, J. A., Vázquez, A., Manassero, M. A., & Acevedo, P. (2002). Persistencia de las actitudes y creencias CTS en la profesión docente [Persistence of STS attitudes and beliefs in the teaching profession]. *Revista Electrónica de Enseñanza de Las Ciencias*, 1(1), 1–27.

AERA, APA & NCME. (2014). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association

Andersen, L., & Ward, T. J. (2014). Expectancy-value models for the STEM persistence plans of ninth-grade, high-ability students: A comparison between Black, Hispanic, and White students. *Science Education*, 98(2), 216–242. https://doi.org/10.1002/sce.21092

Aikenhead, G. S., & Ryan, A. G. (1992). The development of a new instrument: "Views on science-technology-society" (VOSTS). *Science Education*, 76, 477-491. https://doi.org/10.1002/sce.3730760503

Archer, L., Dewitt, J., Osborne, J., Dillon, J., Willis, B., & Wong, B. (2010). "Doing" science versus "being" a scientist: Examining 10/11-year-old schoolchildren's constructions of science through the lens of identity. *Science Education,* 94(4), 617–639.

https://doi.org/10.1002/sce.20399

Aydeniz, M., & Kotowski, M. R. (2014). Conceptual and methodological issues in the measurement of attitudes towards science. *Electronic Journal of Science Education*, 18(3), 1–24.

Ball, C., Huang, KT., Cotten, S.R., & Rikard, R. B. (2017). Pressurizing the STEM pipeline: An expectancy-value theory analysis of youths' STEM attitudes. *Journal of Science Education and Technology,* 26(372). https://doi.org/10.1007/s10956-017-9685-1

Bandura, A., Barbaranelli, C., Caprara, G. V., & Pastorelli, C. (2001). Self-efficacy beliefs as shapers of children' s aspirations and career. *Child Development*, 72(1), 187–206.

Blalock, C. L., Lichtenstein, M. J., Owen, S., Pruski, L., Marshall, C., & Toepperwein, M. A. (2008). In pursuit of validity: A comprehensive review of science attitude instruments 1935-2005. *International Journal of Science Education*, 30(7), 961–977. https://doi.org/10.1080/09500690701344578

Boateng, G. O., Neilands, T. B., Frongillo, E. A., Melgar-Quiñónez, H. R., & Young, S. L. (2018). Best practices for developing and validating scales for health, social and behavioral research: A primer. *Frontiers in Public Health*, 6(149), https://doi.org/10.3389/fpubh.2018.00149

Borsa, J. C., Damásio, B. F., & Bandeira, D. R. (2012). Cross-cultural adaptation and validation of psychological instruments: some considerations. *Paidéia*, 22(53), 423–432.

Campbell, D. T., Stanley, J. C., & Cage, N. L. (1963). *Experimental and quasi-experimental designs for research*. Boston, MA: Houghton, Mufflin and Company.

Catell, R. B. (1966). The scree test for the number of factors. *Multivariate Behavioral Research*, *1*, 245–276.

Choi, N., Fuqua, D. R., & Griffin, B. W. (2001). Exploratory analysis of the structure of scores from the multidimensional scales of perceived self-efficacy. *Educational and Psychological Measurement*, 61(3), 475–489. https://doi.org/10.1177/00131640121971338

Costello, A. B., & Osborne, J. W. (2005). Best practices in exploratory factor analysis: Four recommendations for getting the most from your analysis. *Practical Assessment, Research & Education*, 10(7), 1–9.

Crocker, L., & Algina, J. (1986). *Introduction to classical and modern test theory*. New York: CBS College Publishing.

Churchill, G. A., Jr. (1979). A paradigm for developing better measures of marketing constructs. *Journal of Marketing Research*, 16, 64–73.

DeVellis, R. F. (2017). *Scale development. Theory and Applications*. Los Angeles: SAGE.

Eagly, A. H., & Chaiken, S. (1995). Attitude strength, attitude structure, and resistance to change. In R. E. Petty & J. A. Krosnich (Eds.), *Ogio State University series on attitudes and persuasion, Vol. 4. Attitude strength: Antecedents and consequences* (pp. 413–432). New York: Lawrence Erlbaum Associates, Inc.

Eccles, J. S., Adler, T. F., Futterman, R., Goff, S. B., Kaczala, C. M., Meece, J. L., & Midgley, C. (1983). Expectancies, values, and academic Behaviors. In J. T. Spence (Ed.), *Achievement* *and Achievement Motivation* (pp. 75-146). San Francisco, CA: W. H. Freeman.

Eccles, J. S., & Wigfield, A. (1995). In the mind of the achiever: the structure of adolescents' academic achievement related-beliefs and self-perceptions. *Personality and Social Psychology Bulletin*, 21, 215–225.

Eccles, J. S., Wigfield, A., Harold, R. B., & Blumenfeld, P. B. (1993). Age and gender differences in children's self- and task perceptions during elementary school. *Child Development*, 64, 830–847.

Fabrigar, L. R., Wegener, D. T., Maccallum, R. C., & Strahan, E. J. (1999). Evaluating the use of exploratory factor analysis in psychological research, Psychologial Methods, 4(3), 272–299.

*Fernández-Cézar, R., & Pinto-Solano, N. (2017). Actitud hacia las clases de ciencias naturales en la educación primaria en España [Attitude towards natural Science classes in primary education in Spain]. *Revista Electrónica de Investigación Educativa*, 19(4), 112–123. https://doi.org/10.24320/redie.2017.19.4.1393

*Fernández-Cézar, R., Pinto-Solano, N., & Muñoz-Hernández, M. (2018). ¿Mejoran los proyectos de divulgación con experimentación la actitud hacia las clases de ciencias? [Do outreach projects with experimentation improve attitudes towards Science classes?] *Revista de Educación*, 381(Julio-Septiembre), 285–307. https://doi.org/10.4438/1988-592X-RE-2017-381-389

*García-Ruiz, M., Maciel, S., & Vázquez, A. (2014). La ciencia, la tecnología y la problemática socioambiental: secuencias de enseñanza-aprendizaje para promover actitudes adecuadas en los futuros profesores de Primaria [Science, technology and socio-environmental problems: teaching-learning sequences to promote appropriate attitudes in future primary school teachers]. *Revista Electrónica de Enseñanza de Las Ciencias*, 13(3), 267–291.

Gardner, P. L. (1975). Attitudes to science: A review. *Studies in Science Education*, 2, 1–41.

Gardner, P. L. (1995). Measuring attitudes to science: Unidimensionality and internal consistency revisited. *Research in Science Education*, 25(3), 283–289.

Germann, P. J. (1988). Development of the attitude toward science in school assessment and its use to investigate the relationship between science achievement and attitude toward science in school. *Journal of Research in Science Teaching*, 25, 689–703.

Guo, J., Marsh, H. W., Parker, P. H., Morin, A. J. S., & Dicke, T. (2017). Extending expectancy-value theory predictions of achievement and aspirations in science: Dimensional comparison processes and expectancy-by-value interactions. *Learning and Instruction, 49*, 81-91. https://doi.org/10.1016/j.learninstruc.2016.12.007

Guo, J., Parker, P. D., Marsh, H. W., & Morin, A. J. S. (2015). Achievement, motivation, and educational choices: A longitudinal study of expectancy and value using a multiplicative perspective. *Developmental Psychology*, *51*(8), 1163-1176. http://dx.doi.org/10.1037/a0039440

Hayton, J. C., Allen, D. G., & Scarpello, V. (2004). Factor retention decisions in exploratory factor analysis: A tutorial on parallel analysis. *Organizational Research Methods*, 7(2), 191–205. https://doi.org/10.1177/1094428104263675

Horn, J. L. (1965). A rationale and test for the number of factors in factor analysis. *Psychometrika*, 30, 179–185.

Kaiser, H. (1970). A second generation Little Jiffy. *Psychometrika*, 35, 401–415.

Kennedy, J. P., Lyons, T., & Quinn, F. (2014). The continuing decline of science and mathematics enrolments in Australian high schools. *Teaching Science*, 60(2), 34–46.

Kline, P. (2014). *An easy guide to factor analysis*. New York: Routledge.

Klopfer, L. E. (1971). Evaluation of learning in science. In B. S. Bloom, J. T. Hastings, y G. F. Madaus (Eds.), *Handbook of formative and summative evaluation of student learning*. London: McGraw-Hill.

Kosovich, J. J., Hulleman, C. S., Barron, K. E., & Getty, S. (2014). A practical measure of student motivation: Establishing validity evidence for the expectancy-value-cost scale in middle school. *The Journal of Early Adolescence*, 35(5-6), 790–816. https://doi.org/10.1177/0272431614556890

Lederman, N. G. (1992). Students' and teachers' conceptions of the nature of science: A review of the research. *Journal of Research in Science Teaching*, 29(4), 331-359. https://doi.org/10.1002/tea.3660290404

Lloret-Segura, S., Ferreres-Traves, A., Hernández-Baeza, A., & Tomás-Marco, I. (2014). El análisis factorial exploratorio de los ítems: Una guía práctica, revisada y actualizada. *Anales de Psicología*, 30(3), 1151–1169. https://doi.org/10.6018/analesps.30.3.199361

Lyons, T., & Quinn, F. (2010). *Choosing science. Understanding the declines in senior high school science enrolments*. Armidale: University of New England.

Lyons, T., & Quinn, F. (2015). Understanding declining science participation in Australia: A systemic perspective. In E. K. Henriksen, J. Dillon, y J. Ryder (Eds.), Under*standing student participation and choice in science and technology education* (pp. 153–168). Springer Netherlands.

*Manassero, M. A., & Vázquez, A. (2001a). Actitudes de estudiantes y profesorado sobre las características de los científicos [Attitudes of students and faculty about the characteristics of scientists]. *Enseñanza de Las Ciencias*, 19(2), 255–268.

*Manassero, M. A., & Vázquez, A. (2001b). Instrumentos y métodos para la evaluación de las actitudes relacionadas con la ciencia, la tecnología y la sociedad [Tools and methods for the assessment of attitudes related to Science, Technology and Society]. *Enseñanza de Las Ciencias*, 20(1), 15–27.

*Manassero, M. A., Vázquez, A., & Acevedo, J. A. (2002). Opiniones sobre la influencia de la ciencia en la cultura [Opinions on the influence of Science on culture]. *Didáctica de Las Ciencias Experimentales y Sociales*, 16, 35–55.

*Manassero, M. A., Vázquez, A., & Acevedo, J. A. (2004). Evaluación de las actitudes del profesorado respecto a los temas CTS: nuevos avances metodológicos [Assessment of teacher attitudes towards STS issues: New methodological advances]. *Enseñanza de Las Ciencias*, 22(2), 299–312.

*Marbà-Tallada, A., & Márquez, C. (2010). ¿Qué opinan los estudiantes de las clases de ciencias? Un estudio transversal de sexto de primaria a cuarto de ESO [What do students think about Science classes? A cross-sectional study of sixth grade of elementary education to fourth grade of secondary education]. *Enseñanza de Las Ciencias*, 28(1), 19–30.

*Mazas, B., & Bravo Torija, B. (2018). Actitudes hacia la ciencia del profesorado en formación de educación infantil y educación primaria [Attitudes towards science of teachers in pre-school and primary education training]. *Profesorado, Revista de Currículum y Formación Del Profesorado*, 22(2), 329–348. https://doi.org/10.30827/profesorado.v22i2.7726

Mau, W. C. (2003). Factors that influence persistence in science and engineering career aspirations. *Career Development Quarterly*, 51(3), 234–243. https://doi.org/10.1002/j.2161-0045.2003.tb00604.x

Messick, S. (1989). Validity. In R.L. Linn (Ed.), *Educational measurement* (3rd ed.), (pp. 13-103). New York: Macmillan

Munby, H. (1983). Thirty studies involving 'Scientific attitude inventory': What confidence can we have in this instrument? *Journal of Research in Science Teaching*, 20, 141–162.

Munby, H. (1997). Issues of validity in science attitude measurement. *Journal of Research in Science Teaching*, 34(4), 337–341.

Newman, I., & McNeil, K. (1998). *Conducting survey research in the social sciences*. New York: University Press of America.

Noll, V. H. (1935). Measuring the scientific attitude. *Journal of Abnormal and Social Psychology*, 30, 145–154.

Nunnally, J. C., & Bernstein, I. H. (1994). *Psychometric theory*. New York: McGraw-Hill.

Osborne, J., Simon, S., & Collins, S. (2003). Attitudes towards science: A review of the literature and its implications. *International Journal of Science Education*, 25(9), 1049–1079. https://doi.org/10.1080/0950069032000032199

Patil, V. H., Singh, S. N., Mishra, S., & Donavan, T. (2008). Efficient theory development and factor retention criteria: Abandon the "eigenvalue greater than one" criterion. *Journal of Business Research*, 61, 162–170. https://doi.org/10.1016/j.jbusres.2007.05.008

Pearl, R. E. (1974). The present status of science attitude measurement; history, theory, and availability of measurement instruments. *School Science and Mathematics*, 74, 375-381.

*Pérez-Franco, D., de Pro Bueno, A., & Pérez Manzano, A. (2018). Actitudes ambientales al final de la ESO. Un estudio diagnóstico con alumnos de Secundaria de la Región de Murcia [Environmental attitudes at the end of the secondary education. A diagnostic study with secondary school students from the Region of Murcia]. *Revista Eureka Sobre Enseñanza y Divulgación de Las Ciencias*, 15(3), 3501. https://doi.org/10.25267/Rev_Eureka_ensen_divulg_cienc.2018.v15.i3.3501

Polit, D. F. (2015). Assessing measurement in health: Beyond reliability and validity. *International Journal of Nursing Studies*, 52(11), 1746–1753. https://doi.org/10.1016/j.ijnurstu.2015.07.002

Potvin, P., & Hasni, A. (2014). Interest, motivation and attitude towards science and technology at k-12 levels: A systematic review of 12 years of educational research. *Studies in Science Education*, 50(1), 85–129. https://doi.org/doi:10.1080/03057267.2014.881626

Ramsden, J. M. (1998). Mission impossible? Can anything be done about attitudes to science? *Interntional Journal of Science Education*, 20, 125–137.

Schibeci, R. A. (1984). Attitudes to science: an update. *Studies in Science Education*, 11, 26–59.

Schreiner, C., & Sjøberg, S. (2004). *ROSE: The Relevance of Science Education. Sowing the seeds of Rose. Background, rationale, questionnaire development and data collection for ROSE (The Relevance of Science Education)-a comparative study of students' views of science and science education*. Oslo: Acta Didactica.

Sellami, A., El-Kassem, R. C., Al-Qassass, H. B., & Al-Rakeb, N. A. (2017). A path analysis of student interest in STEM, with specific reference to Qatari students. *Eurasia Journal of Mathematics, Science and Technology Education,* 13(9), 6045–6067. https://doi.org/10.12973/eurasia.2017.00999a

Shirley, R. L., & Koballa, T. R. Jr. (1992). A decade of attitude research based on Hovland's elarning theory. *Science Education*, 76(1), 17-42.

Summers, R., & Abd-El-Khalick, F. (2018). Development and validation of an instrument to assess student attitudes toward science across grades 5 through 10. *Journal of Research in Science Teaching*, 55(2), 172–205. https://doi.org/10.1002/tea.21416

Thomas, J. A., & Strunk, K. K. (2017). Expectancy-value and children's science achievement: Parents matter. *Journal of Research in Science Teaching*, 54(6), 693-712, https://doi.org/10.1002/tea.21382

*Vázquez, A., Acevedo, J. A., Manassero, M. A., & Acevero, P. (2006). Evaluación de los efectos de la materia CTS de bachillerato en las actitudes CTS del alumnado con una metodología de respuesta múltiple [A multiple responde methdology evaluation of the effects of the baccalaureate STS subject on students' STS attitudes]. *Revista Eureka Sobre Enseñanza y Divulgación de Las Ciencias*, 3(3), 317–348. https://doi.org/10.25267/rev_eureka_ensen_divulg_cienc.2006.v3.i3.01

*Vázquez, A., & Manassero, M. A. (1998). Actitudes hacia la influencia de la cultura en la ciencia [Attitudes towards the influence of culture on Science]. *Cultura y Educación*, 10(3–4), 143–167. https://doi.org/10.1174/113564098320762085

*Vázquez, A., & Manassero, M. A. (1999a). Actitudes hacia la influencia de la sociedad en la ciencia y la tecnología [Attitudes towards the influence of society on Science and Technology]. *Arbor*, 162(637), 45–72.

*Vázquez, A., & Manassero, M. A. (1999b). Características del conocimiento científico: creencias de los estudiantes [Characteristics of scientific knowledge: Students' beliefs]. *Enseñanza De Las Ciencias*, 17(3), 377–395.

*Vázquez, A., & Manassero, M. A. (2004). Imagen de la ciencia y la tecnología al final de la educación obligatoria [Image of science and technology at the end of compulsory Education]. *Cultura y Educación*, 16(4), 385–398.

*Vázquez, A., & Manassero, M. A. (2005). La ciencia escolar vista por los estudiantes [School science as seen by students]. *Bordón. Revista de Pedagogía*, 57(5), 125–144.

*Vázquez, A., & Manassero, M. A. (2008a). El declive de las actitudes hacia la ciencia de los estudiantes: un indicador inquietante para la educación científica [Declining attitudes towards science among students: A disturbing indicator for Science education]. *Revista Eureka Sobre Enseñanza y Divulgación de Las Ciencias*, 5(3), 274–292. https://doi.org/10.25267/rev_eureka_ensen_divulg_cienc.2008.v5.i3.03

*Vázquez, A., & Manassero, M. A. (2008b). La vocación científica y tecnológica de las chicas en secundaria y la educación diferenciada [The scientific and technological vocation of girls in secondary education and differentiated education]. *Bordón. Revista de Pedagogía*, 60(3), 149–163.

*Vázquez, A., & Manassero, M. A. (2009a). Factores actitudinales determinantes de la vocación científica y tecnológica en secundaria [Attitudinal factors determining the scientific and technological vocation in secondary school]. *Cultura y Educacion*, 21(3), 319–330. https://doi.org/10.1174/113564009789052280

*Vázquez, A., & Manassero, M. A. (2009b). La relevancia de la educación científica: actitudes y valores de los estudiantes relacionados con la Ciencia y la Tecnología [The relevance of science education: Student attitudes and values related to Science and Technology]. *Enseñanza de Las Ciencias*, 27(1), 33–48.

*Vázquez, A., & Manassero, M. A. (2009c). La vocación científica y tecnológica: predictores actitudinales significativos [The scientific and technological vocation: Significant attitudinal predictors]. *Revista Eureka Sobre Enseñanza y Divulgación de Las Ciencias*, 6(2), 213–231. https://doi.org/10.25267/rev_eureka_ensen_divulg_cienc.2009.v6.i2.03

*Vázquez, A., & Manassero, M. A. (2009d). Patrones actitudinales de la vocación científica y tecnológica en chicas y chicos de secundaria [Attitudinal patterns of scientific and technological vocation in secondary school girls and boys]. *Revista Iberoamericana de Educación*, 50(4), 1–15.

*Vázquez, A., & Manassero, M. A. (2010). Perfiles actitudinales de la elección de ciencias en secundaria según el sexo y el tipo de educación [Attitudinal profiles of the choice of sciences in secondary school according to sex and type of education]. *Revista Electrónica de Enseñanza de Las Ciencias*, 9(1), 242–260.

*Vázquez, A., & Manassero, M. A. (2011). El descenso de las actitudes hacia la ciencia de chicos y chicas en la Educación Obligatoria [Declining attitudes towards science among boys and girls in compulsory education]. *Ciência & Educação*, 17(2), 249–268.

*Vázquez, A., Manassero, M. A., & Talavera, M. (2010). Actitudes y creencias sobre naturaleza de la ciencia y la tecnología en una muestra representativa de jóvenes estudiantes [Attitudes and beliefs about the nature of science and technology in a representative sample of young students]. *Revista Electrónica de Enseñanza de Las Ciencias*, 9(2), 333–352.

Wareing, C. (1982) Developing the WASP: Wareing attitude toward science protocol. *Journal of Research in Science Teaching*, 19(8), 639-645.

Wigfield, A. (2004). Motivation for reading during the early adolescent years. In D. S. Strickland y D. E. Alvermann (Eds.), *Bridging the literacy achievement gap in grades* 4–12 (pp. 56–69). New York: Teachers College Press.

Wigfield, A., & Cambria, J. (2010). Students' achievement values, goal orientations, and interest: Definitions, development, and relations to achievement outcomes. *Developmental Review*, 30(4), 1–35. https://doi.org/https://doi.org/10.1016/j.dr.2009.12.001

Wigfield, A., & Eccles, J. S. (1992). The development of achievement task values: A theoretical analysis. *Developmental Review*, 12, 265–310.

Wigfield, A., & Eccles, J. S. (2000). Expectancy-value theory of motivation. *Contemporary Educational Psychology*, 25, 68–81.

Wigfield, A., & Eccles, J. S. (2002). The development of competence beliefs, expectancies for success, and achievement values from childhood through adolescence. In A. Wigfield y J. S. Eccles (Eds.), *Development of achievement motivation* (pp. 91–120). San Diego: Academic Press.

Xu, J. (2017). Homework expectancy value scale for high school students: Measurement invariance and latent mean differences across gender and grade level. *Learning and Individual Differences*, *60*, 10-17. https://doi.org/10.1016/j.lindif.2017.10.003

Electronic Supplementary Material 1: Development and Validation Approach

Online resource for the parent article entitled: "Development and validation of the SUCCESS instrument: towards a valid and reliable measure of expectancies of success in Science in Elementary Education".

Introduction

In this online resource, we provide full details of the method and analysis procedure used during the development and validation of the SUCCESS instrument, aimed at measuring Spanish students' expectancies of success in the school science subject.

Phase 1: Item generation

Development of initial pool of items

To create an initial list of items, authors reviewed extant expectancies of success instruments, including the original instruments first advanced by Eccles and colleagues (Eccles et al. 1983; Eccles, Wigfield, Harold, and Blumenfeld 1993; Eccles and Wigfield 1995), and other recent research measuring expectancies for success. Authors reviewed up to 43 studies focused on measuring expectancies of success in different educational domains, and jointly decided to include those items that required less adaptation to the focus of this study (i.e. school science), drawn from six existing studies.

Phase 2: scale development

Content validity

Method and procedure. Content validity refers to whether the items included in the instrument are adequate and relevant measures of the construct under study. More specifically, content validity ensures the "(…) adequacy of content coverage and relevance for multi-item measures of a construct" (Polit 2015, p. 1750). A panel of experts composed of two university professors, one with extensive experience in science education and another with expertise in psychology research, and six elementary school teachers assessed the extent to which the items were adequate to measure students' expectations of success in school sciences according to the theoretical framework adopted, using a Likert scale of 5 responses (0 - not appropriate at all; 1- not appropriate; 2 - somewhat appropriate; 3 - appropriate; 4 - very appropriate). The panel of experts have met and jointly decided the reasons for rejecting and retaining the items, for which an agreement of at least 80% was established among experts in the evaluation of each item (Stemler 2004), which means that only those items evaluated with a score of 3 and 4 by all components of the expert panel were retained.

Cognitive interviews with target audience

Method and procedure. The content validity of the selected items was further examined by a panel composed of the target population. Twelve students enrolled in grades 3, 4, 5 and 6 of Elementary Education (three for each grade) were interviewed using Think-aloud interviews (Beatty and Willis 2007) to assess students' understanding and interpretation of the items. The protocol consisted of providing each student with a copy of the retained items, asking them to explain aloud what they were thinking when reading and answering each item. Only those items that did not show difficulties for students in all school years were retained.

Phase 3: Psychometric evaluation

Item quality

Method and procedure. The initial quality of the items was examined through descriptive statistics. Ideally, each response options should be endorsed, and the mean of each item must be close to the intermediate value of the scale (DeVellis 2017), and items should have an adequate level of skewness and kurtosis, with values in the range of ±2 (Gravetter and Wallnau 2014).

Construct validity

Method and procedure. Construct validity refers to "(…) the degree to which evidence about a measure's scores supports the inference that the construct has been appropriately represented." (Polit 2015, p. 1750). To establish the construct validity of the SUCCESS instrument, we examined its structural validity, which is one of the most important form of construct validity, that "Tests whether a measure captures the hypothesized dimensionality of a construct, using factor analysis" (Polit 2015, p. 1750).

The Kaiser-Meyer-Olkin (KMO) measure with KMO criterion > .80 and Bartlett's sphericity test with a significance value of $p < .05$ (Barlett 1950,1954) were used to determine the sampling and correlation matrix adequacy for factor analysis. The appropriate sample size was determined using the 20:1 criterion, which determines that for each item there must be at least 20 valid responses. This criterion has been shown to be very useful in producing accurate factorial solutions (Costello and Osborne 2005).

Items were subjected to both the Principal Component Analysis (PCA) and Principal Axis Factoring (PAF) extraction method, as the latter is more advisable for extracting underlying factors (de Winter and Dodou 2012; Gaskin and Happell 2014; Henson and Roberts 2006; Howard 2016; Norris and Lecavalier 2010). According to the theoretical framework adopted, only one factor is expected to be extracted, so factors were not rotated. To ensure that the number of factors extracted was adequate, Kaiser's criterion > 1 (Kaiser 1970; Williams, Onsman, and Brown 2010), Catells' scree test (Catell 1966), and parallel analysis (Horn 1965) were used. Parallel analysis is the test that provides the greatest reliability for factor extraction decisions (Choi, Fuqua and Griffin, 2001; Hayton, Allen and Scarpello 2004), and consists of comparing and maintaining only those factors whose eigenvalues exceed those obtained from a randomly generated data matrix of the same size and number of variables as the original dataset (Hayton et al. 2004). Only items with communalities and loadings of at least .30 were further retained. When items were deleted, the extraction procedure was repeated until the most parsimonious factor solution was established.

The statistical software SPSS v.24 (IBM 2016) was used for the principal component (PCA) and principal axis factoring (PAF) analyses, and the Monte Carlo PCA (Watkins 2000) software for parallel analysis.

Convergent validity

Method and procedure. The convergent validity of the SUCCESS instrument was also examined, which is another construct validity form that examines the extent to which the items of a construct are correlated with each other (Trochim and Donnelly 2006). More specifically, convergent validity was examined using Spearman's correlation between the items of the extracted factor using the statistical package SPSS v.24 (IBM 2016), since the assumption of normal distribution of the sample is violated, according to the Kolmogorov-Smirnov test.

Internal consistency reliability

Method and procedure. Reliability refers to whether "(…) scores for people who have not changed are the same for repeated measurements, under several situations" (Polit y Yang 2016, p. 25). One of the main reliability forms is the internal consistency of an instrument, which examines the "(…) degree to which the items on a scale are measuring the same underlying construct" (Polit 2015, p. 1747). Reliability of the SUCCESS instrument was determined through two internal consistency tests. We estimated the Cronbach α, with a value of $\alpha > .70$ as acceptable cutoff to demonstrate good internal consistency (Field 2009; Nunnally and Bernstein 1994), and the item-total correlation, with a criterion of $r > 0.30$ to consider satisfactory results (Field 2009), using the statistical package SPSS v.24 (IBM 2016).

References

Barlett, M. S. (1954). A note on the multiplying factors for various chi square approximations. *Journal of the Royal Statistical Society, Series B*, 16, 296–298.

Bartlett, M. S. (1950). Test of significance in factor analysis. *British Journal of Statistical Psychology*, 3, 77–85.

Beatty, P. C., & Willis, G. B. (2007). Research synthesis: The practice of cognitive interviewing. *Public Opinion Quarterly*, 71(2), 287–311. https://doi.org/10.1093/poq/nfm006

Catell, R. B. (1966). The scree test for the number of factors. *Multivariate Behavioral Research*, 1, 245–276.

Costello, A. B., & Osborne, J. W. (2005). Best practices in exploratory factor analysis: Four recommendations for getting the most from your analysis. *Practical Assessment, Research & Education*, 10(7), 1–9.

DeVellis, R. F. (2017). *Scale development. Theory and applications*. Los Angeles: SAGE.

deWinter, J. C. F., & Dodou, D. (2012). Factor recovery by principal axis factoring and maximum likelihood factor analysis as a function of factor pattern and sample size. *Journal of Applied Statistics*, 39(4), 695–710. https://doi.org/10.1080/02664763.2011.610445

Eccles, J. S., Adler, T. F., Futterman, R., Goff, S. B., Kaczala, C. M., Meece, J. L., & Midgley, C. (1983). Expectancies, Values, and Academic Behaviors. In J. T. Spence (Ed.), *Achievement and Achievement Motivation* (pp. 75-146). San Francisco, CA: W. H. Freeman.

Eccles, J. S., & Wigfield, A. (1995). In the mind of the achiever: the structure of adolescents' academic achievement related-beliefs and self-perceptions. *Personality and Social Psychology Bulletin*, 21, 215–225.

Eccles, J. S., Wigfield, A., Harold, R. B., & Blumenfeld, P. B. (1993). Age and gender differences in children's self- and task perceptions during elementary school. *Child Development*, 64, 830–847.

Field, A. (2009). *Discovering statistics using SPSS*. London: SAGE.

Gaskin, C. J., & Happell, B. (2014). On exploratory factor analysis. A review of recent evidence, an assessment of current practice, and recommendations for future use. *International Journal of Nursing Studies*, 51, 511–521. https://doi.org/10.1016/j.ijnurstu.2013.10.005

Gravetter, F., & Wallnau, L. (2014). *Essentials of statistics for the behavioral sciences*. Belmont, CA: Wadsworth.

Hayton, J. C., Allen, D. G., & Scarpello, V. (2004). Factor Retention Decisions in Exploratory Factor Analysis: A Tutorial on Parallel Analysis. *Organizational Research Methods*, 7(2), 191–205. https://doi.org/10.1177/1094428104263675

Henson, R. K., & Roberts, J. K. (2006). Use of exploratory factor analysis in published research: Common errors and some comment on improved practice. *Educational and Psychological Measurement*, 66(3), 393–416. https://doi.org/10.1177/0013164405282485

Horn, J. L. (1965). A rationele and test for the number of factors in factor analysis. *Psychometrika*, 30, 179–185.

Howard, M. C. (2016). A Review of Exploratory Factor Analysis Decisions and Overview of Current Practices: What We Are Doing and How Can We Improve? *International Journal of Human-Computer Interaction*, 32(1), 51–62. https://doi.org/10.1080/10447318.2015.1087664

IBM, C. (2016). IBM *SPSS Statistics for Windows. Version 24.0*. Armonk, NY: IBM Corp.

Kaiser, H. (1970). A second generation Little Jiffy. *Psychometrika*, 35, 401–415.

Norris, M., & Lecavalier, L. (2010). Evaluating the use of exploratory factor analysis in developmental disability psychological research. *Journal of Autism and Developmental Disorders*, 40, 8–20. https://doi.org/10.1007/s10803-009-0816-2

Nunnally, J. C., & Bernstein, I. H. (1994). *Psychometric Theory*. New York: McGraw-Hill.

Polit, D. F. (2015). Assessing measurement in health: Beyond reliability and validity. *International Journal of Nursing Studies*, 52(11), 1746–1753. https://doi.org/10.1016/j.ijnurstu.2015.07.002

Polit, D. F., & Yang, F. (2016). *Meaurement and the Measurement of Change: A primer for Health Professionals*. Philadelphia: Lippincott Williams & Wilkins.

Stemler, S. E. (2004). A comparison of consensus, consistency, and measurement approaches to estimating interrater reliability. *Practical Assessment, Research & Education*, 9(4), 1–11.

Trochim, W. M., & Donnelly, J. P. (2006). *The research methods knowledge base*. Cincinnati: Atomic Dog Publishing Inc.

Watkins, M. (2000). Determining parallel analysis criteria. *Journal of Modern Applied Statistical Methods*, 5(2), 344–346.

Williams, B., Onsman, A., & Brown, T. (2010). Exploratory factor analysis: A five-step guide for novices. *Journal of Emergency Primary Health Care*, 8(3), 1–13. https://doi.org/10.1080/09585190701763982

Appendix 6: SUCCESS instrument validation study #2

This appendix includes the full-text of the published article about second validation study of the SUCCESS instrument used in this dissertation to assess the effectiveness of the proposed pedagogical framework for STEM-based science education in improving elementary school students expectancy of success in school science.

Toma, R. B. (2019). Measuring elementary students' expectancies of success in school science: Psychometric evaluation of the SUCCESS instrument. *EURASIA Journal of Mathematics, Science and Technology Education, 15*(8), em1733. doi: 10.29333/ejmste/104695

MODESTUM
OPEN ACCESS
EURASIA Journal of Mathematics, Science and Technology Education, 2019, 15(8), em1733
ISSN:1305-8223 (online)
Research Paper https://doi.org/10.29333/ejmste/104695

Measuring Elementary Students' Expectancies of Success in School Science: Psychometric Evaluation of the SUCCESS Instrument

Radu Bogdan Toma [1*]

[1] University of Burgos, Department of Specific Didactics, SPAIN

Received 30 November 2018 ▪ Revised 24 January 2019 ▪ Accepted 26 February 2019

ABSTRACT

Background: The importance of valid and reliable instruments for the assessment of factors affecting students' interest in science encouraged the development and validation of a brief Spanish instrument for the measurement of expectancies of success in school science, named SUCCESS. In this study, the psychometric properties of the SUCCESS instrument are further evaluated using different psychometric tests and a different sample than the one included in the original validation study.

Material and methods: A sample of 313 Spanish elementary school students enrolled in 4th to 6th grade was drawn by means of convenience sampling techniques. Responses were analyzed in terms of construct and criterion validity, and two reliability indices.

Results: Results from confirmatory factor analysis established the unidimensional structure of the instrument, with great model fit indices. Correlation coefficients between the SUCCESS and external measures (i.e. intentions to enroll, enjoyableness, difficulty, auto-efficacy, utility and relevance of school science) provided evidence of criterion validity. Cronbach α and item-total correlation indices supported the internal consistency reliability of the instrument.

Conclusions: Taken together, this study further provide evidence to consider the SUCCESS as a valid and reliable tool for the measurement of Spanish elementary school students expectancies of success in school science.

Keywords: elementary education, expectancy of success, psychometric properties, school science education

INTRODUCTION

Science education worldwide continues to face a steady decline in students interested in scientific-related careers (Archer et al., 2010; Kennedy, Lyons, & Quinn, 2014; Lyons & Quinn, 2010, 2015; Osborne, Simon, & Collins, 2003; Potvin & Hasni, 2014). At the same time, the existence of valid and reliable instruments that would help understand the factors affecting students lack of engagement in science continue to be scarce in the science education literature (Blalock et al., 2008; Gardner, 1975; Munby, 1983, 1997).

This problem is even more pronounced in the Spanish context. Studies conducted in Spain about the factors influencing students' engagement and interest in science mainly focused on attitudes toward science construct (e.g. Marbá-Tallada & Márquez Bargalló, 2010; Pérez Manzano & de Pro Bueno, 2018; Vázquez-Alonso & Manassero–Mas, 2011). Although these studies contributed to the diagnosis of students' unfavorable attitudes towards science, the instruments used have severe limitations when faced with modern psychometric standards, with a great absence of validity and reliability evidence, or with practices in the development of the instrument not consistent with current standards (Toma & Meneses Villagrá, 2019a). For example, studies whose sample includes students from elementary school have mostly used the ROSE (e.g. Marbá-Tallada & Márquez Bargalló, 2010; Vázquez-

✉ rbtoma@ubu.es **(*Correspondence)**

Contribution of this paper to the literature

- Further evaluation of the psychometric properties of a promising expectancy of success instrument.
- The SUCCESS instrument has strong content, construct and criterion validity, and internal consistency reliability.
- This study paves the way for future research related to the measurement of expectancy of success in school science.

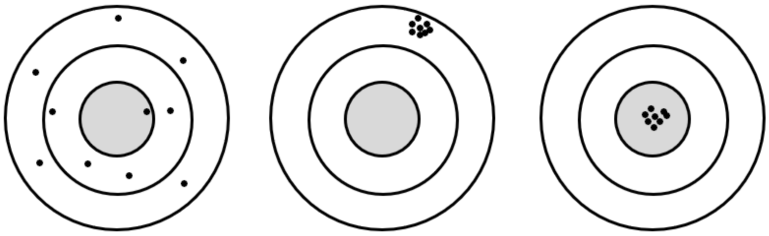


Figure 1. An illustration of validity and reliability, based on Linn and Gronlund (1995). (i) Neither valid nor reliable; (ii) Reliable but not valid; (iii) Both valid and reliable

Alonso & Manassero–Mas, 2011; Vázquez-Alonso & Manassero–Mas, 2008) or the PANA questionnaire (e.g. de Pro & Pérez Manzano, 2014; Pérez Manzano & de Pro Bueno, 2018) without providing information on their psychometric properties (Toma & Meneses Villagrá, 2019a). When validity evidence was provided, there were critical methodological misapplications, like retaining items with cross loadings between factors (i.e. Vázquez-Alonso & Manassero-Mas, 2004), that undermined the quality of the results reported.

In response to this lack of valid instruments, and because persistence has been linked with self-efficacy and expectancies of success (Ceci & Williams, 2007; Chachashvili-Bolotin, Milner-Bolotin, & Lissitsa, 2016; Guo, Parker, Marsh, & Morin, 2015; Sellami, El-Kassem, Al-Qassass, & Al-Rakeb, 2017), Toma and Meneses Villagrá (*in press*) have developed a short instrument to measure elementary school Spanish students' expectancies of success in school science (SUCCESS). Given Blalock et al. (2008) concerns about the lack of validity and reliability of measurement instruments used in science education research, who noted that among weak psychometric properties and methodological issues in the development and validation of instruments there was a proliferation of instruments that where only used in a single study with no follow-up psychometric re-analysis, in this study further psychometric analysis of the SUCCESS instrument is provided. More specifically, this study addressed one of the concerns raised by Blalock et al. (2008) about the need to subject measurement instruments to several psychometric evaluations before they can be presented as valid and reliable.

Therefore, in order to establish a cohesive and robust instrument, we follow Blalock et al. (2008) call related to the need for replication studies to generalize findings and refine instruments by submitting them to both exploratory (EFA) and confirmatory factor analysis (CFA). This practice is consistent with the Standards for Educational and Psychological Testing (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education [AERA, APA, & NCME], 2014) which states that the validation of an instrument is an ongoing process that does not end with the original validation study. Consequently, while Toma and Meneses Villagrá (*in press*) evaluated the psychometric properties of the SUCCESS instrument through different reliability and validity tests, the instrument would benefit of further analysis to establish its validity and reliability beyond the original sample included in the validation study.

BRIEF REVIEW OF PSYCHOMETRIC THEORY

Psychometric theory is concerned with the properties of measurements instruments used in social sciences and involves the development and refinement of approaches to measure latent traits (Nunnally & Bernstein, 1994). The two aspects most relevant to psychometric assessment are validity and reliability (**Figure 1**). Validity relates to whether an instrument measures the trait it should be measuring, and reliability focused on the extent to which an instrument is free of measurement error (Mokkink et al., 2010).

The validity of an instrument is determined upon three types of validity indices (Mokkink et al., 2010a; Mokkink et al., 2010a, 2010b). Firstly, *content and face validity* examine if the construct is adequately represented by the included items in the instrument, and whether these items are correctly interpreted and understood by the population under study. Secondly, *construct validity* determines if the instrument is consistent with theoretical expectations and with initial hypothesis. In other words, an instrument would demonstrate construct validity if it

Table 1. The SUCCESS instrument

Spanish	English
1. Las clases de Ciencias de la Naturaleza me resultan muy difíciles para mí (r)	1. School Science is very hard for me (r)
2. No soy un estudiante al que se le dé bien Ciencias de la Naturaleza (r)	2. I am not a student who does well in school science (r)
3. Soy muy bueno en Ciencias de la Naturaleza	3. I am very good at school science
4. Soy capaz de obtener buenas notas en Ciencias de la Naturaleza	4. I can get good grades in school science
5. Aprendo cosas muy rápidamente en Ciencias de la Naturaleza	5. I learn things quickly in school science
6. Puedo hacer bien las tareas de Ciencias de la Naturaleza	6. I can do my science schoolwork well

(r) refers to items that must be reverse scored

captures the dimensionality of the trait being studied, if it does not measure aspects that should not be measuring, and if it can discriminate between groups that have been hypothesized to be different regarding the focal trait (Polit, 2015). Lastly, *criterion validity* measures if the proposed new instrument is consistent with existing instruments that can be considered "gold standards" in the measurement of that particular trait under study (Polit & Yang, 2016). However, in the absence of a gold standard, it should be examined to what extent the new instrument is correlated to external measures of constructs with which conceptual convergence is expected (Polit, 2015).

The reliability domain measures mainly two aspects related to the internal consistency and the measurement error of change scores. Internal consistency examines the degree to which "(…) the items on a scale are measuring the same underlying construct" (Polit, 2015, p. 1747). The measurement error measures the variation in a set of scores to examine the extent to which "(…) scores are stable and reproducible" (Polit, 2015, p. 1748), with the assumption that an reliable instrument should be able to discriminate between "(…) random temporal fluctuations (unreliability) and true change on an attribute" (Polit, 2015, p. 1749). Likewise, a reliable instrument would correctly measure a trait that has not changed over time.

THE SUCCESS INSTRUMENT

More than two decades ago, Bandura (1977, 1997) proposed in his social cognitive theory that self-efficacy is one of the key motivational aspect that affects individual's choice, achievement, and persistence (Bandura, Barbaranelli, Caprara, & Pastorelli, 2001). Bandura defined self-efficacy as individuals' beliefs in their own ability to effectively perform a given task or problem. He distinguished between "outcomes expectations", which are individuals' beliefs about whether a specific behavior or task will lead to a certain result or outcome (e.g. studying daily will improve science achievement), and "efficacy expectations", which relates to individual's beliefs about whether he/she will be able to perform that specific behavior or task (e.g. being able to study long enough to improve science achievement).

Closely related to Bandura's (1997) theory is Eccles and colleague's expectancy-value model of achievement motivation (EVT, Eccles et al., 1983; Eccles, Wigfield, Harold & Blumenfeld, 1993; Eccles & Wigfield, 1995; Wigfield & Cambria, 2010; Wigfield & Eccles, 1992, 2000, 2002). The EVT model posits that choice, motivation and persistence in a given task are directly influenced by individuals' expectancies of success and the values related to that specific task. The first key construct of the theory is the "expectancies of success", defined by Eccles et al. (1983) as an individual's beliefs about the success he will have in a given task or activity, whether in the immediate future or in the long term. Although conceptually the expectancies of success refers to individual's internal beliefs related to both the ability to perform a task in the present and in the future, this differentiation has not been yet empirically distinguished (Eccles et al., 1993; Wigfield & Eccles, 2000).

The SUCCESS instrument is rooted in the Expectancy-Value (EVT) model of achievement motivation postulated by Eccles and colleagues (Eccles et al., 1983, 1993). More specifically, this instrument is composed by a total of six Likert type items that focuses on the first key construct of the EVT theory, called "expectancies of success". Consistent with the EVT model, it is postulated that individuals will tend to become involved and persist in those tasks or activities in which they feel highly competent. Therefore, high levels of expectancies of success among young students may lead to the development of positive beliefs about their ability to succeed in future science courses or subjects, which may promote the selection of additional science subjects in secondary education or extracurricular activities in science, and eventually the enrolment in a science-related career (Andersen & Ward, 2014). **Table 1** displays SUCCESS items in both English and Spanish language.

Previous Psychometric Evaluation of the SUCCESS Instrument

The SUCCESS instrument has been validated by addressing concerns that have been highlighted as compromising the validity and reliability of many existing instruments in science education (Aydeniz & Kotowski, 2014; Blalock et al., 2008). Thus, the items included in this instrument have been selected from an initial pool that was examined by a panel of experts composed of science educators and psychology university professors and elementary education teachers, and have been tested through think-aloud interviews (Beatty & Willis, 2007) with elementary education students, which provides strong evidence of content and face validity.

Subsequently, the instrument was administered to 418 Spanish elementary school students and the construct validity and reliability of the instrument were examined. The results of exploratory factor analyses reported an unidimensional structure consistent with the theoretical expectations postulated by Eccles et al. (1983). Factorial loadings where higher than .40 and with satisfactory reliability indices (Cronbach α = .761), given the young nature of the sample studied (Newman & McNeil, 1998). To date, this is the first instrument in Spanish that has been validated following modern psychometric recommendation for the specific measurement of elementary school Spanish students' expectancies of success in school science.

Therefore, while the original validation study provided evidence for face, content and construct validity, as well as for the reliability of the SUCCESS instrument, in this study, the psychometric properties of the SUCCESS instrument are further examined by submitting it to construct validity through CFA, criterion validity using external measures, and various reliability indices.

METHOD

Sample

The sample was drawn from 15 elementary schools situated in the province of Burgos, located in the north of Spain. The initial sample was composed of 313 students, however, since goodness of fit statistics are known to be sensitive to abnormal data, Mahalanobis distance was used to detect multivariate outliers. After excluding 32 questionnaires identified as multivariate outliers, a final sample size of 281 elementary school students enrolled in 4^{th} (n = 66), 5^{th} (n = 96) and 6^{th} grades (n = 119) was established. Half of the sample size were girls (50.2%), and the mean age of the participants was 10.35 years old (*SD* = 1, range 8-12 years).

The sample size reached approximately a 47:1 distribution (47 valid cases for each item included in the instrument). Therefore, the sample size is well above the minimum criterion set at 5:1, 10:1 and 20:1 in the literature for factor analysis (de Winter & Dodou, 2012).

Measures

The Spanish-School Science Attitude Survey (S-SSAS) adapted by Toma and Meneses Villagrá (2019a) from the Kennedy, Quinn, & Taylor (2016) instrument was used as an external measure. The S-SSAS is a short instrument composed of 7 Likert-type items and 3 semantic differential items that measures the six most common attitudes toward school science constructs explored in the literature. Four constructs (intention to enrol in future school science, enjoyableness of school science, perceived difficulty of school science, and perception of self-efficacy in school science) are examined through single-item measures, and the remaining two constructs (usefulness of school science, and relevance of school science) are measured through two and four items, respectively. **Table 2** displays S-SSAS instrument items in both English and Spanish language.

In the original validation study (Toma & Meneses Villagrá, 2019a) the S-SSAS reported a Cronbach α of .704, with item-total correlations ranging from .243 to .560. In subsequent studies (Toma & Meneses Villagrá, 2019b), the S-SSAS instrument reported a Cronbach α of .763, with item-total correlation ranging from .313 to .552. For the present study, the Cronbach α was .758 and item-total correlation ranged from .341 - .596. Based on these results, it can be concluded that the S-SSAS is a reliable instrument that can be used as an external measure for computing the criterion validity of the SUCCESS instrument. It is worth noting that both S-SSAS and SUCCESS instruments are concerned with the school science domain and not with science in general, which is consistent with the literature suggesting that school science may be a better predictor of students' interest and future enrolment in science studies than science in general (Osborne et al., 2003).

Table 2. The Spanish School Science Attitude Scale (S-SSAS)

(I) Intention to enroll in further science	
1. I am very likely to enroll on a science course in Year 11	1. Es muy probable que me apunte a Ciencias de la Naturaleza en la ESO
(E) Enjoyableness of school science	
2. I think science is (boring – fun)	2. Pienso que Ciencias de la Naturaleza es (aburrida – divertida)
(D) Perceived difficulty of school science	
3. I struggle with completing the assignments for science class	3. Me cuesta terminar las tareas para la clase de Ciencias de la Naturaleza
(S) Perception of self-efficacy in school science	
4. I think I am very good at science	4. Pienso que soy muy bueno en Ciencias de la Naturaleza
(U) Usefulness of science to careers	
5. A job as a scientist would be interesting	5. Un trabajo como científico sería interesante
6. For my planned career, knowledge of school science will be (SD worthless – required)	6. Para mis futuros estudios, el conocimiento de las clases de Ciencias de la Naturaleza es (inútil – útil)
(R) Relevance of school science	
7. Science helps to make life better	7. La ciencia ayuda a mejorar la vida
8. I want to learn about plants in my area	8. Quiero aprender sobre las plantas de mi entorno
9. For my everyday life, I think school science is (irrelevant – relevant)	9. Para mi vida diaria, creo que Ciencias de la Naturaleza es (poco importante – muy importante)
10. I want to learn about electricity and how it is used in the home	10. Quiero aprender sobre la electricidad y saber cómo se usa en una casa

Data Analysis

Construct validity

Construct validity was examined using CFA, which is a structural equation modelling technique (Brown, 2006). Exploratory factor analysis (EFA) aims to find a model that fits the data, so different alternative models are specified until establishing a final model that fits the data and that has theoretical support (Schumacker & Lomax, 2010). However, the CFA approaches seek to statistically verify the significance of the factorial model that has been previously established through EFA, thus providing more evidence of construct validity (Schumacker & Lomax, 2010). Therefore, CFA is appropriate when the structure of the latent construct is known based on theory, empirical research, or both (Byrne, 2010), just as in the case of the SUCCESS instrument.

For the CFA, the maximum likelihood estimation method (ML) was used through the AMOS v.23 software (Arbuckle, 2014). Following consensual recommendations in the literature (Brown, 2006; Hair, Black, Babin, Anderson, & Tatham, 2010; Kline, 2005), the following six fit indices were examined to establish model fit:

- Chi-square (χ^2). The χ^2 measures the extent to which the variance and covariance matrices differ between the observed and the implicit model. A non-significant p value is desired, however, χ^2 is sensitive to sample size, so a sample greater than 200 cases drastically increases the probability of obtaining a significant p value.
- Comparative Fit Index (CFI). The CFI indicates the degree of fit of the proposed model compared to a null model, based on the sample size. CFI values of ≥ .90 indicate an appropriate fit.
- Goodness of Fit Index (GFI). The GFI measures the relative amount of variance and covariance that is explained by the proposed model. GFI values of ≥ .90 indicate an appropriate fit.
- Adjusted Goodness of Fit Index (AGFI). The AGFI is similar to the GFI, however it measures the relative amount of variance and covariance adjusted to the degrees of freedom of the proposed model. AGFI values of ≥ .90 indicate an appropriate fit.
- Root Mean Square Residual Index (RMR). The RMR index represents the average value of all standardised residuals. RMR values of ≤ .05 indicate an appropriate fit.
- Root Mean Square Error of Approximation (RMSEA). The RMSEA examines the degree of error in the proposed model considering the sampling approximation error. RMSEA values of ≤ .08 indicate an appropriate fit.

Criterion validity

The Pearson correlation coefficient was used to examine the correlation between the SUCCESS instrument and six other external measures with which conceptual convergence is expected, using the statistical package SPSS v. 24 (IBM, 2016).

Table 3. Model fit indices

Fit indices	Recommended values	1 factor model	Evaluation
χ^2	$p > .05$	$\chi^2 = 15.274$, $p = .054$	Adequate
CFI	≥ .90	.980	Optimal
GFI	≥ .90	.984	Optimal
AGFI	≥ .90	.958	Optimal
RMR	≤ .05	.034	Optimal
RMSEA	≤ .08	.054	Optimal

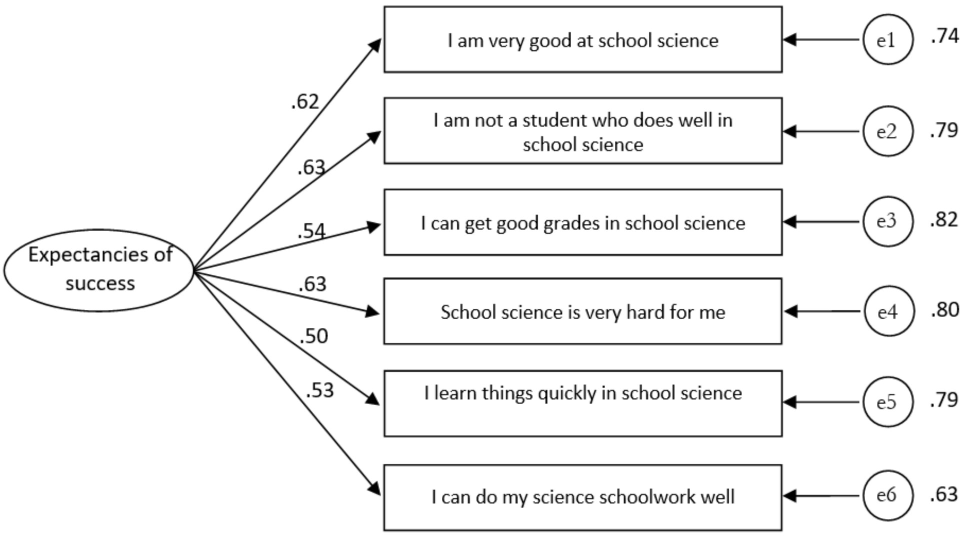


Figure 2. Confirmatory factor analysis results

Reliability

The reliability of the SUCCESS instrument was examined through the Cronbach alpha coefficient with a minimum criterion of $\alpha > .70$ (Nunnally & Bernstein, 1994), and through item-total correlation with a minimum criterion of $r > .30$ using the statistical package SPSS v. 24 (IBM, 2016).

FINDINGS

Construct Validity

CFA results revealed that the hypothesized single-factor model obtained an optimal fit in five of the six model fit indices. The index obtaining only adequate fit was χ^2 with a value of $p = .054$, although this is a parameter that is affected by the sample size, as indicated previously.

The degree of fit of the proposed model was adequate, as shown by the CFI value greater than .90. The relative amount of variance explained by the model was also adequate, obtaining values greater than .90 in both GFI and AFGI indices. The fit of the model was also appropriate when considering the overall amount of error, with the RMR reaching a value lower than .05 and the RMSEA a value lower than .08. These results are reported in **Table 3**.

All items loadings were statistically significant. As can be seen in **Figure 2**, the lowest value was obtained for the item "I learn things quickly in school science", with a value of $\lambda = .50$. The strongest loadings were obtained for both reversed items "I am not a student who does well in school science" and "School science is very hard for me", with a value of $\lambda = .63$. These results confirm the initial hypothesis that the SUCCESS instrument consists of a unidimensional scale that measures the expectancies of success in school science.

Criterion Validity

Pearson correlation coefficient revealed that the SUCCESS instrument was significantly correlated to six external attitudes toward school science measures. The correlation ranged from $r = .193$ to $r = .740$. More specifically, a significant, small and positive correlation was found between students' expectancies of success and their

Table 4. Pearson correlation coefficient between the SUCCESS instrument and external measures

External measures	SUCCESS instrument
Intentions to enrol	.193[1]
Enjoyableness	.353
Difficulty	- .485
Auto-efficacy	.740
Perceived utility	.201
Perceived relevance	.322

[1]All correlation were significant at $p < .01$ level

Table 5. Reliability results

SUCCESS items	Item-total *r*	Cronbach α after removing items
I am very good at school science	.602	.721
I am not a student who does well in school science (r)	.598	.720
I can get good grades in school science	.458	.757
School science is very hard for me (r)	.493	.748
I learn things quickly in school science	.446	.759
I can do my science schoolwork well	.532	.739

(r) refers to items that must be reverse scored

intentions to enroll in further science courses. A statistically significant, large and positive correlation between students' expectancies of success and self-efficacy and a significant, moderate and negative correlation between students' expectancies of success and their perceived difficulty of school science was identified. Finally, a significant, moderate and positive correlation was found between students' expectancies of success and their perceived relevance of school science, and perceived usefulness of school science. These results are reported in **Table 4**.

Reliability

Cronbach's α was .775 for the entire instrument, with ranges between .720 and .759 if items were deleted. Since the Cronbach α does not improve considerably after deleting items, it means that all items included in the instrument are relevant and measure the same underlying construct. These results, which are above the minimum .70 criteria required for exploratory studies (Nunnally & Bernstein, 1994), coincide with the original validation study.

In addition, the items included in the SUCCESS instrument are highly correlated with each other, with item-total correlation ranging from .446 to .602. These results also satisfy the minimum criterion of $r > .30$, and likewise coincide with those obtained in the original validation study. Taken together, these results provide additional evidence of the internal consistency of the SUCCESS instrument, confirming that it is a instrument with satisfactory levels of reliability. **Table 5** shows the internal consistency results of the SUCCESS instrument.

DISCUSSION

This study provides further evidence on the psychometric properties of the SUCCESS instrument. In relation to reliability, the internal consistency results are in line with the results of the original validation study and are above minimum recommended values for exploratory studies (Nunnally & Bernstein, 1994). Regarding validity evidences, the structural validity of the SUCCESS instrument was confirmed with fit model indices results above the recommended values in the CFA literature (Brown, 2006; Hair et al., 2010; Kline, 2005). Likewise, evidence of criterion validity has been provided using external measures that have already been validated in Spanish and postulated as robust instruments. In this regard, the extent to which the SUCCESS instrument correlates with a total of six external measures with which conceptual convergence was expected was assessed.

Firstly, given that expectancies of success greatly influence individuals' decisions regarding career choice and persistence (Lent & Brown, 1994; Lent, Brown, & Hackett, 2002), it was hypothesized that the SUCCESS instrument would be positively related to students' intentions to enroll in future school science courses, assumptions that have been confirmed in this study.

Secondly, since expectancies of success are defined as individuals' beliefs about the success of performing a given task or activity (Eccles et al. 1983), it was hypothesized that the SUCCESS instrument will be negatively related to students perceived difficulty of school science and positively related to students' self-efficacy in studying school science. The results of this study are also in line with these theoretical assumptions.

Finally, since participation in tasks or activities that require knowledge and skills application in demanding situations enhances students' self-efficacy (Van Dinther, Dochy, & Segers, 2011), and since enjoyment, interest and satisfaction have been identified as factors influencing student self-efficacy (Hutchison, Follman, Sumpter, & Bodner, 2006), it was hypothesized that the SUCCESS instrument will be positively related to students enjoyment, perceived usefulness, and perceived relevance of school science. These initial theoretical expectations have also been confirmed in this study.

Taken together, these results confirm that the SUCCESS is a valid and reliable instrument with great construct and criterion validity, as well as with structural and content validity, as assessed and reported in the original validation study (Toma & Menéses Villagrá, *in press*).

IMPLICATIONS AND AVENUES FOR FUTURE RESEARCH

Several implications are derived from this study. To the best of the author's knowledge, the original study (Toma & Menéses Villagrá, *in press*) and this follow-up study represent the first efforts to develop and validate an instrument that specifically tackle Spanish-speaking elementary school students' expectancies of success in school science. Since both studies show that the SUCCESS instrument is valid and reliable according to modern psychometric standards, it can be used by science teachers and researchers to examine whether different activities and teaching methodologies are successful at improving students' expectancies of success in school science. Therefore, the SUCCESS instrument could be used in studies aimed at the development of learning strategies for the improvement of students' expectancies of success in school science, which could tackle students decline in interest toward school science from early stages where scientific vocations can be more easily fostered (Lindahl, 2007).

Another implication derived is related to the need to develop large-scale studies to identify at what stage of elementary education students experience a decline in their expectancies of success. Also, given the short administration time of the proposed instrument, it could be used in conjunction with other relevant measures to identify those variables that may influence students' expectancies of success.

Additionally, although expectancies of success have been linked to persistence (Chachashvili-Bolotin et al., 2016; Guo et al., 2015), it is not clear to what extend high levels of expectancies of success may be related to an actual increase in students desire and interest in enrolling in science courses at later stages of the educational system, aspect that requires future empirical research. In addition, given that gender has been identified as a critical variable in the development of interest and attitudes toward science (Osborne et al., 2003; Sellami et al., 2017; Toma & Greca, 2018), it would be interesting to test in future validation studies the measurement invariance of the SUCCESS instrument and to compare its latent factor mean in girls and boys enrolled in elementary education. This would guarantee the integrity of the results by reducing the possibility of obtaining contradictory results derived from different factorial structures of the SUCCESS instrument in girls and boys.

Furthermore, the test-retest reliability of the SUCCESS should be examined in future research, which will allow the use of this instrument in longitudinal studies that requires several data collection phases. Finally, future studies should focus on validating this instrument for middle and high school students, which would allow cross-sectional and longitudinal studies focusing on those stages of the Spanish educational system in which a greater decrease in students interest in science has been identified (MECD, 2016).

CONCLUSIONS

The results of this second psychometric evaluation of the SUCCESS instrument reveal that the proposed instrument is conceptually consistent with the "Expectancies of success" construct first introduced by Eccles et al. (1983), and methodologically robust in terms of modern validity and reliability psychometric evidences.

REFERENCES

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education (AERA, APA, & NCME, 2014). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association.

Andersen, L., & Ward, T. J. (2014). Expectancy-value models for the STEM persistence plans of ninth-grade, high-ability students: A comparison between black, hispanic, and white students. *Science Education*, *98*(2), 216–242. https://doi.org/10.1002/sce.21092

Arbuckle, J. L. (2014). Amos (Version 23.0). Chicago: IBM SPSS.

Archer, L., Dewitt, J., Osborne, J., Dillon, J., Willis, B., & Wong, B. (2010). “Doing” science versus “being” a scientist: Examining 10/11-year-old schoolchildren’s constructions of science through the lens of identity. *Science Education*, *94*(4), 617–639. https://doi.org/10.1002/sce.20399

Aydeniz, M., & Kotowski, M. R. (2014). Conceptual and methodological issues in the measurement of attitudes towards science. *Electronic Journal of Science Education Electronic Journal of Science Education*, *18*(3), 1–24.

Bandura A, Barbaranelli C, Caprara G.V., & Pastorelli C. (2001). Self-efficacy beliefs as shapers of children’s aspirations and career trajectories. *Child Development, 72*, 187–206. https://doi.org/10.1111/1467-8624.00273

Bandura A. (1997). *Self-efficacy: The exercise of control*. New York: Freeman.

Bandura, A. (1977). Self-efficacy: Toward a unifying theory of behavioral change. *Psychological Review*, *84*(2), 191-215. https://doi.org/10.1037/0033-295X.84.2.191

Beatty, P. C., & Willis, G. B. (2007). Research synthesis: The practice of cognitive interviewing. *Public Opinion Quarterly*, *71*(2), 287–311. https://doi.org/10.1093/poq/nfm006

Blalock, C. L., Lichtenstein, M. J., Owen, S., Pruski, L., Marshall, C., & Toepperwein, M. A. (2008). In pursuit of validity: A comprehensive review of science attitude instruments 1935-2005. *International Journal of Science Education*, *30*(7), 961–977. https://doi.org/10.1080/09500690701344578

Brown, T. A. (2006). *Confirmatory factor analysis for applied research*. New York: Guilford Press.

Byrne, B. M. (2010). *Structural equation modeling with AMOS: Basic concepts, applications, and programming* (Vol. 2). New York: Routledge Taylor & Francis Group. https://doi.org/10.4324/9781410600219

Ceci, S. J., & Williams, W. M. (2007). *Why aren’t more women in science. Top researchers debate the evidence*. Washington, DC: American Psychological Association. https://doi.org/10.1037/11546-000

Chachashvili-Bolotin, S., Milner-Bolotin, M., & Lissitsa, S. (2016). Examination of factors predicting secondary students’ interest in tertiary STEM education. *International Journal of Science Education*, *38*(3), 366–390. https://doi.org/10.1080/09500693.2016.1143137

de Pro, A., & Pérez Manzano, A. (2014). Actitudes de los alumnos de Primaria y Secundaria ante la visión dicotómica de la Ciencia. *Enseñanza de Las Ciencias*, *32*(3), 111–132. https://doi.org/10.5565/rev/ensciencias.1015

de Winter, J. C. F., & Dodou, D. (2012). Factor recovery by principal axis factoring and maximum likelihood factor analysis as a function of factor pattern and sample size. *Journal of Applied Statistics*, *39*(4), 695–710. https://doi.org/10.1080/02664763.2011.610445

Eccles, J. S., & Wigfield, A. (1995). In the mind of the achiever: the structure of adolescents’ academic achievement related-beliefs and self-perceptions. *Personality and Social Psychology Bulletin*, *21*, 215–225. https://doi.org/10.1177/0146167295213003

Eccles, J. S., Wigfield, A., Harold, R. B., & Blumenfeld, P. B. (1993). Age and gender differences in children’s self- and task perceptions during elementary school. *Child Development*, *64*, 830–847. https://doi.org/10.2307/1131221

Eccles, J., Adler, T., Futterman, R., Goff, S., Kaczala, C., Meece, J., & Midgley, C. (1983). Expectancies, values, and academic behaviors. In J. Spence (Ed.), *Achievement and achievement motivation* (pp. 75–146). San Francisco, CA: W. H. Freeman.

Gardner, P. L. (1975). Attitudes to science: A review. *Studies in Science Education*, *2*, 1–41. https://doi.org/10.1080/03057267508559818

Guo, J., Parker, P. D., Marsh, H. W., & Morin, A. J. S. (2015). Achievement, motivation, and educational choices: A longitudinal study of expectancy and value using a multiplicative perspective. *Developmental Psychology*, *51*(8), 1163–1176. https://doi.org/10.1037/a0039440

Hair, J. F., Black, W. C., Babin, B. J., Anderson, R. E., & Tatham, R. L. (2010). *Multivariate data anaysis (Vol 7.)*. Upper Saddle River: Pearson Prentice Hall.

Hutchison, M. A., Follman, D. K., Sumpter, M., & Bodner, G. M. (2006). Factors influencing the self-efficacy beliefs of first-year engineering students. *Journal of Engineering Education*, *95*(1), 39–47. https://doi.org/10.1002/j.2168-9830.2006.tb00876.x

IBM, C. (2016). IBM SPSS Statistics for Windows. Version 24.0. Armonk, NY: IBM Corp.

Kennedy, J. P., Lyons, T., & Quinn, F. (2014). The continuing decline of science and mathematics enrolments in Australian high schools. *Teaching Science*, *60*(2), 34–46.

Kennedy, J., Quinn, F., & Taylor, N. (2016). The school science attitude survey: A new instrument for measuring attitudes towards school science. *International Journal of Research & Method in Education*, *39*(4), 422–445. https://doi.org/10.1080/1743727X.2016.1160046

Kline, R. B. (2005). *Principles and practice of structural equation modeling*. New York: The Guilford Press.

Lent, R. W., & Brown, S. D. (1994). Toward a unifying social cognitive theory of career and academic interest, choice, and performance. *Journal of Vocational Behavior*, *45*, 79–122. https://doi.org/10.1006/jvbe.1994.1027

Lent, R. W., Brown, S. D., & Hackett, G. (2002). Social cognitive career theory. In D. Brown, L. Brooks, & Associates (Eds.), *Career choice and development* (4th ed., pp. 255–311). San Francisco, CA: Jossey-Bass.

Lindahl, B. (2007). Longitudinal study of students' attitudes towards science and choice of career. Paper presented at the *80th session of the International Conference of the National Association for Research in Science Teaching*. New Orleans.

Linn, R. L., & Gronlund, N. E. (1995). *Measurement and assessment in teaching*. New Jersey: Prentice-Hall Inc.

Lyons, T., & Quinn, F. (2010). *Choosing science. Understanding the declines in senior high school science enrolments*. Armidale: University of New England.

Lyons, T., & Quinn, F. (2015). Understanding declining science participation in Australia: A systemic perspective. In E. K. Henriksen, J. Dillon, & J. Ryder (Eds.), *Understanding student participation and choice in science and technology education* (pp. 153–168). Springer Netherlands. https://doi.org/10.1007/978-94-007-7793-4_10

Marbá-Tallada, A., & Márquez Bargalló, C. (2010). ¿Qué opinan los estudiantes de las clases de ciencias ? Un estudio tranversal de sexto de Primaria a cuarto de ESO. *Enseñanza de Las Ciencias*, *28*(1), 19–30.

MECD. (2016). *Datos y cifras del sistema universitario español. Curso 2015/2016*.

Mokkink, L. B., Terwee, C. B., Knol, D. L., Stratford, P. W., Alonso, J., Patrick, D. L., … de Vet, H. C. (2010). The COSMIN checklist for evaluating the methodological quality of studies on measurement properties: A clarification of its content. *BMC Medical Research Methodology*, *10*(22), 1–8. https://doi.org/10.1186/1471-2288-10-22

Mokkink, L. B., Terwee, C. B., Patrick, D. L., Alonso, J., Stratford, P. W., Knol, D. L., … de Vet, H. C. W. (2010a). The COSMIN study reached international consensus on taxonomy, terminology, and definitions of measurement properties for health-related patient-reported outcomes. *Journal of Clinical Epidemiology*, *63*(7), 737–745. https://doi.org/10.1016/j.jclinepi.2010.02.006

Mokkink, L. B., Terwee, C. B., Patrick, D. L., Alonso, J., Stratford, P. W., Knol, D. L., … de Vet, H. C. W. (2010b). The COSMIN checklist for assessing the methodological quality of studies on measurement properties of health status measurement instruments: An international Delphi study. *Quality of Life Research*, *19*(4), 539–549. https://doi.org/10.1007/s11136-010-9606-8

Munby, H. (1983). Thirty studies involving 'Scientific Attitude Inventory': What confidence can we have in this instrument? *Journal of Research in Science Teaching*, *20*, 141–162. https://doi.org/10.1002/tea.3660200206

Munby, H. (1997). Issues of validity in science attitude measurement. *Journal of Research in Science Teaching*, *34*(4), 337–341. https://doi.org/10.1002/(SICI)1098-2736(199704)34:4<337::AID-TEA4>3.0.CO;2-S

Newman, I., & McNeil, K. (1998). *Conducting survey research in the social sciences*. New York: University Press of America.

Nunnally, J. C., & Bernstein, I. H. (1994). *Psychometric theory*. New York: McGraw-Hill.

Osborne, J., Simon, S., & Collins, S. (2003). Attitudes towards science: A review of the literature and its implications. *International Journal of Science Education*, *25*(9), 1049–1079. https://doi.org/10.1080/0950069032000032199

Pérez Manzano, A., & de Pro Bueno, A. (2018). Algunos datos sobre la visión de los niños y de las niñas sobre las ciencias y del trabajo científico. *IQual. Revista de Género e Igualdad*, *0*(1), 18–31. https://doi.org/10.6018/iQual.306091

Polit, D. F. (2015). Assessing measurement in health: Beyond reliability and validity. *International Journal of Nursing Studies*, *52*(11), 1746–1753. https://doi.org/10.1016/j.ijnurstu.2015.07.002

Polit, D. F., & Yang, F. (2016). *Meaurement and the measurement of change: A primer for health professionals*. Philadelphia: Lippincott Williams & Wilkins.

Potvin, P., & Hasni, A. (2014). Interest, motivation and attitude towards science and technology at K-12 levels: A systematic review of 12 years of educational research. *Studies in Science Education*, *50*(1), 85–129. https://doi.org/10.1080/03057267.2014.881626

Schumacker, R. E., & Lomax, R. G. (2010). *A beginner's guide to structural equation modeling. Routledge* (Vol. 47). New York: Routledge Taylor & Francis Group.

Sellami, A., El-Kassem, R. C., Al-Qassass, H. B., & Al-Rakeb, N. A. (2017). A path analysis of student interest in STEM, with specific reference to Qatari students. *Eurasia Journal of Mathematics, Science and Technology Education*, *13*(9), 6045–6067. https://doi.org/10.12973/eurasia.2017.00999a

Toma, R. B., & Greca, I. M. (2018). The effect of integrative STEM instruction on elementary students' attitudes toward science. *EURASIA Journal of Mathematics, Science & Technology Education*, *14*(4), 1383–1395. https://doi.org/10.29333/ejmste/83676

Toma, R. B., & Meneses Villagrá, J. A. (2019a). Validation of the single-items spanish-School science attitude Survey (S-SSAS) for elementary education. *PLoS ONE* *14*(1), e0209027. https://doi.org/10.1371/journal.pone.0209027

Toma, R. B., & Meneses Villagrá, J. A. (2019b). Prefrencia por contenidos científicos de física o de biología en Educación Primaria: un análisis cluster, *Revista Eureka sobre Enseñanza y Divulgación de las Ciencias, 1*(1), 1104, https://doi.org/10.25267/Rev_Eureka_ensen_divulg_cienc.2019.v16.i1.1104

Toma, R. B., & Meneses Villagrá, J. A. (in press). Development and validation of the SUCCESS instrument: towards a valid and reliable measure of expectancies of success in School Science.

Van Dinther, M., Dochy, F., & Segers, M. (2011). Factors affecting students' self-efficacy in higher education. *Educational Research Review*, *6*(2), 95–108. https://doi.org/10.1016/j.edurev.2010.10.003

Vázquez-Alonso, Á., & Manassero–Mas, M. A. (2008). El declive de las actitudes hacia la ciencia de los estudiantes: un indicador inquietante para la educación científica. *Revista Eureka sobre Enseñanza y Divulgación de Las Ciencias*, *5*(3), 274–292. https://doi.org/10.25267/Rev_Eureka_ensen_divulg_cienc.2008.v5.i3.03

Vázquez-Alonso, Á., & Manassero-Mas, M. A. (2004). Imagen de la ciencia y la tecnología al final de la educación obligatoria. *Cultura y Educación*, *16*, 385-398. https://doi.org/10.1174/1135640042802473

Vázquez-Alonso, Á., & Manassero–Mas, M. A. (2011). El descenso de las actitudes hacia la ciencia de chicos y chicas en la Educación Obligatoria. *Ciência & Educação*, *17*(2), 249–268. https://doi.org/10.1590/S1516-73132011000200001

Wigfield, A., & Cambria, J. (2010). Students' achievement values, goal orientations, and interest: definitions, development, and relations to achievement outcomes. *Developmental Review*, *30*(1), 1–35. https://doi.org/10.1016/j.dr.2009.12.001

Wigfield, A., & Eccles, J. S. (1992). The development of achievement task values: a theoretical analysis. *Developmental Review*, *12*, 265–310. https://doi.org/10.1016/0273-2297(92)90011-P

Wigfield, A., & Eccles, J. S. (2000). Expectancy-value theory of motivation. *Contemporary Educational Psychology*, *25*(1), 68–81. https://doi.org/10.1006/ceps.1999.1015

Wigfield, A., & Eccles, J. S. (2002). The development of competence beliefs, expectacies for success, and achievement values from childhood through adolescence. In A. Wigfield y J. S. Eccles (Eds.), *Development of achievement motivation* (pp. 91–120). San Diego: Academic Press. https://doi.org/10.1016/B978-012750053-9/50006-1

http://www.ejmste.com

Appendix 7: S-SSAS instrument validation study

This appendix includes the full-text of the published article about the validation study of the S-SSAS instrument used in this dissertation to assess the effectiveness of the proposed pedagogical framework for STEM-based science education in improving elementary school students' attitudes towards school science.

Toma, R. B., & Meneses Villagrá, J. A. (2019). Validation of the single-items Spanish-School Science Attitude Survey (S-SSAS) for elementary education. *PLoS ONE, 14*(1), e0209027. doi: 10.1371/journal.pone.0209027

PLOS | ONE

RESEARCH ARTICLE

Validation of the single-items Spanish-School Science Attitude Survey (S-SSAS) for elementary education

Radu Bogdan Toma*, Jesús Ángel Meneses Villagrá

Departamento de Didácticas Específicas, Universidad de Burgos, Burgos, Castilla y León, Spain

* rbtoma@ubu.es

OPEN ACCESS

Citation: Toma RB, Meneses Villagrá JÁ (2019) Validation of the single-items Spanish-School Science Attitude Survey (S-SSAS) for elementary education. PLoS ONE 14(1): e0209027. https://doi.org/10.1371/journal.pone.0209027

Editor: Juan Carlos Perez Gonzalez, Universidad Nacional de Educacion a Distancia (UNED), SPAIN

Received: July 19, 2018

Accepted: November 27, 2018

Published: January 2, 2019

Data Availability Statement: All relevant data are within the paper and its supporting information files.

Funding: This study has received funding from the 2017-2021 edition of the University of Burgos pre-doctoral research grant to RBT and by the Spanish Ministry of Economy, Industry and Competitiveness (MINECO) through the project EDU2017-89405-R to JAMV. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Abstract

The development of positive attitudes toward science is one of the main priorities in science education. However, there is a lack of reliable and valid instruments to measure Spanish-speaking elementary students' attitudes towards school science. In this study, the translation and validation of the Spanish School-Science Attitude Survey (S-SSAS) is reported. The instrument was administered to 643 students enrolled in 3rd to 6th elementary grades. Psychometric evaluation of the S-SSAS provided sound evidence for validity (face, content, construct and criterion) and reliability (internal consistency and temporal stability). Content validity was confirmed through a panel of experts who reached great consensus in linking items to attitudinal constructs, with an ICC = .956. Think-aloud interviews confirmed that students have easily understood and correctly interpreted all items included, thus providing face validity for the S-SSAS. Consistent with theoretical expectations, predictive validity ranged between -.334 to 543 and concurrent validity was examined through S-SSAS correlation with two external measures of conceptual convergence that ranged from.301 to .560, thus confirming criterion validity. Construct validity was assessed by obtaining consistent results with the original scale in terms of reporting no statistically significant differences in attitudinal profiles towards school science between girls and boys and between students from urban and rural schools. Cronbach αfor the entire scale was .704, with item-total correlation ranging from .243 to .560, which reports acceptable internal consistency. Temporal stability with a 10-days span was good, with ICC = .873 and r = .464–790. Taken together, these results indicate that the Spanish single-items School-Science Attitude Survey is easy to administer and equally interpreted by both girls and boys enrolled in rural and urban elementary schools, thus being a valid and reliable instrument for measuring attitudes towards school science.

Introduction

In the last decade, there have been significant changes in science education post-compulsory courses enrolments, with a steady decline in students interested in Science, Technology, Engineering and Mathematics (STEM) disciplines [1–3]. Therefore, the promotion of positive

Competing interests: The authors have declared that no competing interests exist.

attitudes towards STEM disciplines is considered nowadays a priority objective in science education [4–7].

To date, attempts to measure attitudes have mainly focused on the administration of Likert-type scales that not always reported adequate psychometric properties and that often were too extensive to be implemented at elementary education stage. For example, the instruments with the strongest psychometric properties from Blalock et al. review [8] were validated with secondary school students [9–11] and posed administration problems due to their length [10]. Nonetheless, there are few newer instruments to measure attitudes toward science specifically developed and validated for elementary education: the 30-item BRAINS instruments [12], the 28-item Three-Dimension Elementary Science Attitude Survey [13], the 28-items attitude toward STEM instrument [14] or the 30-item Attitudes toward Science Class instrument [15]. However, some of these scales are focused on the measurement of attitudes towards science in general and are still to extensive, especially for longitudinal studies where educational interventions intended to improve attitudes are the main focus and therefore multiple data collection are needed [16].

This problem is further accentuated for scales in Spanish, with a great absence in the literature of validated scales for elementary stages. For example, some studies [17,18] have used extensive questionnaires in Spanish in both elementary and secondary schools, however, authors did not report on their psychometric properties. On the other hand, although Navarro, Förster, González, & González-Pose [19] have validated Fraser's TOSRA scale [11,20] for Spanish-speaking students, the questionnaire is made up of a total of 70 Likert-type items, is focused on measuring attitudes towards science in general, and has also been validated only with students in higher grades than the Primary Education stage. Therefore, there is a need for valid and reliable quantitative instruments of easy and quick administration that facilitates the study of students' attitudes toward school science at this stage. Consequently, the aim of this paper is to study the translation and validation of the Spanish single item instrument proposed by Kennedy, Quinn, & Taylor [21] called School Science Attitude Survey (SSAS). Considering the potential usefulness of brief measures of attitude toward school science, this study aims to provide sound evidence for the translation procedure used (face & content validity), for construct and criterion validity, and for internal consistency, sensitivity and temporal stability reliability.

Theoretical underpinning of attitudes toward science

Although studies about students' attitudes toward science have been mounting in the last decades, what is meant by attitudes toward science is still "(. . .) somewhat nebulous, often poorly articulated and not well understood" [6] p.1049, and no clear definition have been provided yet. Klopfer [22] offered insight about the concept under study by categorizing the attitude construct as a set of affective behaviors toward science as an enterprise, scientist, scientific inquiry, scientific careers and towards science-related activities in general. Gardner [23] provided further clarity by addressing the differences between «scientific attitudes», conceived as those elements inherent to scientific thinking and research, and «attitudes toward science», conceptualized as the sociological, psychological and affective conceptions and beliefs about science. Research developed at elementary level reported the existence of many sub-constructs underlying the attitudes toward science construct, including (i) students affective feelings and cognitive judgments of science [13], (ii) unfavorable outlook of science [24] and (iii) perception of scientists and value of science to society [25], among many others.

Recent results of past research indicating that, in general, students tend to have positive attitudes toward science but negative attitudes toward school science, stressed the need of advancing research about factors affecting students' intentions and likehood of enrolling in science related activities and careers. Therefore, there is a shift from focusing on attitudes toward science in general, to focusing on attitudes toward school science, which may be a better predictor of student's behavior than attitudes toward science in general [6]. Consequently, recent studies adopted various psychological theories like the Theory of Planned Behavior (TPB) [26] or the Expectancy-value theory (EVT) [27–31] to further explore student's motivation, choice and persistence in studying science-related subjects and careers. Thus, attitudes are being studied by examining students enjoyment, self-efficacy, perceived difficulty [21] and others behavioral beliefs about the consequences of engaging in science [8,12,24,32].

Literature review

Existing attitude instruments for Spanish-speaking elementary students

Research on attitudes towards science has been dealing with methodological issues caused by a lack of rigor in the development and use of attitude measurement instruments. Munby [33] first concluded that authors do not take the necessary steps to develop valid and reliable attitude toward science instruments. Thus, for example, the presence of discrepant and contradictory results is very common in attitude toward science research and may be explained by the lack of psychometrically robust instruments [34]. In more recent years, Blalock et al. [8] underlined these issues by concluding that the vast majority of attitude instruments used in science education research the area of scientific education are lacking in terms of validity and reliability psychometric properties evidences.

Attitude toward science research conducted in Spain is no stranger to this problem. The most relevant attitude studies in Spain have used questionnaires that have not been subjected to reliability and validity tests and therefore, the extent to which the results reported in these studies are valid and reliable remains unclear (Table 1).

These studies have mainly used the ROSE instrument and the PANA questionnaire. In relation to the ROSE instrument, it is an adaptation of the Schreiner & Sjøberg [45]scale, originally developed in English and subsequently translated into several languages, including Spanish. The PANA questionnaire has been specifically designed by the authors. In both cases, the lack of validity, reliability and resistivity evidences is worryingly absent. None of the studies using the PANA questionnaire have provided information on their psychometric properties, and only one study of the six included in the Table 1 using the ROSE questionnaire provided information on the validity and reliability of the instrument, albeit with clear methodological problems. For example, Vázquez-Alonso & Manassero-Mas [41] have submitted the ROSE questionnaire to Exploratory Factor Analysis (EFA), obtaining cross loadings on three items. Instead of eliminating these items and redoing the factor analysis, the authors decided to keep the items, which is clearly a methodologically inadequate practice [46,47]. In short, these results demonstrate the need to develop psychometrically valid instruments, especially for the Spanish context.

The school science attitude survey (SSAS)

The SSAS [21]is a web-based visual-analogue scale designed to examine student's attitudinal profile (AP) to the area of school science through ten items that addresses the six common attitudinal constructs (AC) used in the literature of attitudes toward science (Table 2).From an initial pool of 46 items based on existing instruments and 22 newly developed, authors selected those that best represented each construct based on interviews with target sample, internal constancy results, and four dimensionality tests.

Table 1. Psychometric quality of the instruments used in Spanish attitude toward science studies.

Instrument	*Authors*	*Items and constructs*	*Grades*	*Reliability*	*Validity*	*Sensitivity*
PANA	De Pro Bueno & Pérez Manzano [17]	6 items measuring 6 constructs	6–10	No	No	No
	Pérez Manzano & De Pro Bueno[35]	17 items measuring 6 constructs	6–10	No	No	No
ROSE	Marbá-Tallada& Márquez Bargalló[36]	16 items measuring 3 constructs	6–10	No	No	No
	Pérez-Franco & De Pro Bueno[37]	19 items measuring 3 constructs	10	No	No	No
	Vázquez-Alonso &Manassero-Mas[38]	24 items measuring 3 constructs	4—undergraduates	No	No	No
	Vázquez-Alonso &Manassero-Mas[39]	80 items measuring 4 constructs	10	No	No	No
	Vázquez-Alonso &Manassero-Mas[40]	149 items measuring 5 constructs	10	No	No	No
	Vázquez-Alonso &Manassero-Mas[41]	16 items measuring 4 constructs	10	Yes	Yes[1]	No
	Vázquez-Alonso &Manassero-Mas[42]	24 items measuring 3 constructs	4–12	Yes	No	No
COCTS	Vázquez-Alonso, Acevedo Díaz, Manassero-Mas & Acevedo Romero[43]	202 items measuring 28 constructs	11–12	No	No	No
WAREING and PAC	Vázquez-Alonso &Manassero-Mas[44]	50 items measuring 10 constructs	8 -undergraduates	Yes	Yes[2]	No

[1]The Exploratory Factor Analysis revealed 3 items with cross loadings between factors.

[2] Authors only provided evidence of content validity.

https://doi.org/10.1371/journal.pone.0209027.t001

Four AC–intentions for future enrolment (I), enjoyableness (E), difficulty (D) and self-efficacy in school science(S)–were unidimensional and the remaining two AC–usefulness (U) and relevance of school science (R)–were found to be multidimensional.More specifically,

Table 2. The original SSAS [21].

Attitudinal construct	*Item*
(I) Intention to enroll in further science	1. I am very likely to enroll on a science course in Year 11(LT)
(E) Enjoyableness of school science	2. I think science is (SD boring–fun)
(D) Perceived difficulty of school science	3. I struggle with completing the assignments for science class (LT)
(S) Perception of self-efficacy in school science	4. I think I am very good at science (LT)
(U) Usefulness of science to careers $= \frac{us+up}{2}$	
(Us)Usefulness of school science to scientific careers	5. A job as a scientist would be interesting (LT)
(Up) Usefulness of school science to personal career choice	6. For my planned career, knowledge of school science will be (SD worthless–required)
(R) Relevance of school science $= \frac{rs+rp}{2}$	
(Rs) Relevance of school science to society	7. Science helps to make life better (LT)
(Rp) Personal relevance of School science $= \frac{Rp1+Rp2+Rp3}{3}$	
(Rp1) What do I want to learn about?	8. I want to learn about plants in my area (LT)
(Rp2)How applicable is school science to my everyday life?	9. For my everyday life, I think school science is (SD irrelevant–relevant)
(Rp3) Biological vs physical science	10. I want to learn about electricity and how it is used in the home

https://doi.org/10.1371/journal.pone.0209027.t002

usefulness of science for a future career in science (U_s) and usefulness of science for personal career choice (U_p) are the two dimensions underlying the usefulness (U) attitudinal construct. Finally, relevance to society (R_s) and personal relevance (R_p) of school science are the two dimensions underlying the relevance (R) attitudinal construct. Both the unidimensional and the dimensions underlying the multidimensional AC are measured through single-item measures with five response options, either Likert-type (totally disagree–totally agree) or semantic differential scales (i.e. fun–boring) with only extreme options being labeled.

The SSAS was selected for its translation and validation for Spanish speaking students for several reasons. Firstly, because it is consistent with recent recommendations on measuring attitudes toward the school science subjectinstead of science in general. In addition, it is an instrument that includes the main attitudinal constructs studied in this line of research. Finally, it is a short and easy to administer instrument, ideally for contexts in which time constraints limit the application of longer instruments, particularly in elementary education with younger students. Although SSAS uses single-items measures to examine all its constructs and sub-dimensions,which is in contrast to trends characterized by measuring a construct across multiple items, the initial validation results of the Kennedy et al. study [21] have shown that the SSAS reports robust results. Although multi-item instruments are more stable, reliable and accurate [48], single-item measures can be as psychometrically valid as long measures. For example, studies that have adapted extensive instruments into single-item scales have shown equally reliable and valid results as its multi-item version, such as the single item self-esteem scale [49]or the SIMP [50], which measures the Big Five personality with one item per construct.

Method

Psychometric properties

The quality of a measurement instrument is assessed through different psychometric tests that examines reliability and validity properties [51–54]. In the literature, different terminology and definitions are used to refer to the psychometric properties that should be examined during the scale validation process. The COSMIN initiative was an international effort in clarifying and standardizing the uses of the different terms related to psychometric properties of measurement instruments. Thus, through an international Delphi study, the COnsensus-based Standards for the selection of health Measurement INstruments (COSMIN) taxonomy was developed [55].

In this study we use the Polit-Yang taxonomy[56,57], which builds on the COSMIN study. In short, in the Polit-Yang taxonomy, reliability is defined as "(. . .) scores for people who have not changed are the same for repeated measurements, under several situations"[56] p. 25. The validity domain is defined as whether an instrument "(. . .) measures the construct(s) it purports to measure" [55] p.743.

Relating reliability, the Polit-Yang taxonomy differs between (i) temporal reliability (i.e. weather scores are stable over time when traits have not changed), (ii), internal consistency (i.e. items measuring the same underlying construct), (iii) measurement error (i.e. error in score not related to true changes in the construct). As for validity, there are three components: (i) content and face validity, referring to "(. . .) the degree to which a content of an instrument adequately reflects the construct being measured" [57] p.1750, (ii) criterion validity (i.e. whether the proposed instrument is correlated with scores of existing instruments measuring the same constructs), and (iii) construct validity (i.e. if the construct under study is appropriately represented and conceptualized). Construct validity can be examined through structural validity tests like exploratory factor analysis (EFA) or confirmatory factor analysis (CFA), and through hypothesis testing validity tests, like convergent validity (i.e. items measuring the same constructs should be highly correlated), discriminant validity (i.e. constructs should measure different traits and therefore items from different constructs should be poorly correlated), and discriminative validity (i.e. whether the instrument discriminate between groups that are known to differ). For translated instruments, the cross-cultural validity should also be examined, which involves translation and back-translation and equivalent testing between the original and the translated version [58]. Finally, in addition to reliability and validity, an instrument should also provide evidence of sensitivity, meaning that it should be able to detect the spectrum of differences in the construct under study [59]. Fig 1 shows the theoretical framework adopted for scale translation and validation.

Sample and procedure. Participants were 643 students (47.9% girls, M_{age} = 10 years, SD_{age} = 1.26, range 8–13) enrolled in 3rd (n = 75), 4th (n = 142), 5th (n = 174) and 6th (n = 252) grades of 21 elementary education schools from Burgos, in Spain. Students were enrolled in both rural (48%) and urban (52%) schools. The Spanish version of the SSAS questionnaire was administered in a paper-pencil format upon students' arrival at a week-long intensive curriculum enrichment program developed at an Innovative Education School Center located in Burgos.

Among all the participants, 117 students were randomly selected (52.1% girls, M_{age} = 9.32 years, SD_{age} = 1.14, range 8–13) for computing the concurrent validity of the scale and 88 students were randomly selected (46.6% girls, M_{age} = 9.55, SD_{age} = 1.41; range 8–12) for assessing the temporal stability of the Spanish SSAS instrument. No significant differences were found in neither age or in attitudes towards school science between the whole sample and both subsamples generated for concurrent and temporal stability validity.

Informed written consent was obtained from all participants' parents.The study has been approved by the University of Burgos Education PhD Doctorate Commission and by the Vice-rectorate for Research and Knowledge Transfer.

Measures. Two external measures were used to examine the criterion validity of the S-SSAS by exploring its relationship with two other attitude scales previously validated in the literature. The first instrument, named Scale of Attitudes Toward Science [60], measures three attitudinal constructs (i.e. positive affect toward science, self-confidence in learning science, students valuing science) through 7 Likert-type items. The second instrument, named «Arabic-Speaking Students' Attitudes toward Science Survey» (ASSASS) [24] measures five attitudinal constructs (i.e. attitudes toward science and school science, unfavorable outlook of

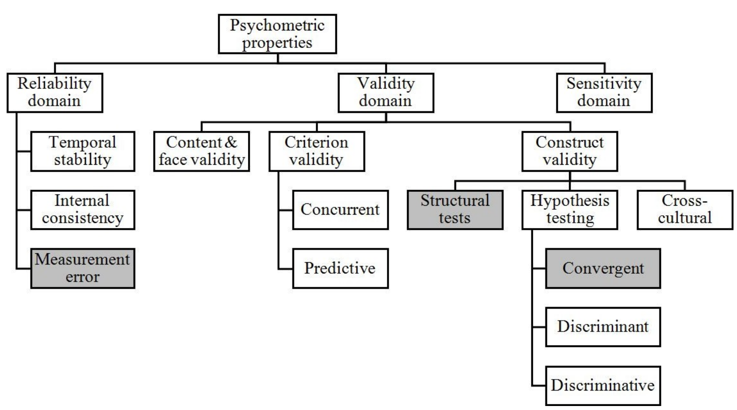


Fig 1. Theoretical framework for scale validation, based on Polit & Yang [56] taxonomy. Psychometric properties in grey quadrants were not measured in this study.

https://doi.org/10.1371/journal.pone.0209027.g001

science, control beliefs about ability in science, behavioral beliefs about the consequences of engaging with science, and intentions to pursue science) through 32 Likert-type items. For this study, the «intentions to pursue science» subscale, composed of six items, was used. These two external measures were selected for computing criterion validity due to items and construct similarities with the S-SSAS (see S1 Table). For example, the «intention to pursue science» subscale of the ASSASS [24]is conceptually identical to the conceptualization of the "intention" construct from the S-SSAS. The «positive affect toward science» construct of the Sabah et al. scale [60] measures how much students enjoy science class (i.e. "I enjoy learning science" p. 695), sharing similarities with the enjoinment construct of the S-SSAS. The «self-confidence in learning science» construct of the Sabah et al. scale (i.e. "I usually do well in science, p. 695) is conceptually similar to the S-SSAS constructs «perceived difficulty of school science» and «perception of self-efficacy in school». Finally, the «students valuing science» construct of the Sabah et al. scale (i.e. I think learning science will help me in my daily life; I need science to learn other school subjects, p. 695) issimilar to the S-SSAS «usefulness of school science» and «relevance of school science».

Results and discussion

Translation and cross-cultural validity

The SSAS was translated into Spanish following a cross-cultural translation procedure [58]. In the first stage, a bilingual professor translated the SSAS from English into Spanish. In the second stage, another bilingual professor back-translated the Spanish version of the scale into the original language (i.e. English). In the third stage, both professors jointly reviewed the equivalence between the original and back-translated version suggesting minor modifications to the Spanish version: «Year 11» from the item assessing intention for future enrolment (i.e. I am very likely to enroll on a science course in Year 11) was translated as «E.S.O», which is the Spanish acronym that refer to the studies that begin at the end of Primary Education (i.e. middle school, students aged 12–16). In the fourth and final stage, six Elementary Education teachers were asked to verify that items wording were appropriated for the reading level of 3rd to 6thelementary grade students. Consequently, the adjective «required» from the semantic differential item assessing usefulness of school science (i.e. For my planned career, knowledge of school science will be worthless/required)was translated as «useful», which is a better Spanish antonym for worthless. Original response format was maintained in all items, with the exception that the questionnaire was applied in a written rather than a web-based format. Original, Spanish and back-translation of the SSAS is outlined in Table 3.

Content and face validity

Content and face validity was determined before large-scale administration using both a panel of experts and target population. An expert committee composed of a university professor in the field of psychology with extensive experience in instrument development and validation, and one university professor with expertise in science teaching and teacher training assessed the content validity of the translated version of the SSAS. Each expert was provided with the ten single-items scales comprising the SSAS and was asked to link each item to the attitudinal constructs or sub-constructs they consider to be measuring, and to evaluate how well each item represented the intended AC (i.e. 1-bad; 2-good; 3-great). Intraclass correlation (ICC) was used to examine agreement between experts in linking items to AC and for evaluating the representativeness of each item, which seems to be the most appropriate statistical method [61,62]. Following Koo & Li [63] guidelines for selecting and reporting Intraclass Correlation Coefficients, ICC estimates and their 95% confident intervals were calculated using the SPSS

Table 3. Original, translated and back-translated S-SSAS.

Original SSAS	*Spanish SSAS*	*Back-translation into English*
1. I am very likely to enroll on a science course in Year 11	Es muy probable que me apunte a Ciencias de la Naturalezaen la ESO	It is very likely that I will enroll in School Science course in ESO / Year 11.
2. I think science is (boring–fun)	Pienso que Ciencias de la Naturalezaes (aburrida–divertida)	I think school science is (boring–fun)
3. I struggle with completing the assignments for science class	Me cuesta terminar las tareas para la clase de Ciencias de la Naturaleza	I have difficulties in completing my homework for school science class.
4. I think I am very good at science	Pienso que soy muy bueno en Ciencias de la Naturaleza	I think I am very good at school science
5. A job as a scientist would be interesting	Un trabajo como científico sería interesante	A job as a scientist would be interesting.
6. For my planned career, knowledge of school science will be (worthless–required)	Para mis futuros estudios, el conocimiento de las clases de Ciencias de la Naturalezaes (inútil– útil)	For my future studies, knowledge of school science class will be (useless–useful).
7. Science helps to make life better	La ciencia ayuda a mejorar la vida	Science helps to improve life.
8. I want to learn about plants in my area	Quiero aprender sobre las plantas de mi entorno	I want to learn about plants in my surroundings.
9. For my everyday life, I think school science is (irrelevant–relevant)	Para mi vida diaria, creo que Ciencias de la Naturaleza es (poco importante–muy importante)	For my daily life, I think school science is (not important–very important)
10. I want to learn about electricity and how it is used in the home	Quiero aprender sobre la electricidad y saber cómo se usa en una casa	I want to learn about electricity and how it is used in a home.

https://doi.org/10.1371/journal.pone.0209027.t003

v.24 statistical software based on single rater, absolute agreement, two-way mixed-effects model.

Intraclass correlation coefficient (ICC) of .956 with 95% confident interval = .833 - .989 revealed a «good» to «excellent» inter-rater reliability in linking items to attitudinal constructs (AC). However, the psychology professor failed at assigning two items to the correct AC. More specifically, he assigned the item «I want to learn about plants in my area" to the «biological vs physical science» AC, and the item «I want to learn about electricity and how it is used in the home» to the «What do I want to learn about?» AC, instead of the other way around.

ICC coefficient of .800 with 95% confident interval = .408 - .945 revealed a «bad» to «excellent» agreement between the raters in evaluating each item representativeness (Koo and Li, 2016). While the professor with background in science teaching considered item «I want to learn about plants in my area» to be a great representativeness of the «What do I want to learn about?» AC, the psychology professor considered it to be only good. Taken together, these results seem to confirm that the SSAS has acceptable content validity.

After establishing the content validity of the S-SSAS, a cognitive interviewing approach known as Think-Aloud Protocol [64] was used with twelve students (three students from 3rd, 4th, 5th and 6th elementary grades, respectively) to assess face validity by examining student's comprehensibility of the items and if their interpretation was similar to that intended by the researchers. Each student was given a copy of the survey and wasindividually prompted to explain what they think when reading and answering each item.

The Think-Aloud Protocol indicated that students have easily understood and correctly interpreted all items. However, some students found it difficult to understand two words. Thus, the adjectives «irrelevant–relevant» from the semantic differential item assessing personal relevance of school science (i.e. For my everyday life, I think school science is irrelevant/relevant) were worded in Spanish as «not important–very important», fostering understanding of lower grades students. No evidence of administration fatigue was identified, with most students taking approximately ten minutes (M = 9.16, SD = 1.05) to complete the task, time that is considerably reduced when students do not have to orally justify and explain their answers. A table with the researchers' interpretation of the item and examples of students' responses can be found in the S2 Table. Instructions for S-SSAS administration can be found in S1 File.

Criterion validity

Predictive validity. Predictive validity was provided using Pearson product-moment correlation coefficient within S-SSAS attitudinal constructs. More specifically, consistent with the literature, the Spanish version of the SSAS would demonstrate criterion validity if there is a positive correlation between intentions and enjoyableness [65–67], between intentions and self-efficacy [68–70], and between usefulness and relevance [71,72].

As for predictive validity, the relationship within the expected S-SSAS AC reported medium to large Pearson's r correlation coefficients that ranged from -334 to .543 (Table 4). More specifically, there was a strong-positive correlation between «intention to enroll in science courses» and «enjoyableness», r = .543, $p < .01$ and a medium-positive correlation between «intention to enroll in science courses» and «self-efficacy» r = .371,$p < .01$, «perceived usefulness» r = .357, $p < .01$ and «perceived relevance» r = .360, $p < .01$ attitudinal constructs. In addition, «enjoyableness of school science» was positively correlated with «self-efficacy» r = .414, $p < .01$, «perceived usefulness» r = .345, $p < .01$ and «perceived relevance» r = .311, $p < .01$ attitudinal constructs. As expected, «perceived difficulty» was negatively correlated with «self-efficacy» r = -.334, $p< .01$, and «perceived usefulness» was positively correlated with «perceived relevance» r = .431, $p < .01$. Taken together, these results are in line with theoretical expectations and previous literature results, thus confirming the criterion validity of the Spanish version of the SSAS.

Concurrent validity. The relationship between S-SSAS's «intention to enroll in further science»ACand ASSASS's «intention to study science» subscale was measured. The concurrent validity of the remaining S-SSAS's AC's was assessed by studying the relationship between

Table 4. Pearson's *r* correlation for criterion and construct validity and item-total correlation for internal consistency.

Attitudinal constructs	*Criterion and construct validity*										*Internal consistency*
	1	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*	*Item-total r*
1. Intention	-	.543*	.171*	.371*	.357*	.360*	.552*	.338*	.225*	.298*	.560
2. Enjoyableness		-	.178*	.414*	.345*	.311*	.390*	.560*	.468*	.352*	.560
3. Difficulty			-	-.334*	.073	.053	-.107	-.236*	-.522*	-.037	.243
4. Self-efficacy				-	.224*	.232*	.215*	.169	.301*	.180	.501
5. Usefulness					-	.431*	.592*	.380*	.079	.428*	.410
6. Relevance						-	.416*	.356*	.160	.406*	.401
7. ASSASS							-	.397*	.076	.439*	
8. PATS								-	.264*	.268*	
9. SCS									-	.189*	
10. SVS										-	

*. Correlation is significant at the 0.01 level (2-tailed)

ASSASS refers to the intention subscale of Abd-El-Khalick et al. instrument[24]

PATS refers to positive affect toward science subscale of Sabah et al. instrument[60]

SCS refers to pelf-confidence in learning science subscale of Sabah et al. instrument[60]

SVS refers to students' valuing science subscale of Sabah et al. instrument[60]

https://doi.org/10.1371/journal.pone.0209027.t004

«enjoyableness of school science» S-SSAS's AC and Sabah et al.'s «positive affect toward science» subscale; between «perceived difficulty of school science» and «perception of self-efficacy in school» S-SSAS's AC and Sabah et al.'s «self-confidence in learning science» subscale; and finally, between «usefulness of school science» and «relevance of school science» S-SSAS's AC and Sabah et al.'s «students' valuing science» subscale.

In relation to concurrent validity, the relationship between S-SSAS AC's and the scales developed by Abd-El-Khalick et al. [24] and Sabah et al. [60] reported medium to large Pearson's r correlation coefficients that ranged from .301 to .560 for the expected attitudinal constructs (Table 4).

More specifically, there was a strong, positive correlation between «intention to enroll in science courses» AC's of both scales, $r = .552$, $p < .01$, and between «enjoyableness» AC's of both scales, $r = .560$,$p < .01$. Also, there was a strong, negative correlation between the S-SSAS «difficulty» AC and Sabah et al. [60] «self-confidence in learning science» subscale, $r = -.522$, $p < .01$. Finally, there was a medium, positive correlation between the «self-efficacy» AC's of both scales $r = .301$, $p < .01$, «usefulness» AC's of both scales $r = .428$, $p < .01$, and «relevance» AC's of both scales $r = .406$, $p < .01$. These results seem to indicate that the S-SSAS shares satisfactory variance with other validated instruments intended to measure students' attitudes toward science, thus further confirming the criterion validity of the S-SSAS.

Construct validity

Discriminant validity. Discriminant and discriminative tests were used to provide evidence for hypothesis testing construct validity. Discriminant validity was examined using Pearson Product-Moment correlation coefficient within S-SSAS attitudinal constructs. Consistent with recent recommendations, moderately-strong correlation between factors are acceptable[73], however, correlation must be below the .80 cut off to consider that the translated S-SSAS provides evidence for discriminant validity [74].

Relating the discriminant validity of the S-SSAS, Pearson's r correlation coefficient within S-SSAS attitudinal constructs were lower than the .80 cutoff (Table 4), thus discriminant validity is confirmed.

Discriminative validity. For discriminative validity, SSAS results were analyzed in terms of gender (girls and boys) and school type (rural and urban schools) differences to determine if they are consistent with the original SSAS results [21]. Specifically, independent sample t test with Bonferroni correction were computed using the SPSS v.24 statistical software[75].

The independent sample t-test revealed no statistically significant differences in attitudinal profiles towards school science between girls and boys (Table 5), neither between students from rural and urban schools (Table 6).

Taken together, these results are consistent with the original SSAS [21], supporting that the Spanish SSAS items are understood and interpreted like-ways by both boys and girls, regardless of whether they are enrolled in rural and urban schools. This further corroborates the valid applicability of the Spanish SSAS.

Reliability

Internal consistency. Cronbach's α cannot be computed for single-item measures [53], and for multidimensional instruments, it should be computed for every construct [76, 77]. However, since the SSAS scale was originally developed to measures students' attitudinal profile toward the school science construct, Cronbach's α was computed for the entire scale, which should report tentative results about the internal consistency of the instrument. The $\alpha > .70$ threshold for acceptable reliability was used [54]. To further examine internal

Table 5. Attitudinal profiles towards school science according to gender variable *SSAS*.

SSAS constructs	*Gender*	*M*	*SD*	*t-test*		
				t-value	*df*	*p*
Intention	Girls	3.17	1.09	1.41	436	.159
	Boys	3.02	1.15			
Enjoyableness	Girls	3.58	1.15	.58	436	.561
	Boys	3.52	1.18			
Difficulty	Girls	3.77	1.18	127	436	,204
	Boys	3.62	1.29			
Self-efficacy	Girls	3.48	1.05	- .845	436	,399
	Boys	3.57	1.03			
Usefulness	Girls	4.01	.81	.508	436	,612
	Boys	3.97	.93			
Relevance	Girls	3.76	.75	- 1.610	436	,108
	Boys	3.88	.76			

https://doi.org/10.1371/journal.pone.0209027.t005

consistency and provide more robust estimates than the Cronbach α, corrected item-total correlation for each of the six S-SSAS AC were calculated, with correlation above .3 suggesting good internal consistency [54,76, 78].

The Spanish SSAS reported good internal consistency (Table 4), with all corrected item-total correlations but the one measuring perceived difficulty of school science being far above the lower limit of .3 that has been suggested in the literature. Cronbach alpha for the entire scale was moderate (α = .704). This value can be considered acceptable for preliminary research. This results indicates that items included in this instrument does correlate very well with the overal scale assesing attitudes toward school science and therefore confirms that the translated SSAS has aceptable internal consistency.

Temporal stability. To assess the temporal stability of the Spanish version of the SSAS, a subgroup of students (n = 88) ranging from third to sixth grade of elementary education completed the scale at two times with a 10-days interval. A 10-days' time span was used as it's been considered in previous literature as adequate for computing test-retest reliability [79]. The first and second administration scale was done 10 days before and upon student's arrival at the

Table 6. Attitudinal profiles towards school science according to school type variable.

SSAS constructs	*Gender*	*M*	*SD*	*t-test*		
				t-value	*df*	*p*
Intention	Urban	3.08	1.10	-.397	436	.692
	Rural	3.11	1.16			
Enjoyableness	Urban	3.54	1.14	-.146	436	.884
	Rural	3.56	1.21			
Difficulty	Urban	3.71	1.25	.384	436	.701
	Rural	3.66	1.19			
Self-efficacy	Urban	3.46	1.03	-1.584	436	.114
	Rural	3.62	1.05			
Usefulness	Urban	3.97	.87	-.494	436	.622
	Rural	4.01	.88			
Relevance	Urban	3.83	.73	.209	436	.835
	Rural	3.81	.79			

https://doi.org/10.1371/journal.pone.0209027.t006

Table 7. Intraclass correlation between S-SSAS constructs at Time 1 and Time 2.

SSAS AC	*r*	*Intraclass correlation*	*95% Confidence Interval*		*Mean difference (Time 2 –Time 1)*
			Lower Bound	*Upper Bound*	
Intention	.771	.868	.798	.913	.11
Enjoyableness	.464	.597	.374	.739	.39
Difficulty	.556	.716	.566	.814	-.02
Self-efficacy	.595	.748	.615	.835	.01
Usefulness	.631	.776	.657	.853	.02
Relevance	.790	.880	.817	.921	-.05
Total	.779	.873	.805	.917	.07

r refers to Pearson's correlation between SSAS constructs at Time 1 and Time 2

https://doi.org/10.1371/journal.pone.0209027.t007

week-long intensive curriculum enrichment program, respectively. Pearson product-moment correlation coefficient was used to compute students' responses between time 1 and time 2. In addition, as Intraclass correlation (ICC) seems to be more suitable than Pearson correlation coefficient when performing temporal stability reliability analysis [80], ICC was used to further estimate test-retest reliability by taking into account the measurement error [61]. More specifically, ICC estimates and their 95% confident intervals were calculated using the SPSS v.24 statistical software based on single ratter, absolute agreement, two-way mixed-effects model [63].

The relationship between each student S-SSAS global result at Time 1 and Time 2 revealed a significant large Pearson's *r* correlation coefficients, $r = .779$, $p < .01$. According to each AC, the test-retest reliability of the S-SSAS ranged from medium to large significant Pearson's *r* correlation coefficients ($r = .464$ - .790). In Table 6 Pearson's correlations results between SSAS at Time 1 and Time 2 are reported. Examination of this table reveals that «enjoyableness of school science» is the AC with lowest test-retest reliability, and that «relevance of school science to society», and «intention to enroll in further science» AC's reported the best temporal stability reliability.

ICC of .873 with 95% confident interval = .805 - .917 revealed a «good» to «excellent» test-retest reliability for the S-SSAS global results, with a mean difference between the two administrations (Time 2 –Time 1) of only .07. Intraclass coefficient results for each S-SSAS attitudinal construct confirms the temporal stability of the S-SSAS scale, with all but one AC (i.e. enjoyableness of school science) revealing a Intraclass correlation above .7 (Table 7).Among the S-SSAS constructs, «Enjoyableness of school science» has reported less test-retest reliability, suggesting that it may be an attitudinal construct that is more sensitive to classroom variables (e.g. science content, teaching methodology, classroom activities). In contrast, «relevance of school science to society» and «intention to enroll in further science» AC's seems as the most stable over time, coinciding with earlier studies that show how difficult it is to change student's intentions to enroll in a scientific career [81]. Overall, these results indicate that the Spanish version of the SSAS has a good temporal stability with a 10-days span between the first and second administration.

Sensitivity

Scale sensitivity of the translated SSAS was explored by computing (i) the distribution of responses according to the spectrum of responses categories for each item, (ii) mean, standard deviation, observed range, (iii) item variance and (iv) skewness and kurtosis. Response distribution and mean of each item should confirm no evidence of ceiling effect, namely, response distribution should be predominantly close to the center of the range of possible scores or

Table 8. Distribution of responses.

SSAS items	Response category (%)				
	1	2	3	4	5
Item 1	13	8.7	46.1	20.8	11.4
Item 2	8.2	8.2	26.5	34.5	22.6
Item 3	33.1	27.2	21.9	11.2	6.6
Item 4	5.5	8.0	31.7	37.9	16.9
Item 5	7.1	7.8	19.6	26.7	38.8
Item 6	3.4	2.3	15.5	32.4	46.3
Item 7	2.3	4.6	20.1	31.3	41.8
Item 8	6.8	10.7	36.3	29.5	16.7
Item 9	3.0	8.0	32.0	29.0	28.1
Item 10	7.5	11.9	19.6	28.3	32.6

https://doi.org/10.1371/journal.pone.0209027.t008

responses [53]. In addition, items with high variance are desirable, which would confirm that the translated SSAS is capable of discriminate among individuals with different attitudinal profiles towards school science [53]. In relation to skewness and kurtosis, several recommendation from the literature were followed: Kline [82] suggested indices not higher than 3 for skewness and 10 for kurtosis and others have established ±2 as acceptable limits [74,83,84].

The distribution of responses covered all the spectrum of response categories (Table 8). Most of responses were concentrated in the middle range of the scale, however, responses to item 3 (measuring perceived difficulty of school science) were concentrated in the first two responses categories (60.3%), indicating that the majority of students do no perceive school science as a difficulty subject. Similarly, items 6 and 7, measuring usefulness and relevance of school science, respectively, reported responses concentrated in the last two responses categories (78.7% and 73.1%), indicating that in general students perceive school science as useful and relevant.

The mean scores of most items were the middle point of the scale, ranging from 2.31 to 4.16 on a scale from 1 to 5 (Table 9). As shown by the negative skewness of all but item 3, responses were weighted towards the positive end of the 5-point response format. However, the values of skewness and kurtosis fall within the acceptable range of ±2 limits [74,83,84], suggesting that the scale did not report evidence of ceiling effects. Based on these results, the Spanish School Science Attitude survey (S-SSAS) has satisfactory sensitivity.

Table 9. Descriptive statistics of translated SSAS items.

SSAS items	Range	*M*	*SD*	Variance	Skewness	Kurtosis
Item 1	1–5	3.09	1.12	1.267	-.244	-.319
Item 2	1–5	3.55	1.17	1.360	-.631	-.281
Item 3	1–5	2.31	1.23	1.501	.628	-.582
Item 4	1–5	3.53	1.04	1.078	-.572	.046
Item 5	1–5	3.82	1.23	1.504	-.842	-.240
Item 6	1–5	4.16	1	.995	-1.298	1.519
Item 7	1–5	4.06	1	1.006	-.935	.387
Item 8	1–5	3.38	1.1	1.198	-.353	-.336
Item 9	1–5	3.71	1.1	1.107	-.434	-.409
Item 10	1–5	3.67	1.3	1.568	-.642	-.624

https://doi.org/10.1371/journal.pone.0209027.t009

Conclusion

The purpose of this study was to translate the single-item SSAS survey [21]for its use by elementary Spanish-speaking students and to test its validity and reliability. Through a multistage translation approach and multiple psychometric evaluations, the current study provides evidence that the S-SSAS is a valid and reliable instrument for measuring elementary students' attitudes toward school science.

Regarding translation validity, the Think-Aloud structured interviews indicated that the S-SSAS is easily understood and interpreted by Spanish students in 3rd to 6th grades of Elementary Education. Additionally, the panel of experts shared acceptable agreement in the assignments of items to each attitudinal construct. These results provide face and content validity for the scale.

Psychometric evaluation of the S-SSAS indicated an adequate level of internal consistency reliability and that response were well distributed along the response categories, showing great sensitivity and no evidence of extreme response tendency. Intraclass correlation coefficient supported the temporal stability reliability of the translated SSAS, with a 10-days span between first and second administration. Pearson correlation coefficient reported acceptable predictive validity with strong correlation between expected attitudinal constructs based on results reported in specialized literature of attitudes, and also great concurrent validity, thus obtaining high correlation between the attitudinal constructs of the S-SSAS scale and two attitudes measures of conceptual convergence already validated in the literature. In addition, discriminative validity between S-SSAS constructs was confirmed and parametric and non-parametric tests indicated that the translated version of the SSAS replicates similar results to those of its original version in terms of sex and rural or urban school variables, thus confirming discriminative validity and further supporting the applicability of the S-SSAS.

One main limitation should be acknowledged. The concurrent validity was assessed using two other scales that weren't previously validated for Spanish-speaking students. Ideally, S-SSAS results should have been correlated with other Spanish scales in order to obtain more reliable results. However, given the absence of validated Spanish scales for elementary education, we were forced to use two scales that were validated in another language. To counter this limitation, both Abd-El-Khalick et al. [24] and Sabah et al. [60] scales were translated into Spanish using the same multistage translation approach as for the S-SSAS. Future studies should consider subscales from the Spanish version of the TOSRA [19] to further examine the concurrent validity of the S-SSAS. Although the Navarro et al.'s [19] TOSRA was validated with high school students, it's psychometric properties for Spanish speaking students are more likely to be stronger than those of the instruments used in this study. In addition, future studies should administer the original and the Spanish SSAS to a bilingual sample and compute equivalence tests between both scores, thus further examining the S-SSAS applicability.

Taken together, in can be concluded that the Spanish SSAS represents a first effort to provide a valid and reliable measure to examine attitudes toward school science of Spanish-speaking elementary students, especially when there are time constraints for data collection. This instrument is easy to administer and quick to answer because it consists of only ten single-item measures equally interpretable by both girls and boys from rural and urban schools. Therefore, it can be useful in quasi-experimental time-series designs where several data measurement is needed.

Supporting information

S1 Table. Similarities between SSAS and two external measures used for examining concurrent validity.
(PDF)

S2 Table. Comparison between researchers and students interpretation of the S-SSAS items.
(PDF)

S1 File. Administration instructions.
(PDF)

Author Contributions

Conceptualization: Radu Bogdan Toma.

Formal analysis: Radu Bogdan Toma.

Funding acquisition: Jesús Ángel Meneses Villagrá.

Investigation: Radu Bogdan Toma.

Methodology: Radu Bogdan Toma.

Resources: Jesús Ángel Meneses Villagrá.

Supervision: Jesús Ángel Meneses Villagrá.

Validation: Radu Bogdan Toma, Jesús Ángel Meneses Villagrá.

Visualization: Radu Bogdan Toma.

Writing – original draft: Radu Bogdan Toma.

Writing – review & editing: Radu Bogdan Toma, Jesús Ángel Meneses Villagrá.

References

1. Sadler PM, Gerhard S, Zahra H, Tai R. Stability and Volatility of STEM career interest in high school: A gender Study. Sci Educ. 2012; 96: 411–427. https://doi.org/10.1002/sce.21007
2. Smith E. Women into science and engineering? Gendered participation in higher education STEM subjects. Br Educ Res J. 2011; 67: 993–1014. https://doi.org/10.1080/01411926 2010.515019
3. MECD. Datos y cifras del Sistema Universitario español. Curso 2015/2016. Secretaría General Técnica; 2016.
4. Archer L, Dewitt J, Osborne J, Dillon J, Willis B, Wong B. "Doing" science versus "being" a scientist: Examining 10/11-year-old schoolchildren's constructions of science through the lens of identity. Sci Educ. 2010; 94: 617–639. https://doi.org/10.1002/sce.20399
5. Kennedy J, Lyons T, Quinn F. The continuing decline of science and mathematics enrolments in Australian high schools. Teach Sci. 2014; 60: 34–46.
6. Osborne J, Simon S, Collins S. Attitudes towards science: A review of the literature and its implications. Int J Sci Educ. 2003; 25: 1049–1079. https://doi.org/10.1080/0950069032000032199
7. Potvin P, Hasni A. Interest, motivation and attitude towards science and technology at K-12 levels: A systematic review of 12 years of educational research. Stud Sci Educ. 2014; 50: 85–129. https://doi.org/10.1080/03057267.2014.881626
8. Blalock CL, Lichtenstein MJ, Owen S, Pruski L, Marshall C, Toepperwein MA. In pursuit of validity: A comprehensive review of science attitude instruments 1935–2005. Int J Sci Educ. 2008; 30: 961–977. https://doi.org/10.1080/09500690701344578
9. Germann PJ. Development of the attitude toward science in school assessment and its use to investigate the relationship between science achievement and attitude toward science in school. J Res Sci Teach. 1988; 25: 689–703.
10. Noll VH. Measuring the scientific attitude. J Abnorm Soc Psychol. 1935; 30: 145–154.
11. Fraser BJ. Development of a test of science-related attitudes. Sci Educ. 1978; 62: 509–515. https://doi.org/10.1002/sce.3730620411

12. Summers R, Abd-El-Khalick F. Development and validation of an instrument to assess student attitudes toward science across grades 5 through 10. J Res Sci Teach. 2018; 55: 172–205. https://doi.org/10.1002/tea.21416
13. Zhang D, Campbell T. The psychometric evaluation of a three-dimension elementary science attitude survey. J Sci Teacher Educ. 2011; 22: 595–612. https://doi.org/10.1007/s10972-010-9202-3
14. Guzey SS, Harwell M, Moore T. Development of an instrument to assess attitudes toward Science, Technology, Engineering, and Mathematics (STEM). Sch Sci Math. 2014; 114: 271–279. https://doi.org/10.1111/ssm.12077
15. Wang TL, Berlin D. Construction and validation of an instrument to measure taiwanese elementary students' attitudes toward their Science class. Int J Sci Educ. 2010; 32: 2413–2428. https://doi.org/10.1080/09500690903431561
16. Campbell DT, Stanley JC, Cage NL. Experimental and quasi-experimental designs for research. Boston, MA: Houghton, Mufflin and Company; 1963.
17. De Pro A, Pérez A. Actitudes de los alumnos de Primaria y Secundaria ante la visión dicotómica de la Ciencia. Enseñanza las Ciencias. 2014; 32: 111–132. https://doi.org/10.5565/rev/ensciencias.1015
18. Vázquez Á, Manassero MA. El declive de las actitudes hacia la ciencia de los estudiantes: Un indicador inquietante para la educación científica. Rev Eureka sobre Enseñanza y Divulg las Ciencias. 2008; 5: 274–292.
19. Navarro M, Förster C, González C, González-Pose P. Attitudes toward science: Measurement and psychometric properties of the Test of Science-Related Attitudes for its use in Spanish-speaking classrooms. Int J Sci Educ. Taylor & Francis; 2016; 38: 1459–1482. https://doi.org/10.1080/09500693.2016.1195521
20. Fraser BJ. Test of science-related attitudes. Melbourne: Australian Council for Educational Research; 1981.
21. Kennedy J, Quinn F, Taylor N. The school science attitude survey: A new instrument for measuring attitudes towards school science. Int J Res Method Educ. Taylor & Francis; 2016; 39: 422–445. https://doi.org/10.1080/1743727X.2016.1160046
22. Klopfer LE. Evaluation of learning in science. In: Bloom BS, Hastings JT, Madaus GF, editors. Handbook of formative and summative evaluation of student learning. London: McGraw-Hill; 1971.
23. Gardner PL. Attitudes to Science: A review. Stud Sci Educ. 1975; 2: 1–41.
24. Abd-El-Khalick F, Summers R, Said Z, Wang S, Culbertson M. Development and large-scale validation of an instrument to assess arabic-speaking students' attitudes toward science. Int J Sci Educ. 2015; 37: 2637–2663. https://doi.org/10.1080/09500693.2015.1098789
25. Hillman SJ, Zeeman SI, Tilburg CE, List HE. My Attitudes Toward Science (MATS): The development of a multidimensional instrument measuring students' science attitudes. Learn Environ Res. Springer Netherlands; 2016; 19: 203–219. https://doi.org/10.1007/s10984-016-9205-x
26. Ajzen I. The theory of planned behavior. Organ Behav Hum Decis Process. 1991; 50: 179–211.
27. Wigfield A, Eccles JS. The development of achievement task values: a theoretical analysis. Dev Rev. 1992; 12: 265–310.
28. Wigfield A, Eccles JS. Expectancy-value theory of motivation. Contemp Educ Psychol. 2000; 25: 68–81. https://doi.org/10.1006/ceps.1999.1015 PMID: 10620382
29. Wigfield A, Eccles JS. The development of competence beliefs, expectancies for success, and achievement values from childhood through adolescence. In: Wigfield A, Eccles JS, editors. Development of achievement motivation. San Diego: Academic Press; 2002. pp. 91–120.
30. Eccles JS, Wigfield A. In the mind of the achiever: the structure of adolescents' academic achievement related-beliefs and self-perceptions. Personal Soc Psychol Bull. 1995; 21: 215–225.
31. Eccles JS, Wigfield A, Harold RB, Blumenfeld PB. Age and gender differences in children's self- and task perceptions during elementary school. Child Dev. 1993; 64: 830–847. PMID: 8339698
32. Said Z, Summers R, Abd-El-Khalick F, Wang S. Attitudes toward science among grades 3 through 12 Arab students in Qatar: Findings from a cross-sectional national study. Int J Sci Educ. 2016; 38: 621–643. https://doi.org/10.1080/09500693.2016.1156184
33. Munby H. Issues of validity in science attitude measurement. J Res Sci Teach. 1997; 34: 337–341.
34. Munby H. Thirty studies involving 'Scientific Attitude Inventory': What confidence can we have in this Instrument? J Res SciTeach. 1983; 20: 141–162.
35. Pérez Manzano A, De Pro Bueno A. Algunos datos sobre la visión de los niños y de las niñas sobre las ciencias y del trabajo científico. iQualRev Género e Igual. 2018; 0: 18. https://doi.org/10.6018/iQual.306091

36. Marbá-Tallada A, Márquez Bargalló C. ¿Qué opinan los estudiantes de las clases de ciencias? Un estudio transversal de sexto de Primaria a cuarto de ESO. Enseñanza las ciencias. 2010; 28: 19–30.
37. Pérez-Franco D, de Pro-bueno AJ, Pérez-Manzano A. Actitudes ambientales al final de la ESO. Un estudio diagnóstico con alumnos de Secundaria de la región de Murcia. Rev Eureka sobre Enseñanza y Divulg las Ciencias. 2018; 15: 3501. https://doi.org/10.25267/Rev
38. Vázquez Alonso Á, Manassero–Mas MA. El declive de las actitudes hacia la ciencia de los estudiantes: un indicador inquietante para la educación científica. Rev Eureka sobre Enseñanza y Divulg las Ciencias. 2008; 5: 274–292.
39. Vázquez-Alonso Á, Manassero-Mas MA. La relevancia de la educación científica: actitudes y valores de los estudiantes relacionados con la Ciencia y la Tecnología. Enseñanza las Ciencias. 2009; 27: 33–48.
40. Vázquez-Alonso Á, Manassero–Mas MA. Perfiles actitudinales de la elección de ciencias en secundaria según el sexo y el tipo de educación. Rev Electrónica Enseñanza las Ciencias. 2010; 9: 242–260.
41. Vázquez-Alonso Á, Manassero–Mas MA. Imagen de la ciencia y la tecnología al final de la educación obligatoria. Cult y Educ. 2004; 16: 385–398.
42. Vázquez-Alonso Á, Manassero–Mas MA. El descenso de las actitudes hacia la ciencia de chicos y chicas en la educación obligatoria. CiênciaEduc. 2011; 17: 249–268.
43. Vázquez-Alonso Á, Acevedo Díaz JA, Manassero–Mas MA, Acevedo Romero P. Actitudes del alumnado sobre ciencia, tecnología y sociedad, evaluadas con un modelo de respuesta múltiple. RevElectronInvestigEduc. 2006; 8: 1–37.
44. Vázquez-Alonso Á, Manassero–Mas MA. Una evaluación de las actitudes relacionadas con la ciencia. Enseñanza las Ciencias. 1997; 15: 199–213.
45. Schreiner C, Sjøberg S. ROSE: The Relevance of Science Education. Sowing the seeds of Rose. Background, rationale, questionnaire development and data collection for ROSE (The Relevance of Science Education)—a comparative study of students' views of science and science education. Oslo: Acta Didactica; 2004.
46. Costello AB, Osborne JW. Best Practices in exploratory factor analysis: Four recommendations for getting the most from your analysis. Pract Assessment, Res Educ. 2005; 10: 1–9. https://doi.org/10.1.1.110.9154
47. Gaskin CJ, Happell B. International Journal of Nursing Studies On exploratory factor analysis A review of recent evidence, an assessment of current practice, and recommendations for future use. Int J Nurs Stud. 2014; 51: 511–521. https://doi.org/10.1016/j.ijnurstu.2013.10.005 PMID: 24183474
48. Bowling A. Just one question: If one question works, why ask several? J Epidemiol Community Health. 2005; 59: 342–345. https://doi.org/10.1136/jech.2004.021204 PMID: 15831678
49. Robins RW, Hendin HM, Trezniewski KH. Measuring global self-esteem: Construct validation of a single-item measure and the Rosenberg Self-Esteem Scale. Personal Soc Psychol Bull. 2001; 27: 151–161.
50. Woods SA, Hampson SE. Measuring the big five with single items using a bipolar response scale. Eur J Pers. 2005; 19: 373–390.
51. Arias MRM, Lloreda MJH, Lloreda MVH. Psicometría. Alianza Editorial; 2014.
52. Clark LA, Watson D. Constructing validity: basic issues in objective scale development. Psychol Assess. 1995; 7: 309–319. https://doi.org/10.1037/1040-3590.7.3.309
53. DeVellis RF. Scale development. Theory and applications. Los Angeles: SAGE; 2017.
54. Nunnally JC, Bernstein IH. Psychometric theory. New York: McGraw-Hill; 1994.
55. Mokkink LB, Terwee CB, Patrick DL, Alonso J, Stratford PW, Knol DL, et al. The COSMIN study reached international consensus on taxonomy, terminology, and definitions of measurement properties for health-related patient-reported outcomes. J Clin Epidemiol. Elsevier Inc; 2010; 63: 737–745. https://doi.org/10.1016/j.jclinepi.2010.02.006 PMID: 20494804
56. Polit DF, Yang F. Measurement and the measurement of change: A primer for health professionals. Philadelphia: Lippincott Williams & Wilkins; 2016.
57. Polit DF. Assessing measurement in health: Beyond reliability and validity. Int J Nurs Stud. Elsevier Ltd; 2015; 52: 1746–1753. https://doi.org/10.1016/j.ijnurstu.2015.07.002 PMID: 26234936
58. Borsa JC, Damásio BF, Bandeira DR. Cross-cultural adaptation and validation of psychological instruments: some considerations. Paidéia. 2012; 22: 423–432.
59. Erdinç O, Lewis JR. Psychometric evaluation of the T-CSUQ: The Turkish version of the computer system usability questionnaire. Int J Hum Comput Interact. 2013; 29: 319–326. https://doi.org/10.1080/10447318.2012.711702

60. Sabah S, Hammouri H, Akour M. Validation of a scale of attitudes toward science across countries using rasch model: Findings from TIMSS. J Balt Sci Educ. 2013; 12: 692–703.
61. de Vet HCW, Terwee CB, Knol DL, Bouter LM. When to use agreement versus reliability measures. J Clin Epidemiol. 2006; 59: 1033–1039. https://doi.org/10.1016/j.jclinepi.2005.10.015 PMID: 16980142
62. Shrout PE, Fleis JL. Intraclass Correlations: Uses in assessing rater reliability. Psychol Bull. 1979; 86: 420–428. PMID: 18839484
63. Koo TK, Li MY. A guideline of selecting and reporting intraclass correlation coefficients for reliability research. J Chiropr Med. Elsevier B.V.; 2016; 15: 155–163. https://doi.org/10.1016/j.jcm.2016.02.012 PMID: 27330520
64. Beatty PC, Willis GB. Research synthesis: The practice of cognitive interviewing. Public Opin Q. 2007; 71: 287–311. https://doi.org/10.1093/poq/nfm006
65. Demirel M, Dağyar M. Effects of problem-based learning on attitude: A meta-analysis study. Eurasia J Math Sci Technol Educ. 2016; 12: 2115–2137. https://doi.org/10.12973/eurasia.2016.1293a
66. Palmer TA, Burke PF, Aubusson P. Why school students choose and reject science: A study of the factors that students consider when selecting subjects. Int J Sci Educ. Taylor & Francis; 2017; 39: 645–662. https://doi.org/10.1080/09500693.2017.1299949
67. Schroeder CM, Scott TP, Tolson H, Huang T-Y, Lee Y-H. A meta-analysis of national research: Effects of teaching strategies on student achievement in science in the United States. J Res Sci Teach. 2007; 44: 1436–1460. https://doi.org/10.1002/tea.20212
68. Ceci SJ, Williams WM. Why aren't more women in science. Top researchers debate the evidence. Washington, DC: American Psychological Association; 2007.
69. Chachashvili-Bolotin S, Milner-Bolotin M, Lissitsa S. Examination of factors predicting secondary students' interest in tertiary STEM education. Int J Sci Educ. Taylor & Francis; 2016; 38: 366–390. https://doi.org/10.1080/09500693.2016.1143137
70. Betz NE. Career self-efficacy: Exemplary recent research and emerging directions. J Career Assess. 2007; 15: 403–422. https://doi.org/10.1177/1069072707305759
71. Andersen HM, Krogh LB, Lykkegaard E. Identity matching to scientists: Differences that make a difference? Res Sci Educ. 2014; 44: 439–460. https://doi.org/10.1007/s11165-013-9391-9
72. Sellami A, El-Kassem RC, Al-Qassass HB, Al-Rakeb NA. A path analysis of student interest in STEM, with specific reference to Qatari students. Eurasia J Math Sci Technol Educ. 2017; 13: 6045–6067. https://doi.org/10.12973/eurasia.2017.00999a
73. Field A. Discovering statistics using SPSS. London: SAGE; 2009.
74. Brown TA. Confirmatory factor analysis for applied research. New York: Guilford Press; 2006.
75. IBM C. IBM SPSS statistics for Windows. Version 24.0. Armonk, NY: IBM Corp; 2016.
76. Cortina JM. What is coefficient alpha? An examination of theory and applications. J Appl Psychol. 1993; 78: 98–104.
77. Tavakol M, Dennick R. Making sense of Cronbach's alpha. Int J Med Educ. 2011; 2: 53–55. https://doi.org/10.5116/ijme.4dfb.8dfd PMID: 28029643
78. Nunnally JC. Psychometric theory. New York: McGraw-Hill; 1978.
79. Keszei AP, Novak M, Streiner DL. Introduction to health measurement scales. J Psychosom Res. Elsevier Inc.; 2010; 68: 319–323. https://doi.org/10.1016/j.jpsychores.2010.01.006 PMID: 20307697
80. Terwee CB, Schellingerhout JM, Verhagen AP, Koes BW, De Vet HCW.Methodological quality of studies on the measurement properties of neck pain and disability questionnaires: A systematic review. J Manipulative PhysiolTher. 2011; 34: 261–272. https://doi.org/10.1016/j.jmpt.2011.04.003 PMID: 21621728
81. Toma RB, Greca IM. The effect of integrative STEM instruction on elementary students' attitudes toward science. EURASIA J Math Sci Technol Educ. 2018; 14: 1383–1395. https://doi.org/10.29333/ejmste/83676
82. Kline P. The handbook of psychological testing. New York: Routledge; 2000.
83. Gravetter F, Wallnau L. Essentials of statistics for the behavioral sciences. Belmont, CA: Wadsworth; 2014.
84. Trochim WM, Donnelly JP. The research methods knowledge base. Cincinnati: Atomic Dog Publishing Inc; 2006.

Appendix 8: Pre-test data analysis study

This appendix includes the full-text of the published article about the analysis of the pre-test data gathered using the S-SSAS instrument.

Toma, R. B., & Meneses Villagrá, J. A. (2019). Preferencia por contenidos científicos de física o de biología en Educación Primaria: un análisis clúster. *Revista Eureka sobre Enseñanza y Divulgación de las Ciencias, 16*(1), 1104. doi: 10.25267/Rev_Eureka_ensen_divulg_cienc.2019.v16.i1.1104

Preferencia por contenidos científicos de física o de biología en Educación Primaria: un análisis clúster

Radu Bogdan Toma

Departamento de Didácticas Específicas. Facultad de Educación. Universidad de Burgos. Burgos. España.
rbtoma@ubu.es
ORCID: https://orcid.org/0000-0003-4846-7323

Jesús Ángel Meneses Villagrá

Departamento de Didácticas Específicas. Facultad de Educación. Universidad de Burgos. Burgos. España.
meneses@ubu.es
ORCID: https://orcid.org/0000-0003-4839-0418

[Recibido: 8 Junio 2018. Revisado: 9 Octubre 2018. Aceptado: 22 Octubre 2018]

Resumen: Las estadísticas de matriculación en carreras de ciencia, tecnología, ingeniería y matemática (STEM) evidencian una gran infrarrepresentación de las mujeres en la mayoría de estas disciplinas, especialmente en física e ingeniería. En este estudio se analiza el interés de niñas y niños de Educación Primaria (n = 733) por el contenido de disciplinas científicas de física y de biología utilizando un método de análisis cuantitativo basado en técnicas de agrupamiento por conglomerados K-Medias, analizándose las diferencias en las actitudes de los estudiantes hacia la ciencia escolar en función de su interés por estas disciplinas científicas. Los resultados revelan la existencia de un perfil actitudinal sesgado sexualmente, caracterizado por un alto interés por contenidos propios de ciencias biológicas y un bajo interés por el aprendizaje de la física por parte de las niñas, y resultados opuestos en el caso de los niños. Además, en comparación con los niños, las niñas han mostrado menos interés por contenidos de física, han sido mayoría en los perfiles actitudinales caracterizados por un bajo interés general por contenidos científicos y han sido minoría en los perfiles caracterizados por un alto interés por ambas disciplinas. Estos resultados parecen mostrar que el interés por disciplinas científicas de física o de biología empieza a forjarse desde niveles elementales del sistema educativo, hecho que podría dar lugar a la posterior segregación de género en las disciplinas STEM. Se discuten las implicaciones educativas y de investigación de estos resultados.

Palabras clave: actitudes hacia la ciencia, K-medias, conglomerados, ciencias biológicas, ciencias físicas

Preference for physics or biology scientific content in Elementary Education: a cluster analysis

Abstract: Statistics on enrolment in science, technology, engineering and mathematics (STEM) show a large under-representation of women in most of these disciplines. This study analyses the preferences of elementary education girls and boys (n = 733) for the content of physics and biology using a quantitative analysis method based on clustering techniques (K-Means) and by analyzing the differences in their attitudes toward school science according to their preferences. The results reveal the existence of a sexually biased attitudinal profile, characterized by a high preference for the content of the biological sciences and a low preference for learning physics among girls, and opposite results in the case of boys. In addition, in comparison with boys, girls showed less interest in physics science content than boys, were majority in the attitudinal profiles characterized by a low overall interest in science content and minority in the profiles characterized by a high interest in both disciplines. These results seem to show that the preference for scientific disciplines in physics or biology are beginning to form from elementary levels of the educational system, which could lead to the subsequent gender segregation existing in STEM disciplines. The educational and research implications of these findings are discussed.

Keywords: attitudes towards science, K-mean, cluster analysis, biology, physics

Para citar este artículo: Toma, R. B., Meneses Villagrá, J. A. (2019) Preferencia por contenidos científicos de física o de biología en Educación Primaria: un análisis clúster. *Revista Eureka sobre Enseñanza y Divulgación de las Ciencias* 16(1), 1104. doi: 10.25267/Rev_Eureka_ensen_divulg_cienc.2019.v16.i1.1104

Introducción

En 2013 más de la mitad de los títulos de estudios superiores en disciplinas de ciencia, tecnología, ingeniería y matemática (STEM) de los países miembros de la Unión Europea fueron obtenidos por mujeres, sin embargo, apenas un 30% de los graduados en ingeniería eran mujeres (OECD, 2013). Además, en la etapa de Educación Secundaria, cuyas edades comprenden en los 14 y 18 años de edad, las alumnas son una minoría en asignaturas de física, un patrón generalizado en diversos países como Francia (Rapoport y Thibout 2018), Holanda (Buser, Niederle y Oosterbeek 2014) o Australia (Collins, Kenway y McLeod 2000). En términos de ocupación STEM, las mujeres siguen representando un porcentaje muy inferior en comparación con los hombres (Beede *et al.* 2011; Deloitte 2016; INE 2017; NSF 2013; UNESCO 2017). España no es una excepción. Si bien el número de mujeres en estudios universitarios de ciencias ha aumentado recientemente, aún persiste la disparidad entre el porcentaje de mujeres y de hombres en algunas disciplinas STEM. Así, apenas un 25.4% de las matriculaciones en titulaciones universitaria de ingeniería y arquitectura pertenecen a mujeres. Además, apenas un 2.5% del total de egresados en estudios superiores en física y matemáticas son mujeres (INE 2017; MECD 2016). Por lo tanto, la retención de las mujeres en STEM resulta crítica en la actualidad.

En la literatura existen numerosos estudios que indican que el comportamiento y las elecciones de los estudiantes pueden verse moldeadas por sesgos implícitos derivados de la exposición a estereotipos culturales generalizados que conciben la Ciencia como una profesión masculinizada (Dasgupta 2011) y a las mujeres como menos competentes (Eagly y Mladinic 1993), siendo estos estereotipos sostenidos y exhibidos por ambos géneros (Nosek, Banaji y Greenwald 2002). Aunque numerosas investigaciones han confirmado la existencia de sesgos de género en ámbitos familiares y educativos en cuanto a las competencias matemáticas y científicas de las niñas y los niños (Dickhauser y Meyer 2006; Gunderson, Ramirez, Levine y Beilock 2012), aún no resulta claro a partir de qué edad estos estereotipos podrían tener efecto en las vocaciones y el interés de los estudiantes por las disciplinas STEM. Por ello, en este estudio se pretende identificar el interés de las niñas y los niños de Educación Primaria por contenidos de física o de biología y analizar si existen diferencias en sus intereses según la variable sexo y edad. Además, se pretende examinar si existen diferencias en las actitudes hacia la asignatura de Ciencias de la Naturaleza en función de los intereses por sus contenidos. Más específicamente, se pretende descubrir si desde una temprana edad se pueden identificar diferencias en los intereses de los niños y las niñas por disciplinas científicas de física y de biología y, en caso afirmativo, determinar si existen diferencias en las actitudes hacia Ciencias de la Naturaleza en los grupos que tienen mayor interés por una u otra disciplina. Las tres preguntas de investigación que guían el presente estudio son:

P1. ¿Cómo afecta el sexo y el curso escolar a las preferencias por estudiar contenidos científicos de física y de biología?

P2. Las preferencias, ¿simulan patrones de matriculaciones identificadas en etapas secundarias y terciarias?

P3. ¿Existen diferencias en las actitudes de los estudiantes hacia la asignatura de Ciencias de la Naturaleza en función de su preferencia por estudiar contenidos de física o de biología?

De acuerdo con los datos nacionales de matriculación (MECD 2016), se espera identificar un subgrupo de niñas que se caracteriza por un bajo interés por contenidos de física y un alto interés por contenidos propios de las ciencias biológicas, y resultados opuestos en el caso de los niños. Además, en consonancia con investigaciones pasadas que señalan actitudes desfavorables hacia la Ciencia por parte de las niñas, (p. ej.: de Pro Bueno y Pérez Manzano

2014; Denessen, Vos, Hasselman y Louws 2015; Marbá-Tallada y Márquez Bargalló 2010; Vázquez y Manassero 2008), se espera identificar otro subgrupo de niñas con bajo interés tanto por la física como por la biología. Por último, dado el patrón generalizado reportado en la literatura de actitudes hacia la Ciencia señalando diferencias significativas a favor de los niños (Caleon y Subramaniam 2008; DeWitt y Archer 2015; George 2006; Hacieminoglu 2016), se espera encontrar una mayor proporción de niñas que de niños con bajo interés tanto por la física como por la biología y un mayor número de niños que de niñas con un alto interés tanto por la física como por biología.

Justificación

En este apartado se adopta una triple-argumentación para justificar por qué los actuales patrones de matriculaciones en ciencias en las etapas secundarias y terciarias del sistema educativo resultan problemáticos; es decir, por qué resulta preocupante la baja presencia de mujeres en determinadas disciplinas STEM como la física o la ingeniería. Así, entre los muchos argumentos que subrayan la necesidad e importancia de aumentar el número de mujeres matriculadas en las disciplinas científicas en las que su presencia es minoritaria, en este trabajo se adopta como fundamentales los siguientes tres: (I) alfabetización científica integral e inclusiva, (II) necesidad de capital humano científicamente alfabetizado, y (III) diversidad en intereses y visiones científico-tecnológicas.

Alfabetización científica integral e inclusiva

Probablemente uno de los principales argumentos a favor de la inclusión de las niñas en todos los estudios y disciplinas STEM se basa en los derechos humanos (UNESCO 2017). Desde hace décadas, numerosos informes internacionales (AAAS 1993; EC 2013; NRC 2012; OECD y UNESCO 2003; Osborne y Dillon 2008), así como leyes educativas nacionales, como la Ley Orgánica para la Mejora de la Calidad Educativa (LOMCE 2013), vienen resaltando la importancia de alfabetizar científicamente a los estudiantes. La igualdad de género en las disciplinas STEM se postula como importante para garantizar que los niños y las niñas puedan adquirir los beneficios asociados con la alfabetización STEM (UNESCO 2017), algo que no sucederá a menos que se reduzca la gran segregación de género en determinadas disciplinas, como la física. Las escasas matriculaciones en estudios de física por parte de las mujeres resultan aún más preocupantes con la nueva actual ley educativa, en la que los estudiantes configuran su trayectoria académica desde el tercer curso de la etapa Educación Secundaria Obligatoria (ESO) y deben decidir si se matriculan en asignaturas propias de la rama de Ciencias o no. Si desde etapas elementales las niñas no desarrollan actitudes positivas e interés por la física, es muy probable que en Secundaria opten por asignaturas diferentes, por lo que la segregación se produce mucho antes en el sistema educativo, resultando más difícil aún despertar su interés por disciplinas de física en etapas posteriores.

Necesidad de capital humano

En la actualidad, la falta de mujeres en algunas disciplinas STEM supone una situación problemática tanto para la inclusión educativa como también para la consecución de las demandas del mercado laboral (Beede *et al.* 2011; EC 2004). La existencia de ciudadanos cualificados en disciplinas STEM es crucial para la capacidad innovadora y el desarrollo tecnológico de los países. En la actualidad, se estima que existe una ausencia de personal cualificado para cubrir los puestos laborales necesarios en STEM (Berkhout, Sattinger, Theeuwes y Volkerink 2012). La falta de personal cualificado combinado con el envejecimiento de la fuerza laboral en STEM suscita un panorama preocupante (Morrell y Parker 2015). Por lo tanto, el aprovechamiento del progreso científico-tecnológico requiere del mayor número posible de talentos, suponiendo un gran déficit la falta de inclusión de las mujeres en estas disciplinas (Blickenstaff 2005).

Diversidad en intereses y visiones científico-tecnológicas

La escasa representación de las niñas en las disciplinas STEM es problemática no solo por cuestiones de inclusión educativa o necesidad de capital humano, sino también debido a las limitaciones que esto supone al desarrollo de las innovaciones científicas. Dado que los hombres y las mujeres pueden albergar intereses y concepciones diferentes, las innovaciones científico-tecnológicas podrían verse obstaculizadas por la falta de diversidad de género en determinadas disciplinas científicas (Sikora y Pokropek 2012). Por ejemplo, en ingeniería, una disciplina predominantemente masculina, se definió la altura y el peso medio de la población estadounidense masculina como la norma y objeto principal de estudio en pruebas de choque (AGARD 1996). La falta de atención a otros segmentos de la población (por ejemplo, las mujeres) durante la fase de diseño de mecanismos de seguridad resultó en daños no intencionales dado que los cinturones de seguridad convencionales no se ajustan a las necesidades de las mujeres embarazadas, por lo que los accidentes automovilísticos son la principal causa de muerte fetal (Weiss, Songer y Fabio 2001). Una mayor presencia de mujeres en estas disciplinas hubiese dado lugar a una visión más holística durante la fase de diseño, siendo las perspectivas diferenciadas aportadas a la Ciencia por las mujeres necesarias para moldear las innovaciones tecnológicas y el propio conocimiento científico de manera significativa (Heybach y Pickup 2017). Las distintas experiencias e inquietudes de las mujeres pueden dar lugar al desarrollo de nuevas preguntas de investigación y perspectivas investigadoras, siendo numerosas las evidencias que manifiestan el impacto de las mujeres en el desarrollo de diferentes disciplinas científicas y que subrayan la necesidad de mayor inclusión de las mujeres en la Ciencia (Bug 2003; Conkey 2003; Schiebinger 2003). Sirva de ejemplo *Gendered Innovations* (Schiebinger *et al.* 2011), una iniciativa que proporciona numerosos estudios de casos que ilustran la forma en que las aportaciones de género en la ciencia, la medicina, la ingeniería y en investigación medioambiental condujo a nuevos descubrimientos e innovaciones.

Fundamentación teórica: actitudes hacia la Ciencia

La literatura sobre actitudes en la Enseñanza de las Ciencias se ve afectada por la falta de claridad y conceptualización del objeto de estudio, es decir, las actitudes (Osborne *et al.* 2003). Por tanto, la investigación en actitudes hacia la Ciencia se enfrenta a la ausencia de una clara conceptualización de lo que constituye las actitudes hacia la ciencia y también sobre los diferentes términos utilizados para referirse a las actitudes. En relación con la conceptualización de las actitudes, Klopfer (1971) ofreció una visión sobre el concepto objeto de estudio al categorizar las actitudes como un conjunto de comportamientos afectivos hacia la Ciencia, los científicos, la investigación científica, las carreras científicas y hacia las actividades relacionadas con la ciencia en general. Gardner (1975) definió las actitudes hacia la Ciencia como la evaluación favorable o desfavorable de la ciencia. Con respecto a los diferentes términos, Ramsden (1998) llegó a la conclusión de que términos como interés, opiniones, imágenes, creencias y valores se utilizan indistintamente y que sus significados a menudo se superponen. A pesar de los numerosos términos y concepciones sobre las actitudes, éstas son esencialmente una medida de las preferencias y sentimientos expresados hacia un objeto, en este caso, la Ciencia y lo científico (Osborne *et al.* 2003). No obstante, esta conceptualización es demasiado amplia y aún no se ha alcanzado un consenso acerca de qué aspectos constituyen las actitudes hacia la Ciencia, por lo que los estudios anteriores han incorporado muchos componentes como la autoestima, la motivación hacia la Ciencia o el disfrute de la Ciencia, entre muchos otros (Osborne *et al.* 2003).

Mientras que la actitud hacia la Ciencia fue un foco de investigación importante en Didáctica de las Ciencias, las teorías de la psicología social han influido en el estudio de las actitudes hacia la ciencia, por lo que en los últimos años el comportamiento, más que las actitudes en sí, se ha convertido en un foco de interés. Por lo tanto, muchos investigadores adoptaron modelos y teorías conductuales tales como la Teoría de la Conducta Planificada (Ajzen 1991) para comprender mejor la relación entre las actitudes y la conducta de los estudiantes. La teoría de Ajzen (1991) hace una clara distinción entre las actitudes hacia un objeto (es decir, actitudes hacia la ciencia) y las actitudes hacia una acción que debe realizarse con respecto al objeto (p. ej.: actitud hacia matricularse en una carrera en ciencias o actitud hacia participar en actividades científicas extracurriculares).

Por lo tanto, dado que las actitudes hacia la acción son mejores predictoras del comportamiento de las personas que las actitudes hacia el objeto mismo (Ajzen y Fishbein 1980), en este estudio nos enfocamos en examinar la actitud de los estudiantes hacia el estudio del contenido de la física o la biología (actitud hacia la acción a realizar sobre el objeto) en lugar de las actitudes de los estudiantes hacia el contenido de la física o la biología (actitud hacia el objeto). Este enfoque es consistente con modelos como el de la Teoría de Conducta Planeada (Ajzen 1991), y, por lo tanto, nos ayudará a examinar la interacción entre el género, el curso escolar y las actitudes hacia el estudio de la física y la biología.

Método

Participantes y contexto

Los datos de este estudio provienen de un extenso proyecto de investigación enfocado al desarrollo de un modelo de enseñanza STEM integrado para la etapa de Educación Primaria, por lo que la muestra está compuesta por 733 estudiantes (M_{edad} = 10.11, DE_{edad} = 1.18; 46.9% niñas) matriculados en el 3er (n = 65), 4º (n = 149), 5º (n = 232) y 6º (n = 287) curso de Educación Primaria. Los estudiantes provenían de un total de 29 Colegios de Educación Infantil y Primaria (CEIPs) situados en la provincia de Burgos (España), que han participado en un programa intensivo de enriquecimiento curricular de una semana de duración desarrollado en el Centro Rural de Innovación Educativa de Burgos (CRIEB).

Instrumento y procedimiento

Se ha empleado el instrumento *School Science Attitudes Survey* (SSAS) (Kennedy, Quinn y Taylor, 2016) en su versión en castellano (Toma y Meneses Villagrá *en prensa*) (Tabla 1). Se trata de un instrumento diseñado para medir las actitudes de los estudiantes de Educación Primaria hacia la asignatura de Ciencias de la Naturaleza en base a los seis constructos actitudinales (CA) más comunes en la literatura de actitudes hacia la Ciencia. Los CA (I) intenciones de matricularse en Ciencias de la Naturaleza, (II) disfrute de las clases de Ciencias de la Naturaleza, (III) autoeficacia en el aprendizaje de Ciencias de la Naturaleza, y (IV) dificultad percibida de las clases de Ciencias de la Naturaleza poseen una naturaleza unidimensional. Los dos constructos restantes, (V) utilidad y (VI) relevancia de las clases de Ciencias de la Naturaleza, poseen una naturaleza multidimensional. Más concretamente, el constructo 'utilidad de las clases de ciencias' está compuesto por dos dimensiones llamadas 'utilidad para las carreras científicas' y 'utilidad para la elección de otras carreras'. Por último, el constructo "relevancia de las clases de Ciencias de la Naturaleza" está formado a partir de las dimensiones 'relevancia para la sociedad' y 'relevancia para la vida personal'. Tanto los CA unidimensionales como las dimensiones subyacentes a los CA multidimensionales se miden a través de escalas de ítem único, bien de tipo Likert de cinco opciones (totalmente en desacuerdo– totalmente de acuerdo) o de diferencial semántico (p. ej.: divertido – aburrido).

Esta escala ha sido previamente validada con estudiantes españoles de Educación Primaria, mostrando resultados que indican que la versión española del SSAS comparte una varianza satisfactoria con otros instrumentos destinados a medir las actitudes de los estudiantes hacia la Ciencia y su aprendizaje, confirmando así la validez convergente del SSAS en castellano. Se trata de un instrumento de reducida extensión y rápida administración, postulándose en su versión en castellano como un instrumento adecuado para medir las actitudes de los estudiantes hacia las clases de Ciencias (Toma y Meneses Villagrá *en prensa*). En este estudio, la confiabilidad del instrumento se ha medido empleando dos pruebas de consistencia interna. Así, el Alfa de Cronbach ha sido .763, resultados que se encuentran por encima del criterio mínimo de .70 estipulado para estudios exploratorios (Nunnally y Bernstein 1994) y que están en consonancia con recomendaciones relacionadas con el uso de escalas de autoinforme con pocos ítems y para muestras con estudiantes de poca edad (Newman y McNeil 1998). Además, las correlaciones ítem-total oscilaron entre .313 y .552, resultados asimismo superiores al criterio mínimo .30 (Nunnally y Bernstein 1994). Por tanto, en base a ambas pruebas, se puede concluir que la confiabilidad del instrumento empleado es satisfactoria.

El cuestionario empleado fue administrado en presencia del primer autor y en ausencia de los tutores convencionales de aula a la llegada de los estudiantes al CRIEB tras informar del carácter voluntario y anónimo de su participación.

Tabla 1. Constructos e ítems de la escala SSAS

Constructo	Ítems
(I) Intenciones	1. Es muy probable que me apunte a asignaturas de CN al terminar Educación Primaria (LT)
(II) Disfrute	2. Pienso que CN es: (SD aburrida – divertida)
(III) Dificultad	3. Me cuesta terminar las tareas para la clase de CN (LT)
(IV) Autoeficacia	4. Pienso que soy muy bueno en CN (LT)
(V) Utilidad	
Utilidad para estudios de ciencias	5. Un trabajo como científico sería interesante (LT)
Utilidad para estudios personales	6. Para mis futuros estudios, el conocimiento de las clases de CN será: (SD inútil – útil)
(VI) Relevancia	
Relevancia social	7. La ciencia ayuda a mejorar la vida (LT)
Relevancia personal	8. Quiero aprender sobre las plantas de mi entorno (LT)
	9. Para mi vida cotidiana, creo que CN es (SD poco importante – muy importante)
	10. Quiero aprender sobre la electricidad y saber cómo se usa en una casa (LT)

Notas: CN: Ciencias de la Naturaleza; LT: ítems de tipo Likert de 5 opciones (totalmente en desacuerdo – totalmente de acuerdo); SD: ítems de tipo diferencial semántico de 5 opciones de respuesta.

Análisis de datos

El interés por contenidos de biología y de física se ha determinado a partir del análisis de los ítems *8. Quiero aprender sobre las plantas de mi entorno* (biología) y *10. Quiero aprender sobre la electricidad y saber cómo se usa en una casa* (física) de la dimensión 'relevancia para la vida personal'.

En relación con la primera pregunta de investigación, se empleó el estadístico *t* de Student para muestras independiente para comparar el interés de las niñas y los niños por contenidos de física y de biología y varias ANOVAs univariantes para comparar el interés por contenidos de física y de biología en función de la edad. Se empleó una corrección de Bonferroni estableciéndose un valor de significancia más estricto ($p < .01$) y así evitar errores de tipo I.

Para la segunda pregunta de investigación, se empleó un método de análisis cuantitativo basado en técnicas de agrupamiento llamado clúster o de conglomerados. El análisis clúster es una técnica analítica centrada en sujetos que proporciona una manera de examinar cómo se combinan las variables dentro de los individuos (Everitt, Landau, Leese y Stahl 2011). Por lo tanto, el análisis clúster tiene por objeto la clasificación de los sujetos en diferentes grupos que posteriormente son analizados en búsqueda de características similares y distintivas que diferencia a cada subgrupo. En la literatura, se ha comprobado la utilidad del análisis clúster en su aplicación para la identificación de subgrupos significativos en contextos académicos (Battaglia, Di Paola y Fazio 2017; Fazio, Battaglia y Di Paola 2013; Paciello, Ghezzi, Tramontano, Barbaranelli y Fida 2016). Las técnicas de análisis clúster pueden ser divididas en dos grupos: jerárquico y no jerárquico (Everitt *et al.* 2011). En este estudio, se empleó el análisis clúster bietápico con medida de distancia logarítmica de verosimilitud y criterio de información bayesiano (BIC) (Schwarz 1978) para la determinación del número de clústeres a extraer. Posteriormente se empleó un método de clúster no jerárquico, llamado K-Medias para la extracción de los conglomerados que han determinado los perfiles de interés por contenidos de física o de biología de cada estudiante. Asimismo, se realizaron dos pruebas de independencia estadística Chi-cuadrado para examinar si los estudiantes estaban uniformemente representados en los perfiles en función del nivel escolar y el sexo y así determinar si las preferencias de las niñas y de los niños de Educación Primaria simulan patrones de matriculaciones de etapas secundarias y terciarias.

Para la tercera y última pregunta de investigación, se efectuaron una serie de pruebas de MANOVA, ANOVA y *t* de Student con corrección de Bonferroni para analizar las diferencias entre los cuatro perfiles de interés en términos de actitudes hacia la asignatura Ciencias de la Naturaleza. Más específicamente, se examinaron las diferencias entre cada subgrupo en términos de cinco constructos actitudinales (CA): (I) intenciones de matricularse en ciencias, (II) disfrute durante las clases de Ciencias de la Naturaleza, (III) dificultad de las clases de Ciencias de la Naturaleza, (IV) percepción de autoeficacia en Ciencias de la Naturaleza, y (V) utilidad percibida de las clases de Ciencias de la Naturaleza.

Resultados

P1. ¿Cómo afecta el sexo y el curso escolar a las preferencias por estudiar contenidos científicos de física y de biología?

La prueba *t* para muestras independientes señaló que no hay diferencias significativas en el interés por contenidos de biología entre niñas (M = 3.23, SD = 1.12) y niños (M = 3.36, SD = 1.21). Sin embargo, los niños reportaron un interés significativamente mayor por contenidos de física que las niñas, t (731) = -5.33, p < .001. Estos resultados aparecen en la figura 1.

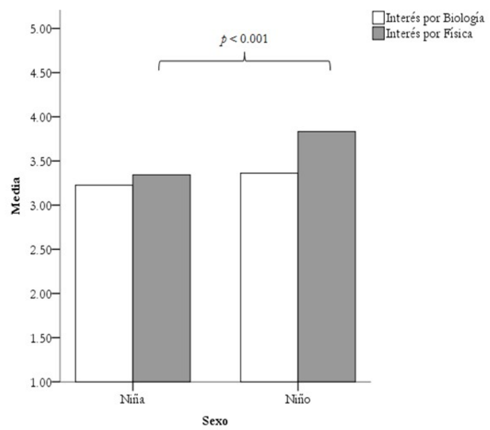


Figura 1. Comparaciones según sexo para interés en biología o física

Al controlar la variable sexo, el ANOVA univariante indicó que las niñas matriculadas en el 6º curso de Educación Primaria poseen un interés significativamente menor por el contenido de biología en comparación con las niñas matriculadas en el tercer curso, $F(3, 340) = 3.9$, $p = .009$. También se ha identificado un descenso en el interés de las niñas por contenidos de física a medida que aumenta el curso escolar, sin embargo, los resultados no alcanzaron significancia estadística. En cuanto a los niños, se ha identificado que a medida que aumenta el curso escolar, su interés por contenidos de biología disminuye y su interés por contenidos de física aumenta, aunque ninguno de los dos resultados es significativo. Estos resultados aparecen en la figura 2.

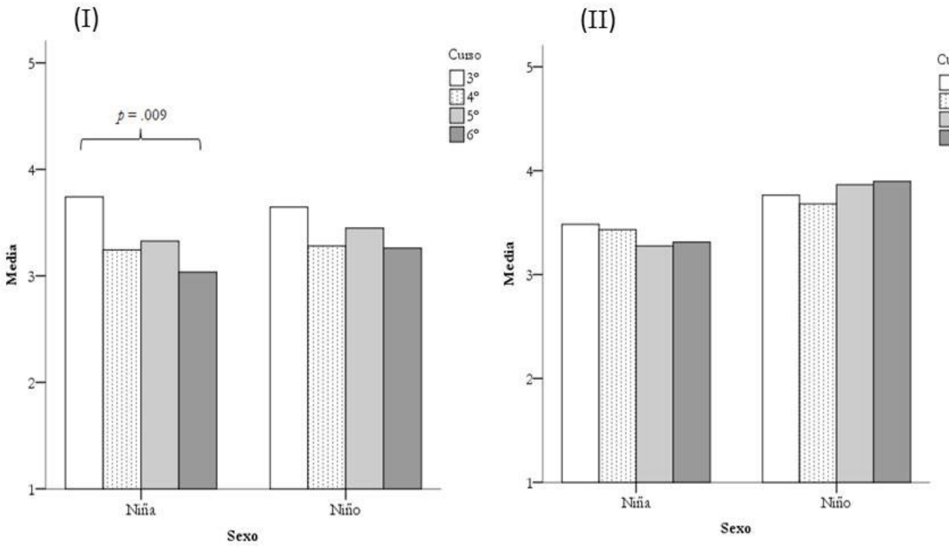


Figura 2. Comparaciones según nivel escolar para interés en (I) biología y (II) física

P2. Las preferencias, ¿simulan patrones de matriculaciones identificadas en etapas secundarias y terciarias?

La prueba clúster bietápica reveló la existencia de siete clústeres con medida Silhouette de cohesión y separación adecuada (.5). No obstante, dado que los clústeres formados con pocos estudiantes pueden resultar poco relevantes desde un punto de vista pedagógico (Battaglia *et al.* 2017), se realizaron sucesivos análisis clúster de tipo bietápico especificando un menor número de clústeres a extraer, obteniéndose que una solución de cuatro subgrupos posee la medida de Silhouete más adecuada (.6), y una estructura de conglomerados razonable (Struyf, Hubert y Rousseeuw 1997). Determinado el número de subgrupos a extraer, se efectuó un análisis K-Medias para extraer perfiles en función del interés por contenidos de física o de biología.

El análisis clúster K-Medias reveló una solución de cuatro clústeres (iteración = 3). El MANOVA reveló diferencias significativas entre los cuatro clústeres en ambos ítems, Wilks Lambda = .100, F (6, 1456) = 525.362, $p < .001$, $\eta^2 = .684$. Dos pruebas ANOVA mostraron que el efecto principal de la pertenencia a los subgrupos fue igualmente significativo tanto por la preferencia de contenidos de biología F (3, 729) = 401.8, $p < .001$, como por física F (3, 729) = 746.9, $p < .001$. Las comparaciones post-hoc de Bonferroni revelaron la existencia los cuatro siguientes perfiles (Figura 3):

i. «Interés por física», con 72 estudiantes (M_{edad} = 10.28; DE_{edad} = 1.10) que mostraron un interés significativamente mayor por contenidos de física que por contenidos de biología.

ii. «Interés por biología», con 221 estudiantes (M_{edad} = 10.02; DE_{edad} = 1.20) con un interés por contenidos de biología significativamente mayor que por los propios de física.

iii. «Bajo interés general», con 85 estudiantes (M_{edad} = 10.29; DE_{edad} = 1.10) que reportaron un bajo interés tanto por contenidos de física como de biología.

iv. «Alto interés general», con 355 estudiantes (M_{edad} = 10.10; DE_{edad} = 1.21) que mostraron un alto interés por ambos contenidos de física y de biología.

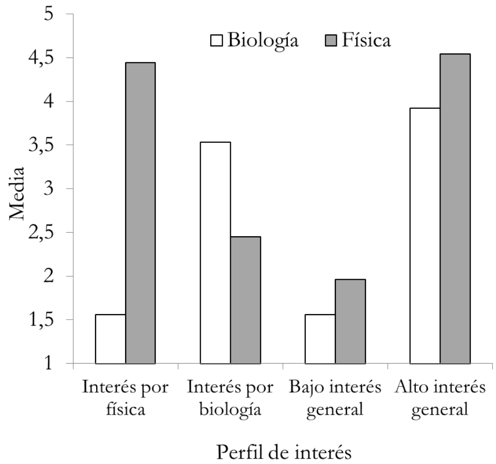


Figura 3. Perfiles según el interés por contenidos de física o de biología

En relación con el nivel escolar, la prueba de independencia estadística Chi-cuadrado reveló que los participantes estaban uniformemente representados en los cuatro perfiles, χ^2 (9, *n* = 733) = 5.125, *p* = .823, phi = .82. Sin embargo, con respecto al sexo de los estudiantes, se encontraron diferencias significativas en la distribución de estos, χ^2 (3, *n* = 733) = 20.843, *p* < .001, phi = .17. Así, se encontró una significativa sobrerrepresentación de las niñas (126 de 221, 57%) con respecto a los niños (95 de 221, 43%) en el perfil «interés por biología», y una significativa infrarrepresentación de las niñas (28 de 72, 38.9%) en el perfil «interés por física» con respecto de los niños (44 de 72, 61.1%). Además, hubo significativamente más niñas (48 de 85, 56.5%) que niños (37 de 85, 43.5%) agrupadas en el perfil «bajo interés general», y significativamente menos niñas (142 de 355, 40%) que niños (213 de 355, 60%) en el perfil «alto interés general». La figura 4 muestra la distribución de los estudiantes según sexo en cada perfil de preferencia.

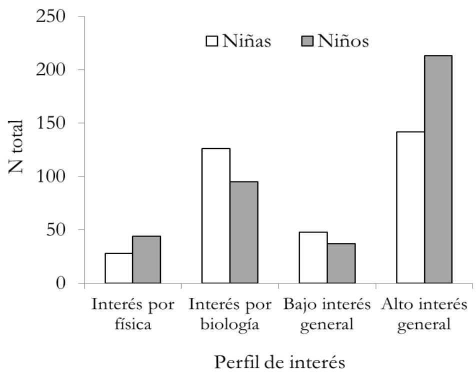


Figura 4. Distribución de los estudiantes en los perfiles derivados según sexo

P3. ¿Existen diferencias en las actitudes de los estudiantes hacia la asignatura de Ciencias de la Naturaleza en función de su preferencia por contenidos de física o de biología?

El MANOVA reveló diferencias significativas en los CA entre los cuatro perfiles de interés, Wilks Lambda = .837, F (1^2, 1921) = 11.142, p < .001, η^2 = .058. Cinco ANOVAs univariantes con corrección de Bonferroni revelaron diferencias significativas para cuatro de los cinco CA: «intención de matricularse en Ciencias de la Naturaleza» F (3, 729) = 29.829, $p < .001$, η^2 = .62, «disfrute durante las clases de Ciencias de la Naturaleza» F (3, 729) = 30.768, $p < .001$, η^2 = .62, «percepción de autoeficacia en Ciencias de la Naturaleza» F (3, 729) = 7.702, $p < .001$, η^2 = .75, y «utilidad de las clases de Ciencias de la Naturaleza» F (3, 729) = 145.204, $p < .001$, η^2 = .16.

Las comparaciones post-hoc de Bonferroni reportaron que los estudiantes del grupo 'alto interés general' son los que poseen significativamente más intenciones de seguir estudios de ciencias y los que han manifestado significativamente mayor disfrute durante las clases de ciencias, especialmente en comparación con los estudiantes del perfil «bajo interés general» e «interés en física». Asimismo, los estudiantes del perfil «alto interés general» mostraron significativamente mayor autoeficacia en ciencias que los otros tres grupos, siendo los estudiantes con 'bajo interés general' los que reportaron menor autoeficacia de todos los grupos. Por último, los estudiantes del perfil «alto interés general» asimismo concibieron las clases de Ciencias de la Naturaleza como significativamente más útiles que los demás perfiles. Estos resultados se recogen en la figura 5.

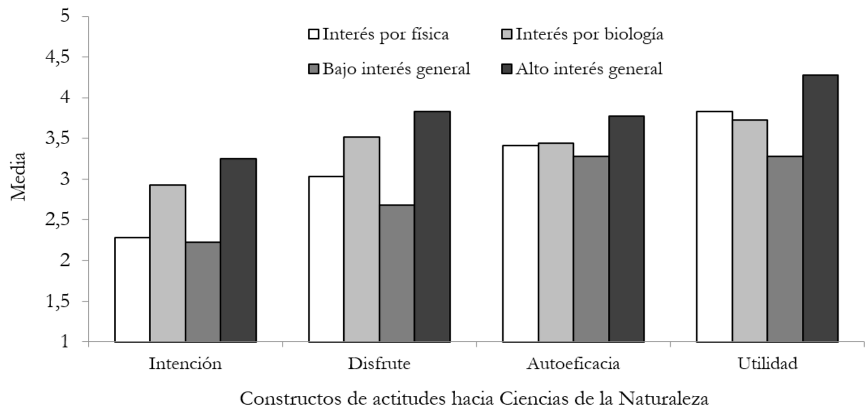


Figura 5. Comparaciones entre actitudes hacia Ciencias de la Naturaleza según el interés por contenidos de biología o de física

Interpretación y discusión

En conjunto, los resultados de este estudio confirman las premisas de inicio. Así, se han identificado al menos cuatro aspectos que resultan de gran preocupación para la educación científica en España.

En primer lugar, se ha identificado un perfil actitudinal caracterizado por un alto interés por contenidos de biología en el que predominaban las niñas y un perfil que se caracteriza por un alto interés por física, representado mayoritariamente por niños. Estos resultados resultan preocupantes porque evidencian que ya desde etapas elementales del sistema educativo se pueden identificar patrones de interés en ciencias muy similares a los registrados por los datos nacionales de matriculación en estudios universitarios (MECD 2016). De este modo, las niñas tienen más interés por contenidos de biología y menos por los propios de física, sucediendo lo contrario en el caso de los niños, siendo ésta una diferencia que se plasma en las estadísticas de matriculaciones en estudios terciarios pero que no ha sido previamente estudiada a nivel de Educación Primaria.

En segundo lugar, se ha observado que las niñas son mayoría en un perfil caracterizado por un bajo interés general, y son minoría en el perfil con un alto interés tanto por contenidos de biología como de física. Estos datos son asimismo preocupantes ya que coinciden con los resultados registrados en Educación Secundaria a nivel nacional (de Pro y Pérez 2014; Vázquez y Manassero 2008) e internacional (Caleon y Subramaniam 2008; Denessen *et al.* 2015; DeWitt y Archer 2015; George 2006; Hacieminoglu 2016) y que muestran una tendencia menor de las niñas por la Ciencia que los niños. Si bien los estudios mencionados han sido desarrollados principalmente con estudiantes de Educación Secundaria, el presente estudio evidencia que este patrón comienza desde la etapa de Educación Primaria.

En tercer lugar, se ha identificado un descenso en el interés de las niñas tanto por contenidos de física como de biología a medida que aumenta el nivel escolar, lo que coincide con estudios anteriores que señalan que las niñas pierden más rápidamente el interés por la Ciencia en cursos superiores en comparación con los niños (Marginson, Tytler, Freeman y Roberts 2013). Estos datos resultan preocupantes debido a que las aspiraciones profesionales podrían estar mayormente formadas a la edad de 13 años, especialmente en las niñas (Lindahl 2007), siendo cada vez más difícil involucrarlas en disciplinas científicas en niveles posteriores.

Por último, al examinar las actitudes de los estudiantes hacia la asignatura de Ciencias de la Naturaleza en función del perfil de interés, se ha observado que aquellos estudiantes con mejores actitudes poseen asimismo un alto interés general tanto por contenidos de física como de biología. La mayoría de las niñas de esta muestra han sido categorizadas dentro del perfil «alto interés general», seguido por «interés en biología», por lo que se podría concluir que en Educación Primaria las actitudes aún son positivas, aunque ya se advierten diferencias de género en cuanto al interés y las preferencias y un descenso en las actitudes, especialmente en las intenciones de seguir estudios de ciencias, que ha sido el constructo actitudinal peor valorado, seguido por el disfrute de las clases de ciencias, resultados coincidentes con investigaciones previas (Denessen *et al.* 2015).

Implicaciones educativas y de investigación

Los resultados de este estudio subrayan la necesidad de diseñar y desarrollar intervenciones educativas desde edades tempranas para despertar el interés y retener a las niñas en la Ciencia, especialmente en aquellas disciplinas en las que se encuentran más infrarrepresentadas, como la física. Estas intervenciones deberían realizarse antes de la etapa de Educación Secundaria, tal y como se viene resaltando en la literatura (Marbá-Tallada y Márquez Bargalló 2010) y además debido a que, tal y como muestran los resultados de este estudio, el interés por la física y la biología parecería forjarse desde la Educación Primaria. Las intervenciones educativas deben abordar al menos dos aspectos esenciales: el uso de metodologías activas para la enseñanza y el aprendizaje de las ciencias, y la reducción de sesgo de género en el rendimiento académico en ciencias.

En cuanto al uso de metodologías activas, la indagación escolar o el aprendizaje basado en problemas han mostrado ser efectivas en aumentar la participación de los estudiantes durante las clases y en mejorar la percepción de utilidad de las ciencias y el disfrute durante las clases de ciencias (Demirel y Dagyar 2016; Romero-Ariza 2017; Schroeder, Scott, Tolson, Huang y Lee 2007; Toma y Greca 2018). Dado que las diferencias entre los niños y las niñas se atribuye a factores de aversión al riesgo y a la competitividad, favoreciendo a los niños (Booth, Cardona-Sosac y Nolen, 2014; Niederle y Vesterlund 2010) y a que las niñas tienden a subestimar sus habilidades en la elección de trayectorias profesionales (Rapoport y Thibout 2018), las metodologías de enseñanza y aprendizaje activas podrían ser efectivas para reducir el nivel de competitividad durante las clases de ciencias estructurando grupos de trabajos que fomenten la participación colaborativa y la búsqueda conjunta de perspectivas y de resolución de problemas (Gillies 2003).

En relación con la existencia de sesgos de género, parecería que las expectativas de los padres y de los maestros en cuanto a las competencias de los niños y las niñas posee sesgos de género y podrían influir negativamente en las actitudes y el rendimiento de las niñas, disminuyendo su autoconcepto (Gunderson *et al.* 2012). Además, se ha comprobado que los maestros evalúan las capacidades de las niñas como inferiores a las de los niños (Dickhauser y Meyer 2006). Si bien estos resultados están más relacionados con las matemáticas, existen evidencias que indican la existencia de asociaciones estereotipadas que vinculan a los hombres con carreras profesionales en ciencia y a las mujeres con las artes y la familia (Nosek, Banaji y Greenwald 2002), así como a las mujeres menos competentes en ciencias que los hombres (Eagly y Mladinic 1993).

Por último, en este estudio se ha identificado que el interés por contenidos de física o de biología comienza a establecerse a partir de la etapa de Educación Primaria. Futuros estudios deberían abordar estos resultados para examinar qué factores intervienen para que desde una edad tan temprana pueda haber una preferencia tan marcada y diferenciada según el género de los estudiantes.

Conclusiones

La presente investigación busca entender mejor a partir de qué edad se produce una segregación horizontal de las mujeres en las disciplinas científicas mediante el análisis de las preferencias de 344 niñas y 389 niños de Educación Primaria por contenidos de física o de biología. Además, se ha examinado las actitudes hacia las clases de Ciencias de la Naturaleza en función de las preferencias por unos u otros contenidos. El análisis de conglomerados reveló la existencia de un perfil actitudinal sesgado sexualmente, caracterizado por un alto interés por el contenido de las ciencias biológicas y un bajo interés por el aprendizaje de la física por parte de las niñas, y resultados opuestos en el caso de los niños. Además, en comparación con los niños, las niñas han sido mayoría en los perfiles actitudinales caracterizados por un bajo interés por la física y la biología, y han sido minoría en los perfiles caracterizados por un alto interés por ambas disciplinas.

El análisis de los constructos actitudinales de los miembros pertenecientes a los diferentes perfiles ilustraron que el grupo de los estudiantes con un alto interés general por los contenidos de física y de biología reportaron mejores actitudes hacia la ciencia escolar que el resto de los grupos, lo que sitúa a las niñas con este tipo de perfil en riesgo de exclusión educativa en ciencias, ya desde la Educación Primaria.

En conjunto, estos resultados señalan que las trayectorias académicas y profesionales de los estudiantes, así como su interés por disciplinas científicas de física o de biología, se empiezan a conformar desde niveles elementales del sistema educativo, suceso que podría dar lugar al posterior descenso de mujeres matriculadas en carreras universitarias de ciencias. La importancia de estos resultados reside en subrayar la necesidad de desarrollar intervenciones educativas desde la etapa de Educación Primaria para reducir las diferencias de género detectadas sobre actitudes hacia la ciencia escolar e interés por disciplinas científicas.

Agradecimientos

Este estudio fue financiado por la edición 2017-2021 del programa de contratos de investigación predoctorales de la Universidad de Burgos y parcialmente financiado por el proyecto MINECO-AEI EDU 2017-89405-R.

Extendemos nuestro agradecimiento de manera especial al Centro Rural de Innovación Educativa de Burgos (CRIEB) y a todos los centros educativos que han formado parte de este estudio.

Referencias

AAAS. (1993) *Benchmarks for scientific literacy*. Washington: American Association for the Advancement of Science.

AGARD. (1996) *Advisory Report 330: Anthropomorphic dummies for crash and escape system testing*. Quebec: Canada Communication Group.

Ajzen, I. (1991) The theory of planned behavior. *Organizational Behavior and Human Decision Processes*, *50*(2), 179–211.

Ajzen, I., Fishbein, M. (1980) *Understanding attitudes and predicting social behavior*. New Jersey: Prentice-Hill.

Battaglia, O. R., Di Paola, B., Fazio, C. (2017) K-means clustering to study how student reasoning lines can be modified by a learning activity based on Feynman's unifying approach. *Eurasia Journal of Mathematics, Science and Technology Education*, 13(6), 2005–2038. doi:10.12973/eurasia.2017.01211a

Beede, D. N., Julian, T. A., Langdon, D., McKittrick, G., Khan, B., Doms, M. E. (2011) Women in STEM: A gender gap to innovation. *Economics and Statistics Administration*, 4–11. doi:10.2139/ssrn.1964782

Berkhout, E., Sattinger, M., Theeuwes, J., Volkerink, M. (2012) *Into the gap. Exploring skills and mismatches.* Recuperado de: https://research.randstad.es/wp-content/uploads/2015/12/into-the-gap.pdf

Blickenstaff, J. C. (2005) Women and science careers: Leaky pipeline or gender filter? *Gender and Education*, 17(4), 369–386. doi:10.1080/09540250500145072

Booth, A., Cardona-Sosac, L., Nolen, P. (2014) Gender differences in risk aversion: Do single-sex environments affect their development. *Journal of Economic Behavior and Organization*, 99, 126–154.

Bug, A. (2003) Has feminism changed physics. *Signs*, 28(3), 881–899. doi:0097-9740/2003/2803-0008

Buser, T., Niederle, M., Oosterbeek, H. (2014) Gender, competitiveness, and career choices. *The Quarterly Journal of Economics*, 129, 1409–1447.

Caleon, I. S., Subramaniam, R. (2008) Attitudes towards science of intellectually gifted and mainstream upper primary students in Singapore. *Journal of Research in Science Teaching*, 45(8), 940–954. doi:10.1002/tea.20250

Collins, C., Kenway, K., McLeod, J. (2000) Factors influencing the educational performance of males and females in school and their initial destinations after leaving school. Recuperado de http://siandvasupport.sydneyinstitute.wikispaces.net/file/view/Factors+influsencing+the+performance+of+males+ and+females+in+school.pdf

Conkey, M. W. (2003) Has feminism changed archaeology? *Signs*, 28(3), 867–880.

Dasgupta, N. (2011) Ingroup experts and peers as social vaccines who inoculate the self-concept: The stereotype inoculation model. *Psychological Inquiry*, 22(3), 231–246.

de Pro Bueno, A., Pérez Manzano, A. (2014) Actitudes de los alumnos de Primaria y Secundaria ante la visión dicotómica de la Ciencia. *Enseñanza de Las Ciencias*, 32(3), 111–132. doi:10.5565/rev/ensciencias.1015

Deloitte. (2016) *Women in STEM. Technology, career pathways and the gender pay gap*. London: Deloitte.

Demirel, M., Dagyar, M. (2016) Effects of problem-based learning on attitude: A meta-analysis study. *Eurasia Journal of Mathematics, Science and Technology Education*, 12(8), 2115–2137. doi:10.12973/eurasia.2016.1293a

Denessen, E., Vos, N., Hasselman, F., Louws, M. (2015) The relationship between primary school teacher and student attitudes towards Science and Technology. *Education Research International*, 1–7. doi:10.1155/2015/534690

DeWitt, J., Archer, L. (2015) Who aspires to a science career? A comparison of survey responses from primary and secondary school students. *International Journal of Science Education*, 37(13), 2170–2192. doi:09500693.2015.1071899

Dickhauser, O., Meyer, U. (2006) Gender difference in young children's math ability attributions. *Psychological Science*, *48*(1), 3–16.

Eagly, A. H., Mladinic, A. (1993) Are people prejudiced against women? Some answers from research on attitudes, gender stereotypes, and judgments of competence. *European Review of Social Psychology*, 5(1), 1–35.

EC. (2004) *Europe needs more scientists: Report by the high-level group on increasing human resources for Science and Technology*. Brussels: European Commission.

EC. (2013) *Special eurobarometer 401: Responsible Research and Innovation (RRI), Science and Technology*. Brussels: European Commission.

Everitt, B. S., Landau, S., Leese, M., Stahl, D. (2011) *Cluster Analysis*. Chichester: John Wiley & Sons, Ltd.

Fazio, C., Battaglia, O. R., Di Paola, B. (2013) Investigating the quality of mental models deployed by undergraduate engineering students in creating explanations: the case of thermally activated phenomena. *Physical Review Physics Education Research*, 9(2), 1–21.

Gardner, P. L. (1975) Attitudes to science: A Review. *Studies in Science Education*, *2*, 1–41.

George, R. (2006) A cross-domain analysis of change in students' attitudes toward science and attitudes about the utility of science. *International Journal of Science Education*, 28(6), 571–589. doi:10.1080/09500690500338755

Gillies, R. M. (2003) Structuring cooperative group work in classrooms. *International Journal of Educational Research*, 39(1–2), 35–49. doi:10.1016/S0883-0355(03)00072-7

Gunderson, E. A., Ramirez, G., Levine, S. C., Beilock, S. L. (2012) The role of parents and teachers in the development of gender-related math attitudes. *Sex Roles*, 66(3–4), 153–166. doi:10.1007/s11199-011-9996-2

Hacieminoglu, E. (2016) Elementary school students' attitude toward science and related variables. *International Journal of Environmental and Science Education*, 11(2), 35–52. doi:10.12973/ijese.2016.288a

Heybach, J., Pickup, A. (2017) Whose STEM? Disrupting the gender crisis within STEM. *Educational Studies*, 53(6), 614–627. doi:10.1080/00131946.2017.1369085

INE. (2017) Mujeres y hombres en España. Educación. Recuperado de https://goo.gl/RC2UKd

Kennedy, J., Quinn, F., Taylor, N. (2016) The school science attitude survey: A new instrument for measuring attitudes towards school science. *International Journal of Research y Method in Education*, 39(4), 422–445. doi:10.1080/1743727X.2016.1160046

Lindahl, B. (2007) Longitudinal study of students' attitudes towards science and choice of career. En proceedings de *80th session of the International Conference of the National Association for Research in Science Teaching*. New Orleans.

Ley Orgánica para la Mejora de la Calidad Educativa (LOMCE) (Ley Orgánica 8/2013, 9 de diciembre). Boletín Oficial del Estado, núm. 295, 2013, 10 de diciembre. Referencia: BOE-A-2013-12886.

Marbá-Tallada, A., Márquez Bargalló, C. (2010) ¿Qué opinan los estudiantes de las clases de ciencias? Un estudio transversal de sexto de primaria a cuarto de ESO. *Enseñanza de las Ciencias*, *28*(1), 19–30.

Marginson, S., Tytler, R., Freeman, B., Roberts, K. (2013) *STEM: Country comparisons. International comparisons of science, technology, engineering and mathematics (STEM) education*. Melbourne: Australian Council of Learned Academies.

MECD. (2016) *Datos y cifras del sistema universitario español. Curso 2015/2016*. Secretaría General Técnica.

Morrell, C., Parker. C. (2015) *Solving the education equation: A new model for improving STEM workforce outcomes through academic equity*. Recuperado de: www.napequity.org/solving-education-equation

Newman, I., McNeil, K. (1998) *Conducting survey research in the social sciences*. New York: University Press of America.

Niederle, M., Vesterlund, L. (2010) Explaining the gender gap in math test scores: The role of competition. *Journal of Economic Perspectives,* 24(2), 129–144. doi:10.1257/jep.24.2.129

Nosek, B. A., Banaji, M. R., Greenwald, A. G. (2002) Harvesting implicit group attitudes and beliefs from a demonstration website. *Group Dynamics: Theory, Research, and Practice*, 6(1), 101–115.

NRC. (2012) *A framework for K-12 science education: Practices, crosscutting concepts, and core ideas*. Washington: The National Academies Press.

NSF. (2013) *Women, minorities, and persons with disabilities in Science and Engineering*. Washington: National Science Foundation.

Nunnally, J. C., Bernstein, I. H. (1994) *Psychometric theory*. New York: McGraw-Hill.

OECD. (2013) *Education at a glance 2013: OECD indicators*. Paris: OECD Publishing.

OECD., UNESCO. (2003) *Literacy skills for the world of tomorrow – Further results from PISA 2000*. Paris: OECD.

Osborne, J., Dillon, J. (2008) *Science education in Europa. Critical reflections*. London: Nuffield Foundation.

Osborne, J., Simon, S., Collins, S. (2003) Attitudes towards science: A review of the literature and its implications. *International Journal of Science Education*, 25(9), 1049–1079. doi:10.1080/0950069032000032199

Paciello, M., Ghezzi, V., Tramontano, C., Barbaranelli, C., Fida, R. (2016) Self-efficacy configurations and wellbeing in the academic context: A person-centered approach. *Personality and Individual Differences*, 99, 16–21. doi:10.1016/j.paid.2016.04.083

Ramsden, J. M. (1998) Mission impossible? Can anything be done about attitudes to science? *Interntional Journal of Science Education*, *20*, 125–137.

Rapoport, B., Thibout, C. (2018) Why do boys and girls make different educational choices? The influence of expected earnings and test scores. *Economics of Education Review*, 62, 205–229. doi:10.1016/j.econedurev.2017.09.006

Romero-Ariza, M. (2017) El aprendizaje por indagación: ¿existen suficientes evidencias sobres sus beneficios en la enseñanza de las ciencias? *Revista Eureka sobre Enseñanza y Divulgación de las Ciencias*, 14(2), 286–299.

Schiebinger, L. (2003) Introduction: Feminism inside the sciences. *Signs*, 28(3), 859–866.

Schiebinger, L., Klinge, I., Sánchez de Madariaga, I., Paik, H. Y., Schraudner, M., Stefanick, M. (2011) *Gendered innovations in science, Health y medicine, engineering and environment*. Recuperado de http://genderedinnovations.stanford.edu/index.html

Schroeder, C. M., Scott, T. P., Tolson, H., Huang, T.-Y., Lee, Y.-H. (2007) A Meta-Analysis of national research: Effects of teaching strategies on student achievement in science in the United States. *Journal of Research in Science Teaching*, 44(10), 1436–1460. doi:10.1002/tea

Schwarz, G. E. (1978) Estimating the dimension of a model. *Annals of Statistics*, 6(2), 461–464.

Sikora, J., Pokropek, A. (2012) Gender segregation of adolescent science career plans in 50 countries. *Science Education*, 96(2), 234–264. doi:10.1002/sce.20479

Struyf, A., Hubert, M., Rousseeuw, P. J. (1997) Integrating robust clustering techniques in S-PLUS. *Computational Statistics & Data Analysis*, 26, 17–37. doi:10.1016/S0167-9473(97)00020-0

Toma, R. B., Greca, I. M. (2018) The effect of integrative STEM instruction on elementary students' attitudes toward Science. *EURASIA Journal of Mathematics, Science and Technology Education,* 14(4), 1383–1395. doi:10.29333/ejmste/83676

Toma, R. B., Meneses Villagrá, J. A. (*en prensa*) Validation of the Spanish single-item School Science Attitude Survey (S-SSAS) for elementary education. *PLOS ONE.*

UNESCO. (2017) *Cracking the code: Girls' and women's education in science, technology, engineering and mathematics (STEM)*. París: United Nations Educational, Scientific and Cultural Organization. Recuperado de: http://unesdoc.unesco.org/images/0025/002534/253479e.pdf

Vázquez, A., Manassero, M. A. (2008) El declive de las actitudes hacia la ciencia de los estudiantes: un indicador inquietante para la educación científica. *Revista Eureka sobre Enseñanza y Divulgación de las Ciencias*, 5(3), 274–292.

Weiss, H., Songer, T., Fabio, A. (2001) Fetal deaths related to maternal injury. *Journal of the American Medical Association*, 15, 1863–1868.

Supplemental material: Preliminary studies

Supplemental material 1: Draw a Scientist Test study

This supplemental material includes the full-text manuscript of first preliminary study about stereotypical image of scientists:

Toma, R. B., Greca, I. M., & Orozco Gómez, M. L. (2018). Una revisión del protocolo Draw-a-Scientist-Test (DAST). *Revista Eureka sobre Enseñanza y Divulgación de las Ciencias, 15*(3), 3104. doi: 10.25267/Rev_Eureka_ensen_divulg_cienc.2018.v15.i3.3104

Una revisión del protocolo Draw-a-Scientist-Test (DAST)

Radu Bogdan Toma[1a], Ileana María Greca[1b], Martha Lucía Orozco Gómez[2c]

[1]*Departamento de Didácticas Específicas, Área de Didácticas Experimentales. Universidad de Burgos, España.*
[2]*Departamento de Ciencias de la Educación, Área de Didáctica y Organización Escolar. Universidad de Burgos, España.*
[a]*rbtoma@ubu.es,* [b]*imgreca@ubu.es,* [c]*mlorozco@ubu.es*

[Recibido: 23 Noviembre 2017. Revisado: 7 Abril 2018. Aceptado: 9 Junio 2018]

Resumen: Durante décadas, numerosas investigaciones han identificado la existencia de imágenes estereotipadas sobre los científicos mediante el empleo del método *Draw-a-Scientist-Test* (DAST). No obstante, las limitaciones metodológicas asociadas al DAST cuestionan los resultados obtenidos y resaltan la necesidad de desarrollar nuevos protocolos para evaluar las concepciones de los estudiantes acerca de los científicos. En este trabajo, tras revisar las deficiencias metodológicas del DAST, se proponen algunas modificaciones al protocolo original probadas en una prueba piloto con 149 estudiantes de 2º a 6º curso de Educación Primer de diversidad cultural. Los resultados iniciales muestran que las modificaciones realizadas ayudan a reducir algunos de los problemas metodológicos identificados en el uso del DAST, permitiendo identificar e interpretar de manera más fiable las concepciones de los estudiantes sobre los científicos.

Palabras clave: imagen de científico, DAST, estereotipos sobre ciencia, educación primaria, diversidad cultural

A revision of the Draw-a-Scientist-Test (DAST)

Abstract: For decades, studies have shown the existence of a stereotypical image of scientist between children among the world using the *Draw-a-Scientist-Test* (DAST) method. However, the methodological limitations associated with the DAST question the results obtained and highlight the need to develop new protocols to evaluate students' conceptions about scientists. In this paper the methodological deficiencies of the DAST instrument are reviewed and some changed made to the original protocol are described. The pilot test carried out with 149 students of cultural diversity enrolled in 2nd to 6th grades of Elementary School shows that this modification allows a more reliable identification and analysis of students' conceptions, reducing some of the methodological problems identified in the use of DAST.

Keywords: image of scientist, Draw-a-Scientist-Test, science stereotypes, primary school, cultural diversity

Para citar este artículo: Toma, R. B., Greca, I. M., Orozco Gómez, M. L. (2018) Una revisión del protocolo Draw-a-Scientist-Test (DAST). *Revista Eureka sobre Enseñanza y Divulgación de las Ciencias 15*(3), 3104. doi: 10.25267/Rev_Eureka_ensen_divulg_cienc.2018.v15.i3.3104

Introducción

Varios estudios muestran que los estudiantes que poseen concepciones negativas sobre los científicos son poco proclives a interesarse por carreras científicas (Finson 2002; McCann y Marek 2016). La exploración y comprensión de estas percepciones es relevante para la comunidad educativa, habiéndose desarrollado numerosas investigaciones de diagnosis y análisis de las concepciones de los estudiantes. El uso del Draw a Scientist Test (DAST) (Chambers 1983) se ha convertido en una práctica común para investigar estas percepciones, especialmente con niños. Usando este método, los investigadores han documentado que los estudiantes tienden a retratar a los científicos como varones de mediana o avanzada edad, con aspecto descuidado, confinados en un laboratorio, rodeados de instrumentos y disoluciones peligrosas y quienes, además, desarrollan experimentos de dudosa moralidad (DeWitt, Archer y Osborne 2013; Ruiz-Mallén y Escalas 2012).

No obstante, los resultados obtenidos están siendo cuestionados por inconsistencias metodológicas relacionadas con los instrumentos y los procedimientos utilizados (Andersen, Krogh y Lykkegaard 2014; Losh, Wilke y Pop 2008; Reinisch, Krell, Hergert, Gogolin y Krüger 2017). Dado que la persistencia de estos estereotipos puede afectar al desarrollo de vocaciones científicas y al interés por la ciencia, es necesario desarrollar nuevos protocolos de administración y análisis para diagnosticar de forma más eficaz las percepciones de las niñas y niños sobre los científicos.

El objetivo de este trabajo es presentar algunas modificaciones realizadas al DAST que han mostrado ser efectivas para superar algunas de las principales limitaciones relacionadas con el instrumento en sus versiones anteriores. Para ello, se realiza en primer lugar una revisión detallada de la literatura especializada en el diagnóstico de las imágenes sobre los científicos, prestando especial énfasis a las limitaciones relacionadas con el uso del DAST en su versión original. Posteriormente, se describen las modificaciones realizadas y se presentan los resultados de un estudio piloto desarrollado con 149 estudiantes de Educación Primaria con diferentes antecedentes culturales y de diferentes edades. Finalmente, se discute la eficacia y las limitaciones de los cambios propuestos y se señalan vías para futuras investigaciones.

Revisión de la literatura

La hipótesis que defiende la existencia de percepciones estereotipadas de los científicos entre los estudiantes fue propuesta inicialmente en la década de los años 50 del siglo XX por Mead y Métraux (1957), cuyo estudio reveló una imagen de científico como "Un hombre que lleva una bata blanca y trabaja en un laboratorio. Es de mediana o avanzada edad y usa gafas. Está rodeado de instrumentos: tubos de ensayo, quemadores bunsen, frascos y botellas, tubos de vidrio y máquinas extrañas con diales. Pasa sus días haciendo experimentos" (Mead y Métraux 1957, pp. 386-387 [traducción propia]). Desde entonces, esta descripción es considerada la imagen estereotipada clásica de un científico.

Décadas más tarde, Chambers (1983) desarrolló el Draw-a-Scientist Test (DAST), un instrumento cuya administración no requería respuestas escritas pues solicitaba a los participantes dibujar un científico. En el análisis posterior, Chambers identificó siete atributos que aparecían sistemáticamente en los dibujos: bata de laboratorio (generalmente blanca), gafas, vello facial (barba, bigote o patillas anormalmente largas), símbolos de investigación (instrumentos científicos), símbolos de conocimientos (libros y librerías), tecnología (productos de la ciencia) y frases relevantes (¡Eureka! y similares). Chambers también identificó la presencia de criaturas místicas y personajes literarios (Frankenstein; Doctor Jekyll y el Señor Hyde), representaciones mágicas de laboratorios alquímicos y científicos locos practicando torturas. Años más tarde, Finson Beaver y Cramond (1995) desarrollaron la rúbrica DAST-C (DAST *checklist*) añadiendo nueve indicadores adicionales que incorporaban la raza y el sexo de los científicos, así como otros elementos repetidamente presentes en estudios anteriores pero que no fueron incorporados a la lista original del instrumento. Adaptaciones más recientes del DAST original son el E-DAST (*Enhanced* Draw-A-Scientist Test), que consiste en la realización de un conjunto de tres dibujos sobre los científicos para estudiar cómo varían las concepciones en sucesivas representaciones (Farland-Smith y McComas 2006) o el mDAST (*modified* DAST), que solicita a los estudiantes incluir explícitamente información sobre la apariencia, el lugar de trabajo y la actividad, permitiendo así obtener aquellos datos menos probables de ser incluidos por los alumnos en sus dibujos (Farland-Smith 2012).

En conjunto, tanto la versión original como las sucesivas adaptaciones del DAST muestran que los estereotipos sobre los científicos son persistentes y que estos se repiten en todos los niveles académicos, acentuándose y consolidándose a medida que aumenta el nivel escolar

(Medina-Jerez, Middleton y Orihuela-Rabaza 2011; Özel 2012; Ruiz-Mallén y Escalas 2012). Así, en los dibujos predomina la presencia de científicos varones, de rasgos caucásicos, con gafas y vello facial, que visten batas de laboratorio y que mayormente realizan trabajos relacionados con la química (DeWitt *et al.* 2013; Finson 2002; Özgelen 2012; Türkmen 2008; Vázquez y Manassero 1998; Zhai, Jocz y Tan 2014). Además, varios estudios indican que los niños poseen concepciones más estereotipadas que las niñas (Newton y Newton 1998; Ruiz-Mallén y Escalas 2012) y que la cultura y procedencia de los participantes tiene un alto impacto en la formación de imágenes de científicos. Por ejemplo, (Farland-Smith 2009) registró diferencias en las representaciones de los estudiantes estadounidenses, que concibieron a los científicos trabajando en laboratorios con productos químicos mientras que las representaciones de estudiantes de origen asiático mostraban científicos en fábricas creando robots.

Limitaciones metodológicas del DAST

En los últimos años, las críticas relacionadas con el uso del DAST para la identificación de concepciones sobre los científicos han ido en aumento, discutiéndose si los estereotipos pueden resultar de la falta de habilidad artística y no de los prejuicios arraigados, o si las imágenes dibujadas reflejan los estereotipos de la sociedad presentes en los medios de comunicación en lugar de la propia idea que poseen los estudiantes sobre los científicos. Losh *et al.* (2008) señalan la falta de consideración del diferente desarrollo madurativo entre niñas y niños como uno de los principales defectos del DAST. Así, las diferencias en los dibujos pueden deberse al mayor control fino de las niñas en comparación con los niños (Boyd y Bee 2006) y no a una diferencia de concepciones, sucediendo lo mismo con los estudiantes con mayor talento artístico o aquellos de niveles superiores.

Otras críticas están relacionadas con la interpretación de los dibujos. Losh *et al.* (2008) encontraron una gran cantidad de representaciones en las que el sexo de los científicos mostraba un elevado grado de ambigüedad. Andersen *et al.* (2014), utilizando entrevistas, obtuvieron que, a diferencia de la mayoría de estudios anteriores que emplearon el DAST, muchas de las opiniones que tenían los estudiantes sobre los científicos eran realistas y los rasgos estereotipados solían ser positivos, como, por ejemplo, considerar que los científicos desean descubrir nuevos conocimientos, son curiosos, lógicos y analíticos.

Otra principal limitación está relacionada con la fiabilidad del protocolo DAST, en particular, con las diferentes formas de administrar el instrumento (solicitar un dibujo o varios) y con las instrucciones facilitadas que podrían introducir sesgo en la respuesta del participante. Maoldomhnaigh y Hunt (1988) solicitaron a 76 estudiantes de quinto nivel de Educación Primaria que realizaran varios dibujos de científicos. La frecuencia en la aparición de aspectos místicos y caricaturescos fue mayor en el segundo dibujo de los participantes, lo que parece mostrar que un mismo estudiante puede tener varias visiones sobre los científicos, y que un solo dibujo no es suficiente para captar la totalidad de las concepciones. Farland-Smith y McComas (2006) utilizaron un enfoque similar y su estudio reveló que en el 76% de los casos, el primer dibujo de los estudiantes difiere significativamente de los posteriores. Maoldomhnaigh y Mhaolái (1990) pidieron a un grupo de participantes dibujar un científico y a otro grupo dibujar un hombre o una mujer científica. Los resultados mostraron que las instrucciones facilitadas alteraron significativamente el tipo de dibujos realizados por los participantes. Mientras que en el primer grupo tan solo el 28% de las niñas han dibujado científicas mujeres, en el segundo grupo el 49% de los dibujos de las niñas representaban a una mujer científica.

En definitiva, el estudio de las concepciones de los estudiantes sobre los científicos ha de tener en cuenta los siguientes aspectos: (I) el desarrollo madurativo es diferentes en los niños y las niñas; (II) las habilidades artísticas son muy diferentes entre el alumnado; (III) las instrucciones en la administración del instrumento deben reducir el sesgo; y, (IV) el uso solo de dibujos puede ser insuficiente para captar la totalidad de la concepción de los estudiantes. Por lo tanto, existe una clara necesidad de desarrollar protocolos alternativos y mejorados para evaluar qué imagen de científico tienen los niños y las niñas de Educación Primaria.

Descripción del nuevo protocolo

Para superar algunas de las limitaciones relacionadas con el DAST, se propone un protocolo multidimensional en el que se estudia la imagen de los estudiantes a través de dos fases: una primera, cuantitativa, que consiste en realizar un dibujo sobre los científicos; y, una segunda fase, cualitativa que consta de preguntas cortas que se han de responder de forma escrita y una prueba de selección de imágenes. A través de estas fases, se pretende abordar cinco aspectos diferentes de las concepciones de los estudiantes: (I) apariencia de los científicos, (II) origen, (III) lugar de trabajo, (IV) actividad, y (V) forma en que realizan dicha actividad. La tabla 1 recoge qué limitaciones de las señaladas en la literatura pretenden reducir cada una de las fases del protocolo propuesto.

Tabla 1. Limitaciones del DAST que se pretenden reducir mediante el protocolo propuesto

Limitaciones del DAST	Dibujo	Preguntas escritas	Selección de imágenes
Diferente desarrollo madurativo (Losh *et al.* 2008)	x	x	x
Habilidades artísticas dispares (Losh *et al.* 2008; Reinisch *et al.* 2017)	x	x	x
Sexo del científico ambiguo (Losh *et al.* 2008; Reinisch *et al.* 2017)		x	x
Edad del científico ambiguo (Reinisch *et al.* 2017)		x	
Etnia/nacionalidad del científico ambigua (Reinisch *et al.* 2017)		x	x
Dibujos socialmente reconocibles (Finson y Pederson 2011)		x	x
Los dibujos proporcionan información limitada (Ehrlen, 2009; Reinisch *et al.*,2017)		x	x
Resultados dependientes de la administración del instrumento (Farland-Smith y McComas 2006; Maoldomhnaigh y Mhaoláin 1990; Maoldomhnaigh y Hunt 1988)	x	x	x
Posible existencia de más de una imagen de científico en cada participante (Farland-Smith y McComas 2006; Finson *et al.* 1995)	x	x	x

Dibujo

Se ha modificado la instrucción original "Dibuja un científico" (Chambers 1983, p. 256 [traducción propia]) por "Imagina que vas de viaje a ver cómo se hace ciencia. Dibuja quién hace ciencia, cómo es, dónde trabaja y qué hace", basada en Farland-Smith (2012). Esta instrucción solicita de manera explícita que se incluya en el dibujo información sobre la apariencia del científico, su trabajo y el lugar en el que lo realiza. Nótese que se ha suprimido la palabra "científico" dado que, en castellano, a diferencia del inglés, esta palabra hace referencia explícita al género masculino. Además, se ha evitado cualquier referencia a personas (p. ej. Dibuja a una persona que hace ciencia) a fin de evitar introducir sesgo en las representaciones y reducir en mayor medida las limitaciones asociadas con resultados dependientes de la forma en que se administra el instrumento (Farland-Smith y McComas 2006; Maoldomhnaigh y Mhaoláin 1990; Maoldomhnaigh y Hunt 1988). Otras modificaciones están relacionadas con la administración del protocolo. Se ha estandarizado los materiales disponibles, entregándose a cada estudiante cinco lápices de colores para reducir los problemas resaltados en la literatura con respecto a la diferencia de detalles dibujados por las niñas y los niños (Losh *et al.* 2008; Reinisch *et al.* 2017) y la falta de material para transmitir las propias concepciones (Reinisch *et al.* 2017).

Para evaluar esta dimensión del protocolo, se emplea una categorización (Tabla 2) adaptada a partir de Farland-Smith (2012), en la que se evalúan tres aspectos (apariencia, ubicación y actividad) de los dibujos en función de cuatro posibles concepciones (sensacionalista, tradicional-estereotipada, alternativa-no estereotipada, y aspectos imposibles de determinar por falta de información o dibujos no interpretables).

Tabla 2: Categorización para analizar los dibujos

Aspectos	**Concepciones**			
	Imposible de determinar	**Sensacionalista**	**Tradicional-estereotipada**	**Alternativa-no estereotipada**
Apariencia	Figuras de palo; otra profesión; no se puede identificar.	Caricatura; monstruo; científico loco.	Varón caucásico; bata de laboratorio; mediana-avanzada edad.	Minorías étnicas; extranjeros; mujeres
Actividad que realizan	No aparece; no se puede identificar	Actividades peligrosas (explosiones, fuego, pociones); torturas.	Propias de laboratorio; experimentos químicos; trabajo individual.	Actividades diferentes a las de un laboratorio convencional (redacción); trabajo grupal.
Lugar de trabajo	No aparece; no se puede identificar.	Sótano; cueva; caverna; lugares secretos.	Laboratorio tradicional, particularmente de química	Lugares diferentes a un laboratorio tradicional (yacimientos arqueológicos)

Las modificaciones realizadas están relacionadas con el apartado actividad que realizan los científicos. Así, mientras que en la rúbrica original se consideraba una concepción alternativa-no estereotipada en tanto el dibujo represente "(…) actividades realistas que reflejan el trabajo de un científico (…) con las herramientas apropiadas y necesarias" (Farland-Smith 2012 p. 116 [traducción propia]); en la adaptación de la categorización se ha estimado más apropiado considerar como concepciones alternativas-no estereotipadas a aquellos dibujos que representen actividades diferentes a las propias de un laboratorio convencional y aquellos dibujos que representan varios científicos trabajando de forma grupal. Las razones que justifican este cambio son las siguientes: dibujar un científico trabajando en un laboratorio, especialmente realizando experimentos de química, es uno de los estereotipos más comunes encontrados en la literatura, con independencia de si se representan los materiales necesarios para dicha actividad. Además, los elementos dibujados pueden poseer diferentes significados para el participante y para el investigador (Reinisch *et al.* 2017), por lo que la interpretación de si se han representado las herramientas necesarias resulta muy ambigua. Otro estereotipo común encontrado en la literatura es considerar al científico como una persona solitaria y aislada socialmente. Por ello, parecería más apropiado incluir en la categoría "tradicional-estereotipada" aquellos dibujos que representen a un científico trabajando de forma individual o realizando experimentos propios de un laboratorio convencional (p. ej. Mezclas y disoluciones, uso de probetas y pipetas), y en la categoría "alternativa-no estereotipada" los dibujos que representen científicos trabajando de forma grupal o actividades diferentes a las de laboratorio (p. ej.: redacción de artículos, búsqueda de información, toma de muestra en la naturaleza). Por último, se ha prescindido del sistema de puntuación original según el cual, en función del grado de estereotipia, los dibujos obtenían una puntuación que variaba entre 0 y 9 puntos totales (Farland-Smith 2012, p. 115-116). Así, cada aspecto de los dibujos es catalogado en cada categoría según las cuatro concepciones descritas anteriormente.

Preguntas escritas

En esta dimensión, se solicita a los participantes contestar de forma escrita a una serie de preguntas relacionadas con el dibujo, lo que permitiría superar hándicaps de habilidades artísticas y de diferente desarrollo madurativo, y, además, proporciona más información que ayuda a interpretar el dibujo realizado, siendo estos aspectos señalados como importantes limitaciones del DAST (Ehrlén 2009; Reinisch *et al.* 2017). Las preguntas, situadas en el reverso del folio bajo el epígrafe "Contesta a las siguientes preguntas sobre tu dibujo", son: (I) ¿Quiénes son?; (II) ¿Cómo son?; (III) ¿Dónde trabajan?; (IV) ¿Qué hacen?; y (V) ¿Qué objetos utilizan? Los estudiantes han dispuesto de una hora para la realización del dibujo y de las preguntas escritas.

Prueba de selección de imágenes

En la tercera dimensión del protocolo propuesto, se emplea una entrevista con fotografías, método útil en la investigación cualitativa con niños (Cappello 2005). La prueba consiste en solicitar a los participantes responder a las preguntas que realiza el entrevistador, seleccionando aquellas (o ninguna) imágenes que representen su respuesta. Para la administración de la prueba, se facilita la siguiente instrucción "Voy a realizarte varias preguntas. Para contestar, debes seleccionar una, más de una, o ninguna de las fotografías que te voy a enseñar. Explica tu elección". A continuación, el investigador realiza una pregunta (tabla 3, segunda columna) y muestra las tres imágenes correspondientes (tabla 3, tercera columna).

Tabla 3. Rúbrica para analizar la prueba de selección de imágenes.

Aspectos	Pregunta	Descripción de las fotografías
Apariencia	¿Quién hace ciencia?	1. Científico loco, varones, realizando experimentos peligrosos. 2. Científico varón, mediana-avanzada edad, con bata blanca, trabajando en un laboratorio. 3. Científica femenina, joven y apariencia normal.
Origen	¿Cómo es?	1. Figura o persona de apariencia ficticia real, con rasgos de mutaciones extrañas/robots. 2. Rasgos caucásicos 3. Personas negras o extranjeros.
Lugar de trabajo	¿Dónde trabaja?	1. Cuevas, sótanos o laboratorios subterráneos y secretos. 2. Laboratorio convencional, principalmente con elementos de química. 3. Cualquier lugar real diferente a un laboratorio convencional, como observatorios o yacimientos arqueológicos.
Actividad	¿Qué hace en su trabajo?	1. Experimentos de dudosa moralidad, como bombas, torturas o pociones. 2. Actividades propias de un laboratorio de química. 3. Actividades propias de otras disciplinas científicas: astronomía, arqueología.
Forma de trabajar	¿Cómo trabaja?	1. Aislado socialmente y con rasgos de estar exhausto. 2. En solitario. 3. De forma grupal.

Nota: 1. Fotografía con concepciones sensacionalistas; 2. Fotografía con concepciones tradicionales-estereotipadas; 3. Fotografía con concepciones alternativas-no estereotipadas.

Las imágenes han sido seleccionadas para representar las tres posibles categorías de concepciones de los estudiantes: sensacionalista, que representa motivos y actividades que no son propias del quehacer científico; tradicional-estereotipado, que recoge visiones estereotipadas comúnmente identificadas en la literatura; alternativas-no estereotipadas, que representa aspectos no estereotipados sobre los científicos y su trabajo. La inclusión de la prueba de selección de imágenes pretende recabar más información sobre las concepciones de los estudiantes y así determinar con mayor exactitud aquellos aspectos señalados como deficiencias metodológicas del DAST. Más específicamente, en relación con las limitaciones sobre el sexo, apariencia y etnia de los científicos (Losh *et al.* 2008, Reinisch *et al.* 2017), en la prueba de imágenes se han incluido fotografías de científicos de ambos sexos (Tabla 3, apariencia), y de rasgos caucásicos, asiáticos y personas negras (Tabla 3, origen). Asimismo, se han incluido imágenes sobre científicos trabajando de forma solitaria y grupal (Tabla 3, forma de trabajar). Además, al permitir al alumnado seleccionar más de una imagen, se supera la limitación relacionada con la coexistencia de más de una concepción en cada participante (Farland-Smith y McComas 2006; Finson *et al.* 1995), lo que permitiría reducir considerablemente la falta de información derivada del dibujo realizado (Ehrlén 2009; Reinisch *et al.* 2017). Las elecciones de los estudiantes se evalúan clasificándose en función de las tres posibles concepciones (sensacionalista, tradicional-estereotipada, alternativas-no estereotipadas).

Prueba piloto

Metodología

Participantes y procedimiento

Para estudiar la efectividad del protocolo propuesto con alumnado de diversas capacidades, destrezas y rendimiento académico, se conformó una muestra de conveniencia de diversidad cultural compuesta por un total de 149 estudiantes de 2º (n = 26), 3º (n = 13), 4º (n = 17), 5º (n = 33) y 6º (n = 60) de Educación Primaria. Se han incluido estudiantes españoles (n = 80), inmigrantes de segunda generación (n = 41) y alumnado español de etnia gitana (n = 28). Los estudiantes españoles provenían de un colegio con bajo porcentaje de estudiantes inmigrantes o de etnia gitana situado en una zona de clase media-alta de la ciudad de Burgos (España) Los inmigrantes de segunda generación asistían a una escuela con una muy alta diversidad cultural, también situado en la ciudad de Burgos. La totalidad de los estudiantes de segunda generación habían nacido en España; sin embargo, sus padres provenían de otros países, la mayoría de ellos del este de Europa. Por último, los alumnos españoles de etnia gitana asistían a una escuela segregada en la que la totalidad de los estudiantes eran gitanos, situada en una zona periférica de Valladolid (España).

Para la fase cualitativa, se conformó un segundo muestreo, seleccionando aleatoriamente un número de participantes de la fase cuantitativa. Inicialmente, se pretendía seleccionar un 20% del número total de participantes, sin embargo, debido a limitaciones de tiempo de los centros educativos, la muestra cualitativa fue menor, con un total de 15 estudiantes (12,7% de la muestra total), de los cuales cinco fueron españoles (3 niñas), seis extranjeros de segunda generación (2 niñas) y cuatro españoles de etnia gitana (2 niñas). El dibujo y las preguntas escritas se realizaron en una misma sesión para todos los participantes. La prueba de selección de imágenes se desarrolló en el mismo día en el que fue administrado el dibujo y las preguntas escritas en el caso de los estudiantes españoles de etnia gitana, y un día más tarde en el caso de los estudiantes españoles y de los extranjeros de segunda generación.

Análisis

Para estudiar la efectividad del protocolo propuesto, se han seguido varios procedimientos diferentes de análisis. En primer lugar, para la fase cuantitativa, todos los dibujos fueron analizados utilizando la tradicional rúbrica DAST-C (Finson *et al.* 1995), otorgando un punto por cada estereotipo identificado en el dibujo de cada participante. En segundo lugar, se analizaron nuevamente los dibujos utilizando la categorización propuesta en este estudio, comparando y discutiendo los resultados obtenidos con respecto a la rúbrica DAST-C.

En base a los resultados reportados en la literatura, se han efectuado, para ambas rúbricas, comparaciones según las variables sexo, etnia/procedencia y nivel escolar de los participantes, empleándose pruebas estadísticas no paramétricas debido a la violación de la presunción de distribución de probabilidad normal ($p < .001$), indicada por la prueba Kolmogorov-Smirnov con corrección de Lilliefors. Además, para determinar la robustez en la utilización de la categorización propuesta se empleó el índice de Kappa para calcular la concordancia inter-observador. Para ello, los primeros dos autores de este estudio evaluaron de forma independiente un conjunto de imágenes seleccionadas de forma aleatoria ($n = 30$).

Finalmente, para la fase cualitativa de la prueba piloto, se han analizado las preguntas escritas y la prueba de selección de imágenes, determinando si coincide con el dibujo realizado y discutiendo aquellos aspectos positivos que añade a la interpretación de los mismos.

Resultados y discusión

Análisis de los dibujos empleando la rúbrica DAST-C

Los dibujos analizados representaron una imagen generalizada del científico como un varón, especialmente en el caso de los niños, de rasgos caucásicos, que está vestido con una bata y que realiza investigaciones en un laboratorio de forma individual. Estos estereotipos persisten en ambos sexos y en todos los niveles escolares, siendo "trabajo individual", "laboratorio", "investigación" y "varón de rasgos caucásicos" los indicadores estereotipados más representados por los participantes (Tabla 4).

Tabla 4. Porcentaje de participantes que han dibujado los indicadores de la rúbrica DAST-C ($n = 149$)

Indicadores estereotipados		**Niños**	**Niñas**	**2º E.P.**	**3º E.P.**	**4º E.P.**	**5º E.P.**	**6º E.P.**
		%	%	%	%	%	%	%
Apariencia	Creatura mística	16.7	8.2	11.8	18.2	0	18.8	12.3
	Bata laboratorio	56.9	65.6	52.9	54.5	62.5	65.6	61.4
	Gafas	44.4	41.0	23.5	36.4	62.5	34.4	49.1
	Vello facial	4.2	0	5.9	0.0	0	3.1	1.8
	Varón	88.9	45.9	47.1	81.8	68.8	62.5	77.2
	Caucásico	83.3	100.0	100.0	90.9	75.0	93.8	91.2
	Edad avanzada	16.7	18.0	5.9	18.2	18.8	9.4	24.6
	Loco-enfadado	25.0	14.8	0.0	36.4	18.8	34.4	15.8
Actividad	Investigación	88.9	90.2	76.5	72.7	87.5	96.9	93.0
	Conocimiento	22.2	23.0	5.9	27.3	25.0	31.3	21.1
	Productos	22.2	11.5	41.2	45.5	18.8	12.5	7.0
	Expresiones	16.7	18.0	23.5	45.5	6.3	6.3	19.3
	Símbolos peligro	15.3	9.8	11.8	9.1	0.0	15.6	15.8
	Individual	88.9	86.9	94.1	72.7	81.3	90.6	89.5
Lugar de trabajo	Secreto	0.0	1.6	0.0	0.0	6.3	3.1	0.0
	Laboratorio	80.6	90.2	64.7	81.8	87.5	84.4	91.2
	Clandestino	2.8	4.9	17.6	9.1	0.0	0.0	1.8
	Bombillas	6.9	6.6	5.9	9.1	6.3	6.3	7.0

En cuanto a la variable sexo, el estadístico U de Mann-Whitney reveló que no hay diferencias significativas en el número total de indicadores estereotipados dibujados por los niños y las niñas, $U = 1811.500$, $z = -.902$, $p = .367$, $r = .08$. No obstante, atendiendo a la tipología de los indicadores, se obtuvo que las niñas han dibujado significativamente más estereotipos relacionados con el *lugar de trabajo* de los científicos en comparación con los niños, U = 1674.500, z = -2.317, p = .02, r = .20. La figura 1 muestra el número medio de estereotipos dibujados por las niñas y los niños.

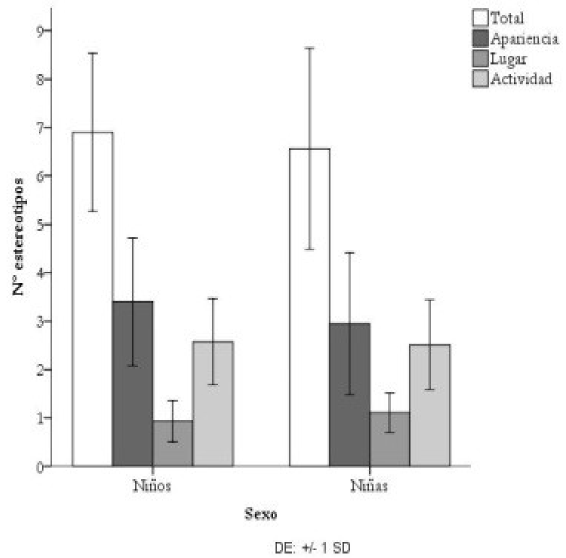


Figura 1. Número medio de estereotipos dibujados según el sexo de los estudiantes

El estadístico Kruskal-Wallis reveló que no hay diferencias significativas en el número total de indicadores estereotipados dibujados por los estudiantes españoles, extranjeros de segunda generación y españoles de etnia gitana, χ^2 (2, n = 127) = 1.920, p = .383. Sin embargo, atendiendo a los estereotipos según su tipología, se registró que los estudiantes de etnia gitana han dibujado un mayor número de estereotipos relacionados con el *lugar de trabajo* de los científicos que los estudiantes extranjeros y que los españoles, χ^2 (2, n = 127) = 6.714, p = .035. La figura 2 muestra el número medio de estereotipos dibujados según el nivel escolar.

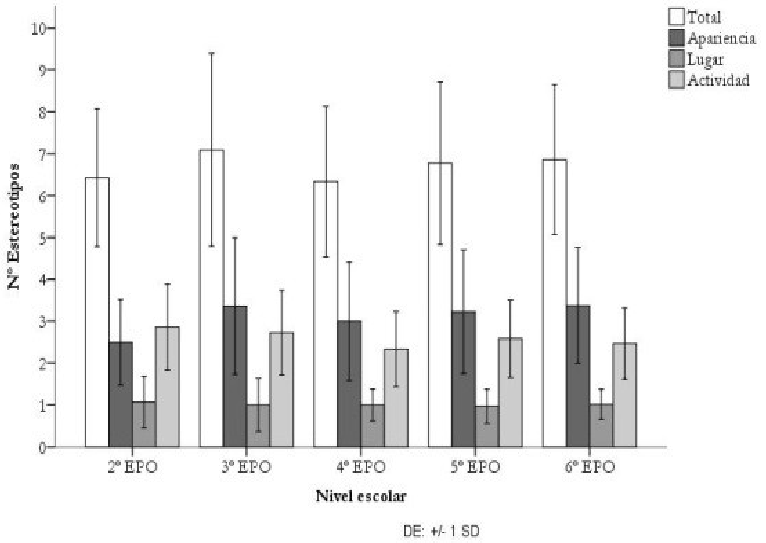


Figura 2. Número medio de estereotipos dibujados según el nivel escolar de los estudiantes

Asimismo, en relación con el nivel escolar, el estadístico Kruskal-Wallis reveló que tampoco hay diferencias significativas en el número total de indicadores estereotipados dibujados, χ^2 (4, n = 127) = 1.386, p =.847. Tampoco se encontraron diferencias significativas en las dimensiones *apariencia de los científicos*, χ^2 (4, n = 127) = 4.679, p = .322, *lugar de trabajo* χ^2 (4, n =

127) = .629, p = .960, ni en la *actividad que realizan* χ^2 (4, n = 127) = 3.463, p = .484. La figura 1 muestra el número medio de estereotipos dibujados por los estudiantes españoles, extranjeros de segunda generación y españoles de etnia gitana.

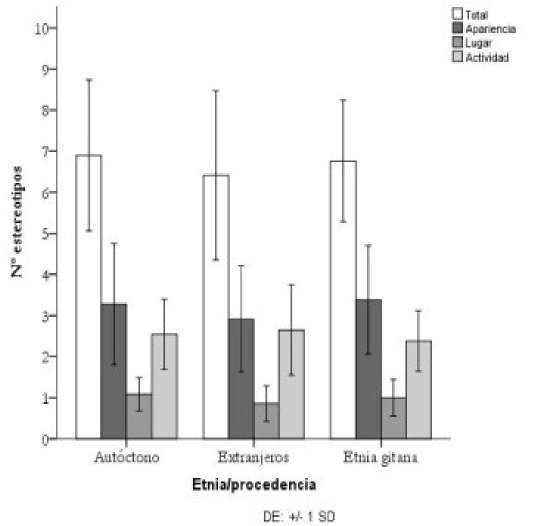


Figura 3. Número medio de estereotipos dibujados según los antecedentes culturales de los estudiantes

Análisis de los dibujos empleando la rúbrica modificada

En general, los dibujos de los estudiantes han mostrado concepciones estereotipadas sobre los científicos. Así, un 57,1% de los dibujos han plasmado concepciones estereotipadas y sensacionalistas sobre la *apariencia de los científicos*, un 68.5% sobre el *trabajo que realizan*, y finalmente un 81,9% sobre el *lugar* en el que desempeñan su labor. Más específicamente, la mayoría de los dibujos han representado científicos varones, de rasgos caucásicos, con bata de laboratorio y mediana-avanzada edad, trabajando de forma individual en un laboratorio convencional realizando actividades propias de laboratorio, como experimentos químicos. La *apariencia de los científicos* ha sido la categoría que ha registrado un mayor número de concepciones alternativas-no estereotipadas (32.9%). Por el contrario, *la actividad de los científicos* ha sido la categoría con un mayor número de dibujos que muestran imágenes sensacionalistas (21.5%). Estos resultados se desglosan a continuación y se recogen en las tablas 5, 6 y 7.

Para la *apariencia de los científicos* (Tabla 5), el test de independencia de Chi-cuadrado reveló la existencia de diferencias significativas en la distribución de los participantes en las cuatro categorías de concepciones sobre los científicos para la variable sexo, χ^2 (3, n = 149) = 13.117, p = .004, y nivel escolar, χ^2 (12, n = 149) = 47.881, p < .001. Atendiendo a la variable etnia/procedencia de los estudiantes, no se identificaron diferencias significativas en su distribución χ^2 (6, n = 149) = 8.844, p = .183. Así, en la categoría tradicional-estereotipada, predominaron los dibujos realizados por los niños (65%), y, por el contrario, han sido las niñas las que han plasmado un mayor número de dibujos categorizados como alternativos-no estereotipados (67.3%), representando minorías étnicas, extranjeros y/o mujeres en lugar de hombres caucásicos o caricaturas/científicos locos. Además, los estudiantes de niveles superiores han sido mayoría en la categoría sensacionalista (52%) y tradicional-estereotipada (51.7%) que los de niveles inferiores, especialmente en comparación con los de 2º de Educación Primaria, con apenas un 12% y 3.3% de dibujos categorizados como sensacionalistas y tradicionales-estereotipados, respectivamente.

Tabla 5. Categorización de los dibujos según la apariencia de los científicos

	Sin determinar (%)	**Sensacionalista** (%)	**Tradicional estereotipada** (%)	**Alternativa no estereotipada** (%)
Niñas	40	36	35	67.3
Niños	60	64	65	32.7
2º E.P.	66.7	12	3.3	22.4
3º E.P.	6.7	8	13.3	4.1
4º E.P.	6.7	0	15	14.3
5º E.P.	0	28	16.7	30.6
6º E.P.	20	52	51.7	28.6
Españoles	40	76	48.3	53.1
Extranjeros	33.3	20	7	30.6
Etnia gitana	26.7	4	0	16.3
Total	*n* =15 (10)	*n* = 25 (16.8)	*n* = 60 (40,3)	*n* = 49 (32.9)

Nota: El porcentaje corresponde a cada una de las categorías (sin determinar, sensacionalista, tradicional estereotipada, alternativa-no estereotipada) y para cada variable (sexo, nivel escolar, etnia/procedencia)

Para la *actividad que realizan los científicos* (Tabla 6), el test de independencia de Chi-cuadrado reveló la existencia de diferencias significativas en la distribución de los participantes en las cuatro categorías de concepciones sobre los científicos para la variable nivel escolar, χ^2 (12, n = 149) = 47.622, p < .001. No se encontraron diferencias según la variable sexo, χ^2 (3, n = 149) = .747, p =.862, ni tampoco según la variable etnia/procedencia de los estudiantes, χ^2 (6, n = 149) = 8.503, p = .204. Más específicamente, en las categorías sensacionalistas y tradicionales-estereotipadas han predominado dibujos de estudiantes de niveles escolares superiores. En otras palabras, los estudiantes de niveles superiores (50%) conciben la actividad de los científicos de una forma más estereotipada que los de niveles inferiores (2.9%). No obstante, también son los estudiantes de niveles superiores quienes poseen visiones más alternativas-no estereotipadas sobre *las actividades que realizan los científicos*, con casi la mitad (46.2%) de los dibujos categorizados como alternativos-no estereotipados perteneciendo al 6º nivel de EP. Si bien no se han registrado diferencias significativas según la etnia/procedencia de los estudiantes, la mayoría de dibujos de la categoría alternativa-no estereotipada para la actividad representada han sido realizados por estudiantes españoles (57.7%) y extranjeros (38.5). Sin embargo, también han sido los españoles quienes han dibujado un mayor número de dibujos sensacionalistas (62.5%).

Tabla 6. Categorización de los dibujos según la actividad que realizan los científicos

	Sin determinar (%)	**Sensacionalista** (%)	**Tradicional estereotipada** (%)	**Alternativa no estereotipada** (%)
Niñas	52.4	40.6	47.1	46.2
Niños	47.6	59.4	52.9	53.8
2º E.P.	47.1	21.9	2.9	19.2
3º E.P.	9.5	15.6	5.7	7.7
4º E.P.	9.5	0	20	3.8
5º E.P.	9.5	28.1	21.5	23.1
6º E.P.	14.3	34.4	50	46.2
Españoles	42.9	62.5	51.4	57.7
Extranjeros	28.6	25	24.3	38.5
Etnia gitana	28.6	12.5	24.3	3.8
Total	*n* = 21 (14.1)	*n* = 32 (21.5)	*n* = 70 (47)	*n* = 26 (17.4)

Nota: El porcentaje corresponde a cada una de las categorías (sin determinar, sensacionalista, tradicional estereotipada, alternativa-no estereotipada) y para cada variable (sexo, nivel escolar, etnia/procedencia)

Para la dimensión *lugar de trabajo de los científicos* (Tabla 7), el test de independencia de Chi-cuadrado reveló la existencia de diferencias significativas en la distribución de los participantes en las cuatro categorías de concepciones sobre los científicos para la variable nivel escolar, χ^2 (12, n = 149) = 46.167, $p < .001$. No se encontraron diferencias según la variable sexo, χ^2 (3, n = 149) = .469, p =.926, ni tampoco según la variable etnia/procedencia, χ^2 (6, n = 149) = 6.187, p =.403. Por tanto, los estudiantes de niveles superiores son quienes asimismo conciben el *lugar de trabajo* de los científicos de una forma más estereotipada en comparación con los de niveles inferiores. Así, casi la mitad de los dibujos (48.5%) que han representado laboratorios de ciencia convencionales pertenecen a estudiantes de 6º nivel. Por otro lado, los estudiantes de niveles inferiores son los que más dibujos sensacionalistas han dibujado (66.7%), representando cuevas, sótanos o lugares secretos. En línea con la *actividad que realizan los científicos*, han sido los estudiantes españoles quienes han mostrado una visión más estereotipada sobre los lugares de trabajo de los científicos, con un 66.7% y 57.4% de los dibujos sensacionalistas y tradicionales-estereotipados dibujados por ellos.

Tabla 7. Categorización de los dibujos según el lugar de trabajo de los científicos.

	Sin determinar (%)	Sensacionalista (%)	Tradicional estereotipada (%)	Alternativa no estereotipada (%)
Niñas	52.4	50	45.5	42.9
Niños	47.6	50	54.5	57.1
2º E.P.	57.1	66.7	5.9	19
3º E.P.	9.5	0	7.9	14.3
4º E.P.	9.5	0	13.9	4.8
5º E.P.	9.5	16.7	23.8	23.8
6º E.P.	14.3	16.7	48.5	38.1
Españoles	42.9	66.7	57.4	42.9
Extranjeros	28.6	33.3	23.8	42.9
Etnia gitana	28.6	0	18.8	14.3
Total	n = 21 (14.1)	n = 6 (4)	n = 101 (67.8)	n = 21 (14.1)

Nota: El porcentaje corresponde a cada una de las categorías (sin determinar, sensacionalista, tradicional estereotipada, alternativa-no estereotipada) y para cada variable (sexo, nivel escolar, etnia/procedencia)

En conjunto, los resultados de la rúbrica propuesta contradicen y complementan algunos datos obtenidos utilizando la rúbrica DAST-C. De este modo, mientras que utilizando la rúbrica DAST-C se ha obtenido que las niñas poseen una visión más estereotipada sobre *el lugar de trabajo de los científicos* en comparación con los niños, los resultados de la dimensión representativa no soportan estas evidencias, sugiriendo que tanto niñas como niños conciben *el lugar de trabajo de los científicos* de manera similar. También se han identificado discrepancias entre las dos rúbricas en cuanto a las variables nivel escolar y etnia/procedencia de los estudiantes. Según la rúbrica DAST-C, los estudiantes de todos los niveles escolares poseen concepciones similares, siendo los de etnia gitana los que conciben de forma más estereotipada el *lugar de trabajo* de los científicos. Estos resultados no coinciden con los obtenidos empleando la nueva rúbrica propuesta, que muestra, por un lado, que los estudiantes de niveles superiores conciben de forma más estereotipada tanto la *apariencia de los científicos*, la *actividad que realizan,* así como *el lugar de trabajo.* Además, empleando la nueva rúbrica propuesta no se han encontrado diferencias entre las concepciones de los estudiantes españoles, de etnia gitana y de los extranjeros.

Por otro lado, las modificaciones realizadas a la rúbrica han permitido establecer nuevas diferencias entre las concepciones de los estudiantes. A diferencia de la rúbrica DAST-C, la dimensión representativa ha señalado diferencias significativas en las concepciones de los estudiantes sobre la apariencia de los científicos, siendo los niños los que han realizado un mayor número de dibujos sensacionalistas (64%) y tradicionales estereotipados (65%). Además, la rúbrica DAST-C no ha encontrado diferencias según el nivel escolar en ninguna de las dimensiones de los dibujos, sin embargo, la nueva rúbrica propuesta ha permitido identificar que a medida que aumenta el nivel escolar también aumentan las imágenes estereotipadas sobre los científicos, tanto en relación su *apariencia*, como sobre la *actividad que realizan* y el *lugar en el que desempeñan su trabajo.*

En cuanto a la categorización de los dibujos siguiendo la rúbrica propuesta, el índice de Kappa mostró un índice alto de concordancia inter-observador (k = .85) (Landis y Koch 1977), sugiriendo que la categorización de los dibujos empleando la rúbrica propuesta es fiable.

Preguntas escritas y prueba de selección de imágenes

En este apartado, se recogen los resultados obtenidos durante la fase cualitativa que respaldan el uso de la nueva rúbrica. Además, se ejemplifican las ventajas que reporta la adición de preguntas escritas y una prueba de selección de imágenes para superar algunas de las principales limitaciones reportadas en la literatura relacionada con el uso del DAST. Estas ventajas son:

- Aportan nueva información que ha sido imposible de determinar a partir del dibujo

En la fase cuantitativa se han identificado dibujos que presentaban habilidades artísticas dispares (Losh *et al.* 2008, Reinisch *et al.* 2017) o que presentaban información limitada o confusa (Ehrlén, 2009; Reinisch *et al.* 2017). Así, la figura 4 representa una visión tradicional-estereotipada de la *apariencia de los científicos*, sin embargo, no incluye suficiente información para determinar el *trabajo que realiza* el científico ni el *lugar en el que lo realiza*. Tanto las preguntas escritas como la prueba de selección de imágenes han ayudado a suplir esta falta de información, identificando que el estudiante posee asimismo una concepción tradicional-estereotipada sobre el *lugar de trabajo* y la *tarea* que realiza. Así, el estudiante ha escrito que "(…) trabaja en un laboratorio utilizando máquinas bien grandes y haciendo experimentos guais", y ha seleccionado la segunda imagen, que representa visiones tradicionales-estereotipadas, de cada uno de los apartados de la prueba de selección de imágenes (ver tabla 3).

Otros ejemplos de dibujos con información limitada o confusa son las figuras 5 y 6. Si bien en la figura 5 se identifica una concepción tradicional-estereotipada del *lugar de trabajo*, no se puede saber con certeza si la concepción de su *apariencia* y *actividad que realiza* es sensacionalista o tradicional-estereotipada, y el sexo del científico dibujado resulta ambiguo, siendo esta otra de las principales críticas recurrentes con el uso del DAST (Losh *et al.* 2008; Reinisch *et al.* 2017). Las respuestas de la estudiante ayudaron a clarificar que se trata de un varón y de una concepción sensacionalista, dado que ha argumentado que "Es un científico loco y viejo. Tiene el pelo gris y los ojos negros. A veces es muy torpe y se lía con los componentes. Trabaja en un laboratorio utilizando componentes peligrosos." La prueba de selección de imágenes ha ayudado a confirmar la concepción tradicional-estereotipada sobre el *lugar de trabajo* (seleccionando la imagen que representa un laboratorio) y la concepción sensacionalista sobre la *actividad que realiza*, puesto que ha seleccionado la imagen que representa experimentos de dudosa moralidad, como pócimas, y la primera imagen que representa a un científico con rasgos de estar cansado trabajando de manera aislada socialmente. No obstante, en relación con la apariencia, la alumna ha seleccionado las primeras dos fotografías que representan, por un lado, un científico varón realizando experimentos peligrosos, y por otro, un científico varón, de mediana-avanzada edad, con bata blanca, trabajando en un laboratorio (ver tabla 3), lo que parece sugerir la coexistencia de más de una imagen sobre los científicos que no ha podido ser plasmada mediante un solo dibujo (Farland-Smith y McComas, 2006; Finson *et al.*

1995).

La figura 6 representa una situación similar. Las preguntas escritas ayudaron a confirmar que la alumna posee una visión sensacionalista sobre la apariencia de los científicos y a determinar una visión sensacionalista sobre el *trabajo que realizan* y una visión tradicional-estereotipada sobre el *lugar de trabajo*, argumentando que "Es feo, bajo y loco. Sabe hacer pociones. Utiliza probetas y otros artilugios que están en el laboratorio". En cuanto a la prueba de selección de imágenes, permitió confirmar la visión sobre el *trabajo* que realiza y el *lugar* de trabajo. No obstante, la alumna argumentó que los científicos trabajan tanto de forma individual como en grupo, lo que sugiere una visión alternativa-no estereotipada sobra la *forma de trabajar de los científicos*.

Figura 4. Niño etnia gitana, 3º E.P.

Figura 5. Niña española, 5º E.P.

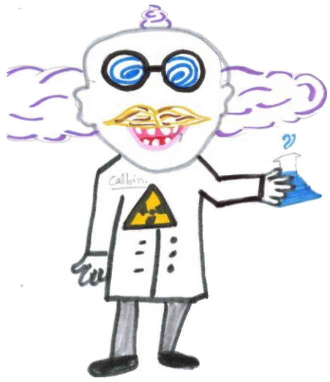


Figura 6. Niña extranjera, 6º E.P.

- Aclaran detalles del dibujo

Además de los ejemplos anteriores, en ocasiones, no se puede interpretar la etnia o la nacionalidad de los científicos dibujados (Reinisch *et al.* 2017), siendo este el caso de la figura 7, en la que no resulta claro si el estudiante ha coloreado a la persona dibujada para representar su etnia o nacionalidad. La prueba de selección de imágenes ha permitido arrojar más luz a este aspecto, sugiriendo que el estudiante posee una visión tradicional-estereotipada sobre la apariencia del científico al haber seleccionado la imagen de un hombre caucásico (pregunta 1) y de rasgos occidentales (pregunta 2). Además, el estudiante no ha aportado más información en las preguntas escritas sobre la nacionalidad o etnia del científico dibujado, lo que parecería sugerir que el participante no posee concepciones alternativas sobre quién hace ciencia, sino que ha coloreado al científico por costumbre o por ser una práctica común, al igual que los estudiantes que realizaron los dibujos de las figuras 5 y 8.

Otra limitación está relacionada con representar a los científicos trabajando de forma solitaria. Investigaciones pasadas han mostrado que esto podría deberse más a la falta de tiempo o de habilidad artística (Reinisch *et al.* 2017), y no tanto por considerar que los científicos son personas solitarias. Por ejemplo, en la figura 6, se ha representado a un solo científico, sin embargo, en la prueba de selección de imágenes la estudiante ha seleccionado también la fotografía que muestra a los científicos trabajando en grupo, argumentando que al su trabajo requiere realizar cosas complicadas, por lo que necesitan ayuda. Si bien las modificaciones realizadas en este estudio con respecto al protocolo DAST no ha solventado esta limitación, sí han permitido identificar la posible coexistencia de más de una concepción sobre la forma en que trabajan los científicos, permitiendo a los estudiantes más oportunidades para compartir sus concepciones, algo que no sucede con la rúbrica DAST original según la cual la forma de trabajar de la científica representada en la figura 6 sería solitaria, concibiéndose que el estudiante posee una visión estereotipada sobre la manera de trabajar de los científicos.

Figura 7. Niño extranjero, 4º E.P.

Figura 8. Niña española, 6º E. P.

Figura 9. Niño español, 6º E.P.

- Aporta más información acerca de imágenes socialmente reconocibles

Finson y Pederson (2011) argumentaron que los estudiantes podrían estar plasmando determinados estereotipos ligados a los científicos con el fin de que su dibujo sea socialmente reconocible. Las preguntas escritas y la prueba de selección de imágenes muestran que la figura 9 es un ejemplo de ello. Así, mientras que el estudiante ha representado concepciones sensacionalistas sobre la apariencia y el trabajo de los científicos, en la dimensión escrita ha justificado que "Tienen muchos estudios. Pueden tener algunos defectos de cualquier tipo. Como todo el mundo. Nadie es perfecto" y, además, en la prueba de selección de imágenes ha seleccionado las fotografías que muestran visiones tradicionales-estereotipadas y alternativas en los apartados "¿Quién hace ciencia?", "¿Qué hace en su trabajo?" y "¿Cómo trabaja?" (ver tabla 3), lo que muestra que su concepción de los científicos es menos estereotipada de lo que plasmó en su dibujo. Así, la inclusión de preguntas escritas y de la prueba de selección de imágenes no solo ha puesto de relieve las limitaciones del protocolo DAST original, sino que además ha permitido que la categorización del dibujo se realice de una forma más objetiva y con mayor precaución.

- Aporta información acerca de aspectos psicológicos de los científicos

Además de concepciones relacionadas con la apariencia o el trabajo de los científicos, los estudiantes poseen creencias sobre la ciencia o el que hacer de los científicos como una carrera difícil de alcanzar y solo destinada para personas muy inteligentes. Por tanto, parecería que los factores psicológicos asociados a la ciencia y a los científicos podrían estar frenando las aspiraciones a seguir una carrera en ciencias (DeWitt y Archer 2015). Así, parecería necesario que un protocolo que aspira a diagnosticar las concepciones que poseen los estudiantes sobre los científicos sea capaz de identificar también aspectos psicológicos sobre los mismos. De este modo, la inclusión de las preguntas escritas ha permitido identificar que estas concepciones relacionadas con componentes psicológicos de los científicos están presentes desde niveles elementales del sistema educativo. A continuación, se recogen algunos extractos que aluden a la capacidad brillante de los científicos o a su inteligencia:

"Es muy listo (…)". Niño de etnia gitana de 6º de Educación Primaria.

"Suelen ser alto [sic], esbeltos, inteligentes, muy estudiosos, experimentadores. Son muy majos y hablan de cosas interesantes. No tienen ningún defecto." Niño extranjero de 6º de Educación Primaria.

"Tienen muchos estudios". Niño español de 6º de Educación Primaria.

"Es muy listo." Niña española de 5º de Educación Primaria.

"Joven, guapa y lista. Siempre aprovecha el tiempo y le encanta hacer experimentos con colores." Niña española de 6º de Educación Primaria.

Conclusiones

En este trabajo se han detallado las modificaciones realizadas al protocolo DAST y se han reportado los resultados obtenidos en un estudio piloto. En general, los resultados sugieren que los cambios propuestos son efectivos en la identificación y análisis de las percepciones e imágenes de los estudiantes acerca de los científicos, reduciendo algunos de los problemas metodológicos identificados en el uso del DAST. La prueba piloto realizada con estudiantes de Educación Primaria de diversidad cultural parece indicar que estos cambios permiten avanzar hacia el desarrollo de un protocolo más fiable y exacto en la identificación de las concepciones que poseen los estudiantes sobre los científicos, aunque aún hay aspectos que necesitan de futuros estudios.

De este modo, en primer lugar, la modificación de la rúbrica de evaluación de los dibujos ha permitido registrar más diferencias entre las concepciones de los participantes de diferente edad y etnia o antecedentes culturales gracias a una evaluación más holística de lo representado y no a la mera cuantificación del número de estereotipos plasmados, permitiendo asimismo superar los problemas derivados de habilidades artísticas dispares entre los estudiantes. Además, la estandarización de su administración y de las instrucciones facilitadas a los estudiantes ha permitido disminuir los problemas identificados en la literatura con la falta de material o tiempo para plasmar las propias concepciones.

En segundo lugar, la utilización de la nueva rúbrica ha mostrado ser robusta, obteniéndose una gran coincidencia entre las categorizaciones realizadas a una muestra de dibujos de forma independiente por dos investigadores diferentes, lo que sugiere que los resultados obtenidos de su uso poseen un alto grado de fiabilidad.

Finalmente, la adición de las preguntas escritas y de la prueba de selección de imágenes ha confirmado los resultados obtenidos mediante la nueva rúbrica, y, además, ha complementado algunas de las interpretaciones. De este modo, se han disminuido algunos problemas metodológicos de versiones anteriores del DAST, consiguiendo (I) añadir más información sobre las concepciones de los estudiantes que no han podido ser determinadas a partir del dibujo, como por ejemplo el sexo y la etnia o nacionalidad de los científicos dibujados, (II) aclarar aspectos confusos del dibujo, como por ejemplo el material que utilizan o el lugar en el que trabajan, y (III) aportando información sobre las concepciones que poseen los estudiantes acerca de los factores psicológicos de los científicos y no solo sobre su apariencia o quehacer científico.

Pese a todo lo anterior, las mejoras propuestas en este estudio no están exentas de limitaciones. Si bien se ha intentado reducir en todo lo posible el sesgo introducido por las instrucciones de administración de los dibujos, la frase empleada ("Imagina que vas de viaje a ver como se hace ciencia. Dibuja quién hace ciencia, cómo es, dónde trabaja y qué hace") podría implícitamente inducir al estudiante a dibujar una sola persona. La prueba de selección de imágenes parece soportar esta presunción, dado que la mayoría de estudiantes han seleccionado la imagen que representa a científicos trabajando de forma grupal, pese a haber dibujado a un solo científico. Por tanto, se trata de un aspecto que requiere de futuros estudios. Otro aspecto está relacionado con las preguntas escritas. Mientras que estas han permitido acceder a las concepciones que poseen los estudiantes sobre las características psicológicas de los científicos, esto se ha dado de forma implícita. Por consiguiente, parecería necesario desarrollar en futuros estudios procedimientos que soliciten de forma explícita información acerca de estos factores tan importantes en el desarrollo de las vocaciones científicas de los estudiantes. Por último, la muestra empleada para la fase cualitativa del estudio ha sido muy irregular en términos de edad de los participantes. Si bien el objetivo principal de este estudio consistió en probar un nuevo protocolo basado en el DAST y no diagnosticar en sí las imágenes que poseen los estudiantes sobre los científicos, la baja muestra de alumnado ha limitado las conclusiones que podrían extraerse de la fase cualitativa. A este respecto, convendría realizar futuros estudios en los que se conforme una muestra cualitativa más amplia y regular, seleccionando un número similar de participantes de cada categoría de concepciones (sensacionalista, tradicional-estereotipada y alternativa-no estereotipada), lo que permitiría caracterizar mejor las diferencias y similitudes entre los estudiantes en cuanto a variables de edad, sexo y etnia o nacionalidad.

En definitiva, en este trabajo se avanza en el desarrollo de protocolos más fiables para la identificación de imágenes sobre los científicos. Las modificaciones realizadas al protocolo DAST original redujo algunas de las deficiencias metodológicas reportadas en la literatura y permitió acceder de una forma más holística y fiable a las concepciones de los estudiantes. La inclusión de preguntas escritas y la prueba de selección de imágenes parecerían ser críticas pues revelaron las concepciones subyacentes que no han podido ser identificadas a partir de los dibujos, lo que ayuda a definir mejor los significados de algunas imágenes ambiguas y a confirmar el significando de otros aspectos plasmados en los dibujos.

Agradecimientos

Queremos agradecer al editor y a los dos revisores anónimos por sus comentarios y sugerencias que han contribuido a mejorar sustancialmente este trabajo. Asimismo, agradecemos especialmente la colaboración de los centros escolares y estudiantes implicados en este estudio.

Referencias

Andersen, H. M., Krogh, L. B., Lykkegaard, E. (2014) Identity Matching to Scientists: Differences that Make a Difference? *Research in Science Education*, 44(3), 439–460. doi: 10.1007/s11165-013-9391-9

Boyd, D., Bee, H. (2006) *Lifespan development*. Boston, MA: Pearson.

Cappello, M. (2005) Photo Interviews: Eliciting Data through Conversations with Children. *Field Methods*, 17(2), 170–182. doi:10.1177/1525822X05274553

Chambers, D. W. (1983) Stereotypic images of the scientist: The draw-a-scientist test. *Science Education*, 67(2), 255–265. doi:10.1002/sce.3730670213

DeWitt, J., Archer, L. (2015) Who Aspires to a Science Career? A comparison of survey responses from primary and secondary school students. *International Journal of Science Education*, 37(13), 2170–2192. doi:10.1080/09500693.2015.1071899

DeWitt, J., Archer, L., Osborne, J. (2013) Nerdy, Brainy and Normal: Children's and Parents' Constructions of Those Who Are Highly Engaged with Science. *Research in Science Education*, 43(4), 1455–1476. doi:10.1007/s11165-012-9315-0

Ehrlen, K. (2009) Drawings as representations of children's conceptions. *International Journal of Science Education*, 31(1), 41–57. doi:10.1080/09500690701630455

Farland-Smith, D. (2009) How Does Culture Shape Students' Perceptions of Scientists? Cross-National Comparative How Does Culture Shape Students' Perceptions of Scientists? Cross- National Comparative Study of American and Chinese Elementary Students. *Journal of Elementary Science Education*, 21(4), 23–42. doi:10.1007/BF03182355

Farland-Smith, D. (2012) Development and field test of the modified Draw-a-Scientist Test and the Draw-a-Scientist Rubric. *School Science and Mathematics*, 112(2), 109–116.

doi:10.1111/j.1949-8594.2011.00124.x

Farland-Smith, D., McComas, W. F. (2006) Deconstructing the DAST: Development of a valid and reliable tool for assessing students' perceptions of scientists. En *Association of Science Teacher Educator Conference*. Clearwater, FL.

Finson, K. D. (2002) Drawing a Scientist: What We Do and Do Not Know After Fifty Years of Drawings. *School Science and Mathematics*, 102(7), 335–345. doi:10.1111/j.1949-8594.2002.tb18217.x

Finson, K. D., Beaver, J. B., Cramond, B. L. (1995) Development and field test of a checklist for the draw-a-scientist test. *School Science and Mathematics*, 95(4), 195–205. doi:10.1111/j.1949-8594.1995.tb15762.x

Finson, K., Pederson, J. (2011) What are Visual Data and What Utility do they have in Science Education? *Journal of Visual Literacy*, 30(1), 66–85. doi:10.1080/23796529.2011.11674685

Losh, S. C., Wilke, R., Pop, M. (2008) Some methodological issues with "Draw a Scientist Tests" among young children. *International Journal of Science Education*, 30(6), 773–792. doi:10.1080/09500690701250452

Maoldomhnaigh, M. Ó., Hunt, Á. (1988) Some factors affecting the image of the scientist drawn by older primary school pupils. *Research in Science and Technological Education*, 6, 159–166.

Maoldomhnaigh, M. Ó., Mhaoláin, V. N. (1990) The perceived expectation of the administrator as a factor affecting the sex of scientists drawn by early adolescent girls. *Research in Science and Technological Education*, 8, 69–74.

McCann, F. F., Marek, E. A. (2016) Achieving Diversity in STEM: The Role of Drawing-Based Instruments. *Creative Education*, 7(15), 2293–2304. doi:10.4236/ce.2016.715223

Mead, M., Métraux, R. (1957) Image of the scientist among high-school students. A pilot study. *Science*, 126(3270), 384–390.

Medina-Jerez, W., Middleton, K. V., Orihuela-Rabaza, W. (2011) Using the DAST-C to explore Colombian and Bolivian Students' Images of scientists. *International Journal of Science and Mathematics Education*, 9(3), 657–690. doi:10.1007/s10763-010-9218-3

Newton, L. D., Newton, D. P. (1998) Primary children's conceptions of science and the scientist: Is the impact of a national curriculum breaking down the stereotype? *International Journal of Science Education*, 20(9), 1137–1149. doi:10.1080/0950069980200909

Özel, M. (2012) Children's images of scientists: Does grade level make a difference? *Kuram ve Uygulamada Egitim Bilimleri*, 12(SUPPL. 4), 3187–3198.

Özgelen, S. (2012) Turkish Young Children's Views on Science and Scientists. *Educational Sciences: Theory & Practice - Special Issue*, Autumn(Special Issue), 3211–3225.

Reinisch, B., Krell, M., Hergert, S., Gogolin, S., Krüger, D. (2017) Methodical challenges concerning the Draw-A-Scientist Test: a critical view about the assessment and evaluation of learners' conceptions of scientists. *International Journal of Science Education*, 39(14), 1952–1975. doi:10.1080/09500693.2017.1362712

Ruiz-Mallén, I., Escalas, M. T. (2012) Scientists Seen by Children: A Case Study in Catalonia, Spain. *Science Communication*, 34(4), 520–545. https://doi.org/10.1177/1075547011429199

Türkmen, H. (2008) Turkish primary students' perceptions about scientist and what factors affecting the image of the scientists. *Eurasia Journal of Mathematics, Science and Technology*

Education, 4(1), 55–61.

Vázquez, Á., Manassero, M. A. (1998) Dibuja un científico: imagen de los científicos en estudiantes de secundaria. *Infancia Y Aprendizaje*, 21(81), 3–26. doi:10.1174/021037098320825226

Zhai, J., Jocz, J. A., Tan, A. L. (2014) "Am I Like a Scientist?": Primary children's images of doing science in school. *International Journal of Science Education*, 36(4), 553–576. doi:10.1080/09500693.2013.791958

Supplemental material 2: Attitude toward Science and Nature of Science views study

This supplemental material includes the full-text manuscript of second preliminary study about elementary school students' attitudes towards science and views of Nature of Science:

Toma, R. B., Greca, I. M., Orozco Gómez, M. L. (2019). Attitudes towards Science and views of Nature of Science among elementary school students in terms of gender, cultural background and grade level variables. *Research in Science & Technological Education, 37*(4), 492-515. doi: 10.1080/02635143.2018.1561433

RESEARCH IN SCIENCE & TECHNOLOGICAL EDUCATION
2019, VOL. 37, NO. 4, 492–515
https://doi.org/10.1080/02635143.2018.1561433

Attitudes towards science and views of nature of science among elementary school students in terms of gender, cultural background and grade level variables

R. B. Toma [a], I. M. Greca [a] and M. L. Orozco Gómez [b]

[a]Departamento de Didácticas Específicas, Área de Didácticas de las Ciencias Experimentales, Universidad de Burgos, Burgos, España; [b]Departamento de Ciencias de la Educación, Área de Didáctica y Organización Escolar, Universidad de Burgos, Bugos, España

ABSTRACT

Background: There is a growing interest in investigating attitudes towards science and views of Nature of Science among elementary grade students in terms of gender, cultural backgrounds, and grade level variables.

Purpose: The purpose of this study is to examine the attitudes toward science and views of Nature of Science among Spanish students, Spanish students of gypsy ethnicity and second-generation Spanish students with east-European heritage, and to determine if their attitudes are related to their views of Nature of Science.

Sample: Data for this study was gathered from seven elementary schools in Spain, forming a convenience sample of 149 students enrolled from 2nd to 5th grade.

Design and Methods: The Nature of Science Instrument (NOSI) and an adaptation of the Test of Science Related Attitudes scale (TOSRA) were used. Follow-up structured interviews were performed with 15 participants.

Results: Regarding gender, boys had better attitudes toward Science than girls but more naïve views of the empirical Nature of Science. In relation to cultural background, second generation Spanish students with east-European heritage reported significantly better attitudes toward Science than Spanish students and Spanish students of gypsy ethnicity. No differences in Nature of Science views were found. Concerning grade level, third graders had more positive attitudes toward Science than fifth and sixth graders and more informed views of the tentative Nature of Science. Finally, no relation between Nature of Science views and attitudes towards Science were identified.

Conclusion: This study stress the need to address the steady decline in positive attitude toward Science and to improve students' views of Nature of Science from early elementary grades, and to use gender and culturally inclusive science teaching strategies.

KEYWORDS
Attitudes towards science; Nature of Science; elementary school; cultural backgrounds; TOSRA

CONTACT R. B. Toma rbtoma@ubu.es Departamento de Didácticas Específicas, Área de Didácticas de las Ciencias Experimentales, Universidad de Burgos, Burgos, España

Supplemental data for this article can be accessed here.

1. Introduction

Many studies have investigated attitudes towards science and views of Nature of Science (NOS) at secondary school level (Abd-El-Khalick and Lederman 2000; Chai et al. 2010; Kalman 2010; Khishfe and Lederman 2007; Vázquez-Alonso et al. 2014), although the same research questions at elementary grade levels are only approached in a few studies. Indeed, literature reviews by Osborne, Simon, and Collins (2003), in which the authors reviewed attitudes toward science studies, and Deng et al's. (2011) NOS literature review has identified few studies focusing on the elementary stage. For example, from among the 105 empirical studies examined by Deng et al. (2011), fewer than 13% included primary school students. Moreover, even more worrying is the exclusion of some social groups from these studies, especially in NOS-related research. In the case of the USA, after examining 112 NOS research studies, Walls (2016, 1557–1558) stated that '(...) white students and teachers emerged as the majority racial group under study' and that NOS research is '(...) operating completely oblivious to the narratives, voices, and stories of Black and other students and teachers of colour, instead relying on and valuing only the voices, narratives and stories provided by White participants'.

Hence, the aim of this study is to determine what attitudes towards science and views of NOS elementary students have, in the case of Spain. The aim is also to study how attitudes and NOS views may be different according to students' gender, cultural backgrounds and school levels. Finally, it is also intended to study whether the attitudes towards science among those students are related to their views of NOS. The results of this study can contribute to clarifying the effect of age, gender and cultural backgrounds on the development of attitudes toward science and NOS views in elementary school students. Additionally, this study can shed more light to the relationship between attitudes towards science and views of NOS, and may therefore be useful in guiding educational interventions focused on preventing the development of negative attitudes towards Science and naïve NOS views from early childhood. The improbability of changes in the aspirations of secondary school students is argued in a substantial corpus of recent studies. In fact, students aged between 10 and 14 years old are in a critical period when interventions can successfully focus on shaping student attitudes and aspirations towards science careers (Tai et al. 2006). Therefore, the elementary stage appears decisive for developing positive attitudes toward science and informed views on NOS. In the next section, the most relevant results from the literature related to elementary students' attitudes towards science and their views of NOS are presented. Information on the Spanish educational system and the cultural diversity of its students is also included.

2. Background

2.1. Studies on the attitudes towards science among elementary students

Research into the attitudes of students towards science points to different variables influencing their development during early childhood. These variables are mainly gender, age, and cultural backgrounds/nationality.

Regarding the first variable, most studies show that boys hold more positive attitudes and aspirations towards science than girls (Caleon and Subramaniam 2008; De Pro Bueno and Pérez Manzano 2014; Denessen et al. 2015; DeWitt and Archer 2015; George 2006; Hacıeminoğlu 2016; Vázquez and Manassero 2008). More specifically, girls appear to experience lower levels of enjoyment when learning science and technology (Denessen et al. 2015) and show greater indecision over their preferences for science-related careers (Caleon and Subramaniam 2008). In contrast, boys have stronger aspirations in science (DeWitt and Archer 2015). Nonetheless, despite boys having more positive attitudes towards science, George (2006) reported that the attitudes of boys in contrast to those of girls, diminishes as grade levels increase. Nevertheless, this gender effect remains a controversial question. Akpinar et al. (2009) found more interest in science among girl participants, and some studies had an absence of this gender effect in the attitudes of elementary students towards science (Khishfe and Boujaoude 2014; Said et al. 2016).

Some studies have revealed consistent data on age and grade level, showing a declining pattern in attitudes towards science as the age and grade level of the students increases (Akpinar et al. 2009; Ali et al. 2013; Denessen et al. 2015; DeWitt and Archer 2015; George 2006; Said et al. 2016; Vázquez Alonso and Manassero-Mas 2008). Students from upper elementary and middle school grades show negative attitudes towards the enjoyment of science lessons (Ali et al. 2013; DeWitt and Archer 2015), especially during 8th grade, when the sharpest decline in attitudes towards science takes place (Akpinar et al. 2009). Two studies reported that while attitudes towards the importance of science remained positive, attitudes towards school science decreased drastically over middle school and high school years (Akpinar et al. 2009; George 2006).

Several studies have explored attitudes toward science based on nationality and cultural background at secondary level (a comprehensive review can be found in Osborne, Simon, and Collins 2003), concluding that students belonging to different cultural groups hold differing perceptions on science, so ethnicity and cultural variables may appear to affect the formation of attitudes toward science. However, few studies have compared student attitudes towards science based on ethnicity or cultural background and existing results are contradictory. For example, DeWitt and Archer (2015) explored variables connected to science aspirations at elementary school level, finding that students from ethnic minority backgrounds appeared to show stronger aspirations for science careers. Khishfe and Boujaoude (2014) showed that students from developing countries appeared to have more positive attitudes towards science than those from developed countries, but Said et al. (2016) obtained opposite results with Qatari students. Although these results show some relation between student attitudes towards science and cultural backgrounds, it is not clear how ethnicity and cultural background variables may influence the development of student attitudes and further studies are needed to determine this relationship.

2.2. Studies on views of the nature of science among elementary students

At elementary level, most articles have focused on creative, tentative, empirical, and subjective constructs of NOS (i.e. Akerson et al. 2014; Cakici and Bayir 2012; Abd-El-Khalick, Masters and Akerson 2015; Quigley, Pongsanon, and Akerson 2010), reporting that, in general, students tend to hold naïve views about this NOS tenets (Cakici and Bayir 2012; Abd-El-Khalick et al. 2015; Khishfe 2008; Khishfe and Abd-El-Khalick 2002; Metin and Leblebicioğlu 2015), even at earlier stages of the educational system (Akerson and Donnely 2010; Akerson et al. 2014).

Few studies have explicitly examined NOS views based on gender, grade level and cultural backgrounds or nationality variables. Therefore, the little available evidence appears to indicate that there is no difference in the NOS views of boys and girls, who seem to both hold naïve views (Kang, Scharmann, and Noh 2005). The results are inconclusive when considering the grade levels. Some studies stated that at higher grade levels, students developed more informed conceptions of NOS especially in the empirical and tentative constructs (Hacıeminoğlu et al. 2015; Yoon, Suh, and Park 2014) and in the subjective, observational and inferential constructs (Yoon, Suh, and Park 2014). But, other studies have shown that students from higher grades have more naïve views on the imaginative and creative aspects of NOS than students from lower grades (Hacıeminoğlu et al. 2015). Others studies reported no clear differences between 6th, 8th and 10th graders NOS views, with the majority of students holding absolutist/empiricists views (Kang, Scharmann, and Noh 2005).

Finally, to the best of the authors' knowledge, there is no study at the elementary level that is directly focused on the impact of nationality and cultural background on the development of students' NOS views, particularly in the Spanish context. However, there are several NOS studies performed with students from different cultural backgrounds. Thus, Yoon, Suh, and Park (2014) explored the perceptions of five NOS constructs (empirical, tentative, observation, subjective, and creative NOS) among Korean students, reporting that as grade level increased, scores for empirical NOS decreased and more students perceived school science as passively listening to a lecture rather than performing lab activities. In addition, none of the 3rd graders held appropriate views on tentative, creative, and subjective aspects of the NOS, and only a few 7th and 10th grade students had more acceptable knowledge of the NOS. Kang, Scharmann, and Noh (2005) also studied Korean students' views of NOS, reporting that, in comparison with students from western countries, Korean students possessed a more pragmatic view of NOS. These results were explained by the characteristics of the Korean culture, which conceives science as '(…) one of the most effective means to be an internationally competitive country' (323–324). Finally, Walls (2012) reported that the 23 African-American students in the 3rd grade of elementary schooling in his study have also naïve views about NOS, since they viewed experimentation as a way for preparing potions and 'mixing colourful liquids', following specific steps, procedures, and rituals. However, the reasons for these views were inconclusive and were not specific to African-Americans students.

2.3. Cultural diversity in the Spanish educational system

Spain has experienced strong migratory flows over the past two decades. In 2016, a total of 4.618.581 immigrants were registered on the Spanish census, representing 9.92% of the total population (INE 2017). These results represent an increase of 8.32% of the foreign population compared to two decades ago. Thus, the Spanish education system has witnessed continuous growth in cultural diversity. Nowadays, 34.3% of the students enrolled in elementary education have an immigrant background (see Supplemental Figure 1), among who only 6.4% will continue on to study at middle and high school (MECD 2017). On the other hand, the gypsy population is estimated at around 725.000 to 750.000, all of whom are living in Spain, which correspond to approximately 1.6% of the total population (ERTF 2016).

In this study, cultural background refers to the sets of knowledge, beliefs and behavioural patterns shared by a social group that characterize them and that therefore differentiate them from other groups. We adopt the conceptualization of culture advanced by Bennet (1990, 64), in which culture is '(…) a system of shared knowledge and belief that shapes human perceptions'. Therefore, cultural backgrounds are a multidimensional system of ideas, public expressions, practices and behaviours of individuals from a particular group (Atran, Medin, and Ross 2005). The Royal Academy of the Spanish Language (RAE) uses the term 'Spanish' to refer to the people that are natives of Spain and the term 'gypsy' to refer to a '(…) person belonging to the people originating from India, spread over several countries, who is largely nomadic and has retained his own physical and cultural characteristics' (RAE 2014). On the other hand, people born in Spain with both parents being not natives of Spain are known as 2nd generation Spanish people or Spanish people with immigrant heritage. Therefore, in this study the following terms are used: Spanish students, to refer to students' natives of Spain that represents the dominant culture; Spanish student of gypsy ethnicity to refer to Spanish gypsy students, and 2nd generation Spanish students to refer to students that were born in Spain and have Spanish nationality but whom parents are both from another country, thus having an immigrant heritage and also pertaining to minority cultures. In this study, we will focus on second-generation immigrants from Eastern European countries (i.e. Spanish students of Romanian, Russian and Bulgarian heritage), because the nationalities of 2nd generation immigrant students enrolled in Elementary Education in Spain are mainly Romanian, Ukrainian, and Russian (MECD 2017).

Second generation Spanish students and Spanish students of gypsy ethnicity have cultural backgrounds (i.e. traditions, habits, religious and moral values, ideas and practices) that all differ from those adopted by the dominant group in the Spanish contexts. Mainly, Spanish people of gypsy ethnicity belong to minority groups at risk of social and educational exclusion, and their culture differs greatly from non-gypsy-Spanish citizens, in particular due to the use of a different language (see ERTF 2016). At the age of 17, more than 58% of Spanish students of gypsy ethnicity drop out of school, and fewer than 1 out of every 10 Spanish students of gypsy ethnicity enroll in high school (see Supplemental Figure 2) (EDUCACNIIE 2014), a percentage that is below the average of other European Union Member states (FRA 2012).

In relation to students with immigrant heritage, longitudinal studies conducted in both the USA (Rumbaut and Portes 2011) and in Spain (Gómez-Quintero and Fernández-Romero 2014; Portes, Aparicio Gómez, and Haller 2013) have shown that while some 2nd generation immigrant students tend to integrate well into the mainstream culture of the new country, others have difficulties with their inclusion in the new dominant culture. So, many 2nd generation Spanish students face educational disadvantage, presenting higher dropout rates, lower educational performance, and greater difficulties accessing higher education in comparison with their Spanish peers (Portes, Aparicio Gómez, and Haller 2013; EC 2008). Although foreign, 2nd generation Spanish students and Spanish students of gypsy ethnicity are not segregated in different schools in the Spanish educational system, there is a tendency towards segregation given that parents from the dominant culture tend to withdraw their children from schools with high percentages of immigrants, Spanish students with immigrant heritage or Spanish students of gypsy ethnicity,1 generating disparities between schools and increasing social and educational inequalities among the population over time (EC 2008).

To the best of the authors' knowledge, there is an absence of studies in the Spanish context on how different cultural backgrounds may affect the attitudes towards science of elementary students and their views of NOS. This gap in the literature also appears to exist at an international level, as pointed out in previous sections.

2.4. Elementary school science education in the Spanish context

The Spanish legislation on Science Education has undergone several changes over recent decades. The LOGSE (1990, *Ley Orgánica de Ordenación General del Sistema Educativo* [Organic Law on General Regulation of the Educational System]) outlined a subject curriculum at elementary school level called 'Knowledge of the Natural, Social and Cultural Environment', designed to improve scientific literacy. Some years later, a new law called the LOE (2006, *Ley Orgánica de Educación* [Organic Law of Education]) emphasized the need to provide science education based on the knowledge of children and using manipulative and reflective activities that foster child motivation for science-related issues. However, despite these demands from the educational system, in their description of the Spanish science education, Porlán et al. (2010) noted that teachers employ rote-learning as the predominant model for science education in Spain.

Recently, the urgent need for active teaching methodologies was highlighted in the new educational law called LOMCE (2013, *Ley Orgánicapara la Mejora de la Calidad Educativa* [Organic Law for the Improvement of Educational Quality]). This new law recognizes the importance of Science Education at an elementary stage, so it divides the subject of 'Knowledge of the Natural, Social and Cultural Environment' introduced in the LOGSE law into two different subjects called 'Social Sciences' and 'Natural Sciences', increasing the number of teaching hours assigned to science education, from one to two teaching hours per week. There is also specific content called 'Initiation in scientific inquiry' that demands teacher participation with students in inquiry-based activities in the school science syllabus of all Primary education grades. Nonetheless, there is still a lack of inquiry-based activities in science classrooms, as highlighted by Gil (2014), who reported that the most common practice during science classes was reading from textbooks and memorizing facts and laws, and that the activities for teaching science used least of all were planning and conducting experiments and research.

3. Design and methods

3.1. Purpose

As indicated earlier, the aims of this study are to determine, in the case of Spain, the attitudes of elementary students towards Science and to establish their views of NOS in relation to gender, cultural background, and school grade variables. Since most science education reform efforts have suggested the need to foster the development of informed conceptions of NOS to ensure the scientific literacy of students (McComas 2015; McComas and Olson 1998), the aim here is to find the relationship between attitudes towards science and views of NOS. These results would be of great interest in order to know which aspects educational interventions should focus in order to improve attitudes towards science. The research questions of this study are:

RQ 1. What effects do gender, cultural background and school grade have on elementary students' attitudes towards Science?

RQ 2. What effects do gender, cultural background, and elementary school grade level have on views of NOS among elementary students?

RQ 3. To what extent do attitudes towards science relate to views of NOS?

3.2. Sample and procedure

Participants in this study were drawn from six elementary schools in Burgos and one from Valladolid, Spain, by means of convenience sampling. Regarding cultural backgrounds, both Spanish, second-generation Spanish students and Spanish students of gypsy ethnicity were included in the sample of this study. Spanish students were recruited from five state schools located in the city of Burgos in middle to upper-class area, with a low percentage of second-generation immigrant students and immigrant students. 2nd generation Spanish students included in this study were born in Spain, but their parents were mostly from Eastern European countries, with both parents born outside Spain. These students were attending a school in Burgos with a high presence of multicultural students (more than 50% of students enrolled were foreigners or 2nd generation Spanish students). The Spanish students of gypsy ethnicity from this sample were attending a segregated school specifically for that ethnicity (100% of students enrolled in this school were of gypsy ethnicity) located in a peripheral zone of Valladolid, a city near Burgos.

Two scales, an adapted version of the Test of Science Related Attitudes (TOSRA) and the Nature of Science Instrument (NOSI) were administered at the beginning of the second semester of the 2015–2016 school year. Students had 60 minutes to answer all the items individually. The first author of this study was present to ensure the questionnaires were administered in the same way in each school. Subsequently, some participants were randomly selected for interview, broadening the quantitative data (n = 15). Although all students included in this sample could write and speak fluently in Spanish (the language in which the questionnaires were administered), some data were missing: no 2nd grade student was capable of completing the scales used in this study, while three 3rd and two 4th graders failed to answer more than half of the items in both scales, leaving blank answers. After discarding these cases (found in all the different groups considered), a total sample was obtained of 118 valid questionnaires from 64 boys and 54 girls. Table 1 shows the gender and the school grade of valid participants and Table 2 displays the gender and the cultural background of the valid participants.

In order to gain richer insights and to elaborate further the quantitative findings by allowing students whose literacy was poor to express themselves more fully, structured interviews were carried out with several students. A qualitative sample was randomly selected from each sample subgroup for the interviews (Patton 2002). Initially, the aim was to select 20% of the total number of participants. However, interviews could only be conducted at the schools during school time, a fact which considerably reduced the available time and the potential advantages of using interviews. Therefore, the qualitative sample was lower, comprising a total of 15 students enrolled in 5th and 6th grades (12.7% of the total sample), of whose 5 were Spanish (three girls), 4 were 2nd generation Spanish students with Romanian heritage (2 girls), and 6 were Spanish students of gypsy ethnicity (2 girls). One Spanish girl and two Spanish boys of gypsy ethnicity that participated in the interview were 5th graders; the rest of the qualitative sample was formed by 6th graders. Interviews were performed from 5 to 7 days after the administration of the quantitative scales.

3.3. Measures

3.3.1. Test of science related attitudes (TOSRA)

The first scale to be administered was the adapted Test of Science related Attitudes (TOSRA) (Fraser 1978, 1981), used to assess seven attitudes towards science constructs. This scale was adapted from the original 70 items to only 14 items (2 for each construct), which was thought more suitable for elementary students. First, all negatively formulated items were discarded, in order to enhance the understanding of participants. Then, from the remaining original items, two authors of this paper independently selected four items for each attitude towards the science constructs under consideration that were considered to be the closest to student's maturational stage. For example, rather than 'Public money spent on science in the last few years has been used wisely' (Fraser 1981, 15), 'Money spent on science is well worth

Table 1. Participants' gender and school year.

Gender	3rd grade	4th grade	5th grade	6th grade	Total
Male	5	11	16	32	64
Female	5	4	16	29	54
Total	10	15	32	61	118

Table 2. Participants' gender and cultural background.

Gender	Spanish	Second generation immigrants	Gypsy Spanish	Total
Male	33	22	9	64
Female	34	8	12	54
Total	67	30	21	118

spending' was preferred, due to its shorter length and less complicated vocabulary, especially in Spanish language. Cohen's Kappa index was calculated, obtaining moderate inter-ratter agreement (k = .75) on the selected items (McHugh 2012). Finally, the items were translated into Spanish and a pilot test was performed (*n* = 24; 4th grade students). Preliminary results showed student fatigue and scale administration was time demanding, so low-reliability items (Cronbach's α < .65) were removed, and a few items were re-phrased to form a final version of 14 total items (two per construct). The Cronbach's α of the final scale administered in this study was .80. This reduced version of the scale has already been used in the literature with Spanish speaking students, reporting reliable results (Toma and Greca 2018).

Attitudes towards science constructs included in the final scale were: (1) social implications of science (e.g. 'It is worth spending money on science so that scientists can make new discoveries'); (2) normality of scientists (e.g. 'Scientists are normal people who look like anyone else'); (3) attitude to scientific inquiry (e.g. 'It is better to discover the answer by experimentation rather than by asking the teacher'); (4) adoption of scientific attitudes (e.g. 'I am curious about the things around me and the world in which I live'); (5) enjoyment of science lessons (e.g. 'Science is the most interesting subject and I really like it'); (6) leisure interest in science (e.g. 'I would like to receive scientific materials to do experiments at home.'); and, (7) career interest in science (e.g. 'When I grow up, I want to study something related to science'). All items were measured on a five-point Likert scale, ranging from '1 = strongly disagree' to '5 = strongly agree'.

3.3.2. Nature of science instrument (NOSI)

The second scale was then administered, the Nature of Science instrument (NOSI), developed by Hacıeminoğlu, Yılmaz-Tüzün, and Ertepınar (2012) for use with elementary school students. This instrument measures students views related to the following NOS tenets: (1) Observation and inferences (e.g. 'Scientists believe that atoms exist from what they already know, but they cannot be certain'); (2) tentative NOS (e.g. 'Everything scientists say in books will never change'); (3) Empirical NOS (e.g. 'Scientists may reach different conclusions when looking at the same data'); and (4) Creative and imaginative NOS (e.g. 'Science could never involve imagination and creativity because this would result in incorrect or wrong findings'). All items were measured on a three-point Likert scale, where '1 = wrong', '2 = do not know', '3 = right'. The same group of 4th grade students (*n* = 24), as with the TORSA, participated in the pilot testing of the scale, after which four items were re-phrased. Cronbach's α was found to be .71, similar to the study by Hacıeminoğlu, Yılmaz-Tüzün, and Ertepınar (2012) where Cronbach's α was .76. In both cases, this value can be considered acceptable for preliminary research (Nunnally 1978).

3.3.3. Structured interviews

For attitude toward science, the first part of the interview consisted of asking the qualitative sample to elaborate on their answers to both items that relate to the

'Enjoyment of science lessons', so as to examine their reasoning and to relate it to the results obtained for this construct on the quantitative scale. Specifically, students were asked to (1) rate and value science subjects in comparison to other school subjects, and to (2) explain their answer to the items 'Science is the most interesting subject and I really like it' and 'We should have more science classes each week' that examines enjoyment of science lessons.

The second part of the interview, again seeking to gain deeper insights into students views about the NOS, consisted of administering a questionnaire based on the Views of Nature of Science questionnaire (Lederman et al. 2002). The questionnaire was administered orally to overcome literacy barriers derived from poor writing abilities, especially in Spanish students of gypsy ethnicity. Only items measuring the same dimensions as the quantitative scale (i.e. NOSI) were selected: (1) observation and inferential NOS (i.e. 'How do scientists develop new ideas that explain our world?'); (2) tentative NOS (i.e. 'After scientists have developed a scientific theory, does that theory ever change?'); (3) empirical NOS (i.e. 'What, in your view, is science?'); and (4) creative and imaginative NOS (i.e. 'Do scientists use their creativity and imagination during their investigations?').

3.4. Data analysis

Two types of analyses were performed. Attitudes towards science and views of NOS scales were analyzed through descriptive and inferencial tests using the SPSS v.24 statistical software package, and the interviews were analyzed by means of conventional content analysis (Hsieh and Shannon 2005). More specifically, in relation to the interviews, the transcripts were edited to facilitate analysis at sentence level. The responses of each student were grouped, through a process of decomposition, by interview question and similarity of response. Once the answers had been grouped under each interview question, subgroups of answers sharing similar information among the participants were created.

Statistical analyses were completed in three steps. First, descriptive analyses with both scales (NOSI and TOSRA) provided simple summaries on participants and measures (i.e. mean scores and standard deviation). Second, significant differences in data between participants by cultural background, school grades levels, and gender variables were analyzed with inferential statistics. Kolmogorov-Smirnov with Lilliefors correction and Shapiro-Wilks tests were run on both scales and their dimensions to calculate whether the variables were adjusted to a normal probability distribution. The results for the variable 'gender' were significant ($p < .001$) in both tests, suggesting a violation of the assumption of normality distribution. Therefore, the gender variable was analyzed using non-parametric statistics, and cultural background ($p = .217$) and grade level ($p = .854$) variables using parametric statistics.

Third, Pearson correlation and multiple regressions tests were used to measure the inter-relations between the NOSI dimensions and the TOSRA constructs.

4. Findings

4.1. RQ 1. What effects do gender, cultural background and school grade have on elementary students' attitudes towards science?

Having considered a maximum score of 5 (positive attitude) and a minimum of 1 (negative attitude), the participants of this study reported a range of indifferent to somewhat negative attitudes towards science. Analyzing TOSRA by its constructs, the highest scores were reported in 'social implication of science' and the lowest scores in 'career interest in science' and 'enjoyment of science lessons'. However, the standard deviation was quite high in both dimensions, indicating that some students were highly interested in a scientific career and thoroughly enjoyed science lessons, while others showed no interest at all in science careers and the science lessons were not enjoyable for them. Table 3 reports the descriptive results from the TOSRA scale.

4.1.1. Gender

In a gender comparison, boys scored higher in attitudes towards science compared to girls. Girls had more negative attitudes than boys in both 'enjoyment of science lessons' and 'career interest in science' dimensions. Analyzing gender effects, the Mann-Whitney U Test indicated that boys and girls differed significantly in their attitudes towards science as measured on the TOSRA scale. On average, boys had better attitudes towards science than girls (U = 1347.5; p = .04; effect-size r = .19).

4.1.2. Cultural background

Comparing by cultural backgrounds, 2nd generation Spanish students reported slightly better attitudes towards science than Spanish students and those of gypsy ethnicity. Second generation Spanish students enjoyed science lessons more than any others and they were also the ones who reported positive attitudes towards a career in science. Spanish students of gypsy ethnicity were the ones that enjoyed science lessons the least, and Spanish students showed the least interest in following a scientific career.

A one-way ANOVA was conducted to compare the effect of cultural background on student attitudes towards Science. The results indicated that Spanish, 2nd generation Spanish students, and Spanish students of gypsy ethnicity differed significantly in their

Table 3. TOSRA scale descriptive results.

	N	SI	NS	SI	SA	ES	LI	CI	TOT
All	118	3.40 (.66)[a]	2.83 (.89)	2.8 (.98)	3.26 (.76)	2.32 (1.10)	2.89 (.91)	2.3 (1.10)	2.84 (.47)
Girls	54	3.30 (.70)	2.74 (.82)	2.71 (1.00)	3.16 (.88)	2.21 (1.10)	2.83 (.91)	2.18 (1.10)	2.74 (.49)
Boys	64	3.45 (.63)	2.90 (.94)	2.90 (.95)	3.35 (.63)	2.41 (1.10)	2.94 (.92)	2.47 (1.00)	2.92 (.44)
3rd graders	10	3.10 (.69)	2.65 (.63)	3.50 (.85)	3.55 (.64)	3.6 (.46)	3.70 (.54)	3.20 (.79)	3.32 (.33)
4th graders	15	3.13 (.81)	2.17 (.75)	2.83 (.98)	3.33 (.59)	3.03 (.92)	3.33 (.90)	2.40 (1.24)	2.89 (.49)
5th graders	32	3.48 (.57)	2.88 (.90)	2.48 (1.01)	3.47 (.74)	2.10 (.87)	2.95 (.79)	2.30 (1.10)	2.81 (.44)
6th graders	61	3.48 (.65)	2.99 (.89)	2.87 (.93)	3.10 (.80)	2.10 (1.10)	2.61 (.92)	2.20 (1.01)	2.76 (.46)
Spanish	67	3.43 (.65)	3.07 (.87)	2.81 (.97)	3.28 (.78)	2.22 (1.10)	2.78 (.86)	2.11 (1.04)	2.81 (.45)
Foreign	30	3.33 (.71)	2.70 (.89)	3.10 (.82)	3.38 (.69)	2.87 (1.08)	3.25 (.93)	2.95 (1.00)	3.10 (.42)
Gypsy	21	3.38 (.67)	2.30 (.68)	2.43 (1.12)	3.05 (.80)	1.86 (.84)	2.74 (.96)	2.17 (.87)	2.66 (.45)

SI (social implication of science); NS (normality of scientists); SI (attitudes to scientific inquiry); SA (adoption of scientific attitudes); ES (enjoyment of science lessons); LI (leisure interest in science); CI (career interest in science); TOT (global results of TOSRA scale).

[a]*M* (SD): Mean (Standard deviation).

attitudes towards science, as measured with the TOSRA scale [$F(2, 115) = 8.763$; $p = .000$; $\eta_p^2 = .132$]. A Bonferroni post-hoc analysis indicated that the attitudes of 2nd generation Spanish students towards science were significantly more positive than the attitudes of both Spanish and Spanish students of gypsy ethnicity. However, no significant differences were found between the attitudes of these two groups towards science.

An in-depth scale analysis reported significant differences in three attitudes towards science constructs: Spanish, 2nd generation Spanish students and Spanish students of gypsy ethnicity differed significantly in their attitudes towards 'normality of scientists' [F (2, 115) = 7.574; $p = .001$; $\eta_p 2 = .116$], 'enjoyment of science lessons' [F (2, 115) = 6.574; $p = .002$; $\eta_p^2 = .103$] and 'career interest in science' [F (2, 115) = 7.588; $p = .001$; $\eta_p^2 = .117$]. A Bonferroni post-hoc analysis indicated that the attitudes of Spanish students towards 'normality of scientists' was significantly more positive than the attitudes of students of gypsy ethnicity; that the attitudes of 2nd generation Spanish students towards 'enjoyment of science lessons' were significantly more positive than the attitudes of Spanish and Spanish students of gypsy ethnicity; and that the attitudes of 2nd generation Spanish students towards 'career interest in science' were significantly more positive than the attitudes of Spanish and Spanish students of gypsy ethnicity. The Bonferroni Post-hoc results are reported in Table 4.

4.1.3. School grade

A one-way ANOVA was conducted to compare the effect of grade level on student attitudes towards science. The results indicated that student attitudes towards science measured on

Table 4. Bonferroni post-hoc results for TOSRA scale according to participants cultural background.

Dependent variable	Cultural background (I)	Cultural background (J)	Mean Difference	Std. Error	*p*. value
TOT	Spanish	Foreign	−.27	1	.022
		Gypsy	.26	.11	.068
	Foreign	Spanish	.27	1	.022
		Gypsy	.52	.13	.000
	Gypsy	Spanish	−.26	.11	.068
		Foreign	−.52	.13	.000
NS	Spanish	Foreign	.40	.19	.098
		Gypsy	.78	.21	.001
	Foreign	Spanish	−.40	.19	.098
		Gypsy	.38	.24	.346
	Gypsy	Spanish	−.78	.21	.001
		Foreign	−.38	.24	.346
ES	Spanish	Foreign	−.64	.23	.017
		Gypsy	.37	.26	.478
	Foreign	Spanish	.64	.23	.017
		Gypsy	1	.29	.003
	Gypsy	Spanish	−.37	.26	.478
		Foreign	−1	.29	.003
CI	Spanish	Foreign	−.84	.22	.001
		Gypsy	−.05	.25	1
	Foreign	Spanish	.84	.22	.001
		Gypsy	.78	.29	.021
	Gypsy	Spanish	.05	.25	1
		Foreign	−.78	.29	.021

TOT (global results of TOSRA scale); NS (normality of scientists); ES (enjoyment of science lessons); CI (career interest in science) the TOSRA scale differed significantly across different grade levels [F (3, 114) = 4.62; p = .004; η_p^2 = .108]. A Bonferroni post-hoc analysis indicated that the attitudes towards science of students in the 3rd grade were significantly more positive than the attitudes of students in 5th and 6th grades.

Further analysis showed that, according to grade levels, student attitudes towards 'normality of scientists' [F (3, 114) = 3.879; p = .011; η_p^2 = .093], 'scientific inquiry' [F (3, 114) = 3.058; p = .031; η_p^2 = .074], 'enjoyment of science lessons' [F (3, 114) = 10.554; p = .000; η_p^2 = .217], 'leisure interest in science' [F (3, 114) = 6.504; p = .000; η_p^2 = .146] and 'career interest in science' [F (3, 114) = 2.722; p = .048; η_p^2 = .067] differed significantly. A Bonferroni post-hoc analysis showed that the attitudes of 6th graders towards 'normality of scientists' was significantly more positive than the attitudes of 4th graders; that the attitudes of students in the 3rd grade towards 'scientific inquiry' was significantly more positive than the attitudes of 5th graders; that students in the 3rd grade enjoyed significantly more science lessons than 5th graders and 6th graders, and the same was true between 4th graders and 5th and 6th grade students; that the attitudes of 3rd graders towards 'leisure interest in science' were significantly more positive than the attitudes of 6th grade students, and the same held true between the attitudes of students in their 4th and in their 6th grade; and finally, that the attitudes of students in the 3rd grade towards 'career interest in science' were significantly more positive than the attitudes of students in their 6th grade. The Bonferroni Post-hoc results are reported in Table 5.

4.1.4. Summary of research question 1

In relation to the first research question, the results show that gender, age and cultural background variables have a significant impact on students' attitudes towards science, favouring students in lower grades of elementary education, 2nd generation Spanish students, and boys in general, who have shown more favourable attitudes towards science, especially in the 'enjoyment of science lesson' attitudinal construct.

Given the low number of students that were interviewed, the interview findings cannot be linked with any significance to gender, grade level or cultural background variables. However, they are presented in order to offer further insight into some aspects of the TOSRA scale. Thus, only the interview results related to the construct 'enjoyment of science lessons' are reported here, because attitudes towards the enjoyment of science lessons are far less positive as grade level increases, so much so that none of the participants interviewed (n = 15) confessed to enjoying their science lessons. Most students (n = 11) said that the science subject was their least favourite and that they were not on good terms with their science teachers. Arguments against science as a subject were: science lessons are boring (10 out of 15 students), a lot of homework (12 out of 15 students), and no experimentation or practical activities during lessons (12 out of 15 students). Three students reported that they usually carried out experiments at school. However, these experiments were either very simple, consisted of only watching the teacher performing the experiment, or involved 'things that I have already done at home when I was a child'. Finally, some students said that they would like to go to science laboratory instead of studying science in their conventional classroom setting. These further insights from the interview tend to support a view that science teaching at elementary level in this sample of schools in Spain is still very traditional and teacher-centred.

Table 5. Bonferroni Post-Hoc results for TOSRA scale according to participants grade level.

Dependent Variable	Grade (I)	Grade (J)	Mean Difference	Std. Error	*p*. value	Dependent Variable	Grade (I)	Grade (J)	Mean Difference	Std. Error	*p*. value
TOT	3rd	4th	.43	.18	.124	ES	3rd	4th	.57	.40	.933
		5th	.52	.16	.012			5th	1.52	.35	.000
		6th	.56	.15	.002			6th	1.53	.33	.000
	4th	3rd	−.43	.18	.124		4th	3rd	−.57	.40	.933
		5th	.08	.14	1			5th	.96	.30	.013
		6th	.13	.13	1			6th	.97	.28	.005
	5th	3rd	−.52	.16	.012		5th	3rd	−1.52	.35	.000
		4th	−.08	.14	1			4th	−.96	.30	.013
		6th	.05	1	1			6th	.01	.21	1
	6th	3rd	−.56	.15	.002		6th	3rd	−1.53	.33	.000
		4th	−.13	.13	1			4th	−.97	.28	.005
		5th	−.05	1	1			5th	−.01	.21	1
NS	3rd	4th	.48	.35	1	LI	3rd	4th	.37	.35	1
		5th	−.23	.31	1			5th	.75	.31	.105
		6th	−.34	.29	1			6th	1.09	.29	.002
	4th	3rd	−.48	.35	1		4th	3rd	−.37	.35	1
		5th	−.71	.27	.057			5th	.38	.27	.948
		6th	−.83	.25	.007			6th	.72	.25	.026
	5th	3rd	.23	.31	1		5th	3rd	−.75	.31	.105
		4th	.71	.27	.057			4th	−.38	.27	.948
		6th	−.12	.19	1			6th	.34	.19	.434
	6th	3rd	.34	.29	1		6th	3rd	−1.09	.29	.002
		4th	.83	.25	.007			4th	−.72	.25	.026
		5th	.12	.19	1			6th	−.34	.19	.434
SI	3rd	4th	.67	.39	.540	CI	3rd	4th	.80	.42	.366
		5th	1	.35	.024			5th	.90	.38	.106
		6th	.63	.33	.332			6th	1	.35	.032
	4th	3rd	−.67	.39	.540		4th	3rd	−.80	.42	.366
		5th	.35	.30	1			5th	.10	.32	1
		6th	−.04	.28	1			6th	.20	.30	1
	5th	3rd	−1	.35	.024		5th	3rd	−.90	.38	.106
		4th	−.35	.30	1			4th	−.10	.32	1
		6th	−.38	.21	.407			6th	.10	.23	1
	6th	3rd	−.63	.33	.332		6th	3rd	−1	.35	.032
		4th	.04	.28	1			4th	−.20	.30	1
		5th	.38	.21	.407			6th	−.10	.23	1

TOT (global results of TOSRA scale); NS (normality of scientists); SI (attitudes to scientific inquiry); ES (enjoyment of science lessons); LI (leisure interest in science); CI (career interest in science)

4.2. RQ 2. What effects do gender, cultural background, and elementary school grade level have on views of NOS among elementary students?

According to the NOSI scale, participants are poorly informed about NOS-related constructs. Regarding the NOSI-related dimensions, more informed views of NOS were obtained in the empirical NOS dimension and more naïve views in the tentative NOS construct. Table 6 reports the descriptive results from the NOSI scale.

4.2.1. Gender

Boys scored lower on NOS constructs compared to girls. The Mann-Whitney U test showed no significant differences in NOS views by student gender when measured on the NOSI scale (U = 1650; p = .671; effect-size r = −.03). However, analyzing each NOSI scale construct, the results showed that girls had significantly more informed views of empirical NOS than boys (U = 1336.5; p = .026; effect-size r = .19).

Table 6. NOSI scale descriptive results.

	N	Observation & inferences	Tentative	Empirical	Imagination & creativity	Total
All	118	2.22 (.48)[a]	1.99 (.63)	2.62 (.41)	2.10 (.29)	2.20 (.22)
Girls	54	2.14 (.49)	1.96 (.58)	2.70 (.39)	2.10 (.31)	2.21 (.22)
Boys	64	2.30 (.47)	2.01 (.65)	2.55 (.41)	2.10 (.27)	2.20 (.21)
3rd graders	10	2.15 (.53)	2.57 (.45)	2.20 (.55)	2.10 (.30)	2.23 (.22)
4th graders	15	1.97 (.35)	2.27 (.59)	2.29 (.47)	2.10 (.40)	2.16 (.23)
5th graders	32	2.22 (.46)	1.84 (.52)	2.73 (.30)	2.19 (.26)	2.24 (.22)
6th graders	61	2.30 (.50)	1.90 (.63)	2.72 (.33)	2.01 (.25)	2.19 (.21)
Spanish	67	2.28 (.44)	1.84 (.61)	2.73 (.34)	2.08 (.28)	2.21 (.20)
Foreign	30	2.27 (.54)	2.12 (.58)	2.40 (.50)	2.07 (.31)	2.19 (.26)
Gypsy	21	2.00 (.50)	2.25 (.59)	2.60 (.36)	2.07 (.30)	2.22 (.20)

[a]*M* (SD): Mean (Standard deviation).

4.2.2. Cultural background

NOS constructs were very similar between participants with different cultural backgrounds, with the Spanish students obtaining slightly higher scores. A one-way ANOVA was conducted to compare the effect of cultural background on students' views of NOS, obtaining that cultural background variables have not a significant impact on students NOS views, neither in the overall score nor in each NOSI dimensions [$F(2, 115) = .150$; $p = .861$; $\eta_p^2 = .003$].

4.2.3. School grade

Views on NOS remained very similar across grade levels, with a slight decrease in scores in the older grade levels. Interestingly, as the grade level increased, students gave lower ratings to the tentativeness of science and rated its empirical character more highly. Table 4 shows the descriptive results for the NOSI scale. A one-way ANOVA was conducted to compare the effect of grade level on student views of NOS. No effect of grade level on student views of NOS was found [F (3, 114) = .570; $p = .636$; $\eta_p^2 = .015$]. However, separate analysis of each NOSI dimension, by grade levels, showed that views of the tentative NOS construct differed significantly among students [F (3, 114) = 5.615; $p = .0001$; $\eta_p^2 = .129$). Thus, the Bonferroni post-hoc test reported that the tentative views of NOS among students in the 3rd grade were significantly more informed than the views of students in the 5th and 6th grade. The Bonferroni Post-hoc results are reported in Table 7.

4.2.4. Summary of research question 2

In relation to the second research question, the results show that gender and school year have a significant impact on student's views of NOS. Cultural backgrounds variables have found to not significantly impact students' NOS views. Thus, girls have more informed views about the empirical aspect of NOS and that students developed more naïve views about the tentative aspect of NOS as grade level increase. Content analysis of student responses during the interview supported the naïve views of NOS reflected on the NOSI scale. In relation to empirical NOS (i.e. 'What, in your view, is science?'), students held mixed views and considered science as a tool used by people to learn about the world and nature. Students' defined science as ways used to 'learn about nature, plants and trees', to 'make discoveries' or to 'investigate things that are very

Table 7. Bonferroni post-hoc results for tentative NOS dimension according to participants grade level.

Grade (I)	Grade (J)	Mean Difference	Std. Error	*p*. value
3rd graders	4th graders	.30	.24	1
	5th graders	.73	.21	.004
	6th graders	.67	.20	.007
4th graders	3rd graders	−.30	.24	1
	5th graders	.43	.18	.116
	6th graders	.37	.17	.192
5th graders	3rd graders	−.73	.21	.004
	4th graders	−.43	.18	.116
	6th graders	−.07	.13	1
6th graders	3rd graders	−.67	.20	.007
	4th graders	−.37	.17	.192
	5th graders	.07	.13	1

difficult for most people'. Student responses to observation and inferential NOS and to creative and imaginative NOS revealed naïve constructs. For example, the answers of respondents to the interview questions 'How do scientists develop new ideas that explain our world?', and 'Do scientists use their creativity and imagination during their investigations?', students consistently referred to experimentation as the basis of scientific knowledge and that experiments are performed following some specific steps and techniques that are unique to scientists and that must be learned before becoming a scientist. For example, one student said that scientists 'look at some cells of an animal to see if they can invent a medicine or something to cure diseases. They go to their laboratories; when they discover something, they call everybody'. Another participant argued that scientists 'use animals as guinea pigs to make medicines'. Finally, students consider scientific knowledge as 'discoveries never seen before' and that 'no one knows about', which suggest naïve conceptions because it considers science as mysterious discovering or objective truths that are not influenced by social, cultural or personal aspects like creativity or imagination.

4.3. RQ 3. To what extent do attitudes towards science relate to views of NOS?

A Pearson correlation was performed to establish whether student attitudes towards science and their views of NOS were related: no significant relation [r(118) = .225, $p > .05$] was found. A multiple regression was conducted to see whether NOS constructs could predict students' attitudes towards science. Here, it was found that student views of NOS explained a non-significant amount of the variance in student attitudes towards science [$F = (4, 113) = 1.158$, $p = .333$; $R^2_{adjusted} = .005$). The analysis showed that none of the NOSI scale dimensions do significantly predict the values of student attitudes towards science: empirical NOS (Beta = −.076; *t*(−.811); $p = .419$); observation and inferential NOS (Beta = .128; *t*(1.375); $p = .172$); tentative NOS (Beta = .105; *t*(1.117); $p = .266$); and creative and imaginative NOS (Beta = .031; *t*(.335); $p = .739$).

5. Discussion

This study reports results related to students' attitudes towards Science and their views of Nature of Science according to gender, cultural background and grade level variables. Taken together, these results suggest that gender, cultural background, and school grade had an impact on student attitudes towards science. For example, boys from this sample had better attitudes towards science than girls; 2nd generation Spanish students reported more enjoyment of science lessons and more interest in science-related careers than Spanish and Spanish students of gypsy ethnicity; and at higher grade levels, student attitudes towards science decreased, especially after 3rd grade of elementary education. Specifically, students from higher grade level reported less enjoyment of science lessons, less interest in science-related activities during leisure time, and fewer aspirations to follow a scientific career. However, the effect size of the differences was small in all cases. This result suggests that in our sample, gender, cultural background and school grade variables do not have as much impact on student attitudes towards science as might be expected or reported in other studies (DeWitt and Archer 2015; Khishfe and Boujaoude 2014). However, gender results from this study are similar to previous results obtained in the literature showing better attitudes towards science in boys (Caleon and Subramaniam 2008; De Pro Bueno and Pérez Manzano 2014; Denessen et al. 2015; DeWitt and Archer 2015; George 2006; Hacıeminoğlu 2016; Vázquez Alonso and Manassero-Mas 2008). Nevertheless, the differences found in this study between students' attitudes are not as great as in the studies cited above. As for cultural background, the results reported in this study relating to 2nd generation Spanish students are similar to those previously reported in the literature showing that ethnic minority students have better attitudes towards scientific careers (DeWitt and Archer 2015). However, it must be highlighted that the cultural background and the nationality of the students included in this study differ from those students in the DeWitt and Archer (2015) study; therefore, this comparison should be made with caution until new research can further clarify these findings. Still, in this study, the attitudes of 2nd generation Spanish students towards scientific careers in this sample appear to confirm that students from developing countries have higher science-related aspirations (2nd generation Spanish students from our sample were mostly from east-European countries: e.g. Romania, Bulgaria), as stated previously by Khishfe and Boujaoude (2014).

The results of attitudes towards science according to grade level are similar to previous results in the literature, confirming the steady decline in positive attitudes towards Science as grade levels increase (Akpinar et al. 2009; Ali et al. 2013; Denessen et al. 2015; DeWitt and Archer 2015; George 2006; Said et al. 2016; Vázquez Alonso and Manassero-Mas 2008). Participants studying elementary grades in Spain appear not to enjoy science lessons, in a similar way to UK students (DeWitt and Archer 2015). The interview results showed dissatisfaction among elementary students in relation to science teaching methodology and the activities performed during science lessons. This finding is similar to Hacıeminoğlu's (2016) study where the author stated that rote-learning might lead to negative attitudes towards science. So, although children appeared to begin school with positive attitudes towards science, as they progressed in schooling, they began to lose interest, probably due to traditional science-teaching methodology that is teacher-centered and that relies mainly on textbooks and 'chalk-and-talk' methods.

Regarding NOS, this study has shown that gender and cultural background have no effect on student views of NOS. In general, students included in this study showed naïve views of NOS, similar to those obtained by previous studies (e.g. Yoon, Suh, and Park 2014; Quigley, Pongsanon, and Akerson 2010). Comparing this results with other studies using the NOSI instrument (Hacıeminoğlu, Yılmaz-Tüzün, and Ertepınar 2012), sixth-grader Spanish students held views on NOS that were more informed than the views of Turkish students in terms of 'Empiricism' and 'Observation and Inference' and more naïve in terms of 'Tentativeness' and 'Imagination and Creativity'. An interesting result from our study is that, although no significant differences were found in students general NOS views based on grade level, students from upper grades had less informed views on 'tentative' NOS than students from lower grades, which may be a result of the use of traditional teaching strategies and teachers lack of NOS knowledge. As stated in previous studies that explored NOS instruction for improving student understanding of NOS (i.e. Akerson and Donnely 2010; Akerson et al. 2014; Abd-El-Khalick et al. 2015), NOS constructs are more likely to change when explicit-reflective NOS strategies are used during science lessons. However, in Spain, elementary teachers do not have specific training in science, and their knowledge of scientific issues and science-teaching strategies is very limited (García-Carmona and Acevedo Díaz 2016; Toma, Greca, and Meneses Villagrá 2017; Vázquez-Alonso et al. 2014). Given this lack of knowledge, teachers are unlikely to include more informed NOS views during their science lessons, which may contribute to no improvements in the naïve NOS views of students throughout elementary education.

According to the correlation analysis, there seems to be no relation between student views of Nature of Science and their attitudes towards Science. Although the results are inconclusive, it seems that both attitudes towards Science and NOS views develop independently throughout Elementary education grades. Further studies should explore this relationship including a more representative sample and also specific NOS and attitudes instruments developed and validated using Spanish-speaking students in order to obtain more valid and reliable results.

There are some limitations that may be found in this study. First, the sample was quite small and students' distribution by grade level and cultural background was irregular. A larger sample, with similar numbers of students at each grade level and from each cultural background would be preferable. The authors also acknowledge that grouping students according to cultural background may be limited, especially in the case of 2nd generation Spanish students. Future investigation is suggested to address this limitation. Second, the greater part of the data was collected from responses to self-administered scales, which may not be ideal for lower grade students. An attempt was made to overcome this limitation by performing interviews with selected students. However, due to their availability and the time constraints imposed by school centres, the authors were unable to take maximum advantage of the interviews. Even so, the results from the interviews appear to support the results obtained from the self-administrated questionnaires.

6. Implications for science education

There are several implications for Science education that are linked to this study. Firstly, there is concern over the steady decline in positive attitudes towards science that elementary students experience as they become older and their grade-levels at school increase. In this regard, numerous studies have pointed to the need for more active teaching pedagogies. A meta-analysis revealed that problem-based learning is an effective teaching strategy for fostering positive attitudes towards school science among students (Demirel and Dağyar 2016). More recently, Hellgren and Lindberg (2017) concluded that authentic experiences that create opportunities for students to connect school science contexts to authentic science may improve the motivation of students towards learning science.

Secondly, gender inclusion strategies to reduce the gender gap (Scutt et al. 2013) should be included during science classes. Many studies argue that Science pedagogy and curricula need to be adapted in order to address the interests and learning methods of girls and to improve their self-efficacy in Science (Baker 2013). Jones et al. (2000) pointed out that girls are less competitive than boys, tend to follow the instructions of the teacher and classroom guidelines more than boys, and that girls may benefit more from cooperative-based learning activities. Additionally, gender discrepancies in science education are more likely to be reduced by offering similar experiences and opportunities to both girls and boys, and by engaging them in laboratory-type and hands-on activities (Cavallo and Laubach 2001). Also, gender-inclusive science curricula tend to include real-life context-based activities that foster collaboration and communication between peers in supportive environments, as pointed out by Brotman and Moore (2008) in their review of science education literature related to the inclusion of girls in science. Additionally, preventive educational measures should be adopted to assist girls from an early age in shaping their self-efficacy, academic performance and expectations of success in science, and in improving their scientific identity through mentoring programs and exposure to scientific role models (Cadaret et al. 2017).

Thirdly, teachers should plan science lessons acknowledging the differences that arise from cultural and gender variables between students, otherwise science reforms may not fit all student needs. For example, science teachers should foster inclusive participation of both boys and girls, ensuring that girls do not feel inhibited by boys and feel that they can participate in hand-on activities (Parker and Rennie 2002). In addition, teachers should not pass on unconscious messages about gender expectations in science achievement, as past research reported that teachers' expectations influence girls' performance (Elwood 2005) and they also should plan their classes attending to the possible existing differences in learning between boys and girls. For example, Chetcuti (2009) found that science teachers consider that boys tend to be more participative and engage more in hands-on activities, and that girls are less competitive and prefer to actively listen to explanations. In relation to culturally inclusive science teaching, Yoon, Kim, and Martin (2016) proposed a culturally inclusive science teaching model (CIST) that could be used in classes with culturally and linguistically diverse students as found in this study.

Finally, the results of this study have stressed the need to improve knowledge of the Nature of Science from early educational grades. In this regard, there is evidence on the effectiveness of an explicit-reflective instruction approach for this matter (Akerson and Donnely 2010; Akerson et al. 2014; Cakici and Bayir 2012; Abd-El-Khalick et al. 2015; Khishfe 2008; Quigley, Pongsanon, and Akerson 2010).

7. Conclusions

The aim of this study was to contribute to the efforts in determining the influence of cultural background, gender, and grade-level on the attitudes of elementary students towards science and on their views of NOS. Identifying and understanding the relationship between these traits will be vital for attaining scientific literacy for all students.

Finally, the need to investigate the attitudes of elementary school students towards science in greater depth has also been stressed, especially attitudinal constructs related to the enjoyment of science lessons and leisure interests in science; two constructs that assist scientific and technological vocation and that may awaken science-related career aspirations.

Acknowledgments

This study was partially funded by the Spanish Ministry of Economy and Competitiveness through the research project MINECO EDU2017-89405-R.

We would like to acknowledge the reviewers for their comments and suggestions that have contributed substantially to improving this work.

Disclosure statement

No potential conflict of interest was reported by the authors.

Funding

This work was supported by the Ministerio de Economía y Competitividad [EDU2017-89405-R].

ORCID

R. B. Toma http://orcid.org/0000-0003-4846-7323
I. M. Greca http://orcid.org/0000-0003-3674-7985
M. L. Orozco Gómez http://orcid.org/0000-0001-5547-8712

References

Abd-El-Khalick, F., and N. G. Lederman. 2000. “The Influence of History of Science Courses on Students’ Views of Nature of Science.” *Journal of Research in Science Teaching* 37 (10): 1057–1095. doi:10.1002/(ISSN)1098-2736.

Abd-El-Khalick, F., H. Masters, and V. L. Akerson. 2015. “Using History of Science to Teach Nature of Science to Elementary Students.” *Science and Education* 24 (9–10): 1103–1140. doi:10.1007/s11191-015-9783-5.

Akerson, V., and L. A. Donnely. 2010. “Teaching Nature of Science to K2 Students: What Understandings Can They Attain?.” *International Journal of Science Education* 32 (1): 97–124. doi:10.1080/09500690902717283.

Akerson, V. L., V. Nargund-Joshi, I. Weiland, K. Pongsanon, and B. Avsar. 2014. “What Third-Grade Students of Differing Ability Levels Learn about Nature of Science after a Year of Instruction.” *International Journal of Science Education* 36 (2): 244–276. doi:10.1080/09500693.2012.761365.

Akpinar, E., E. Yildiz, N. Tatar, and Ö. Ergin. 2009. “Students’ Attitudes toward Science and Technology: An Investigation of Gender, Grade Level, and Academic Achievement.” *Procedia - Social and Behavioral Sciences* 1 (1): 2804–2808. doi:10.1016/j.sbspro.2009.01.498.

Ali, M. M., R. Yager, E. Hacieminoglu, and I. Caliskan. 2013. “Changes in Student Attitudes regarding Science When Taught by Teachers without Experiences with a Model Professional Development Program.” *School Science and Mathematics* 113 (3): 109–119. doi:10.1111/ssm.12008.

Atran, S., D. Medin, and N. Ross. 2005. "The Cultural Mind: Environmental Decision Making and Cultural Modeling within and across Populations." *Psychological Review* 112 (49): 744–776. doi:10.1037/0033-295X.112.4.744.

Baker, D. 2013. "What Works: Using Curriculum and Pedagogy to Increase Girls' Interest and Participation in Science." *Theory into Practice* 52: 14–20. doi:10.1080/07351690.2013.743760.

Bennet, C. I. 1990. *Comprehensive Multicultural Education: Theory and Practice*. Boston: Allya & Bacon.

Brotman, J. S., and F. M. Moore. 2008. "Girls and Science: A Review of Four Themes in the Science Education Literature." *Journal of Research in Science Teaching* 45 (9): 961–1002. doi:10.1002/tea.20241.

Cadaret, M. C., P. J. Hartung, L. M. Subich, and I. K. Weigold. 2017. "Stereotype Threat as a Barrier to Woman Entering Engineering Careers." *Journal of Vocational Behavior* 99: 40–51. doi:10.1016/j.jvb.2016.12.002.

Cakici, Y., and E. Bayir. 2012. "Developing Children's Views of the Nature of Science through Role Play." *International Journal of Science Education* 34 (7): 1075–1091. doi:10.1080/09500693.2011.647109.

Caleon, I. S., and R. Subramaniam. 2008. "Attitudes Towards Science of Intellectually Gifted and Mainstream Upper Primary Students in Singapore." *Journal of Research in Science Teaching* 45 (8): 940–954. doi:10.1002/tea.20250.

Cavallo, A. M. L., and T. A. Laubach. 2001. "Students' Science Perceptions and Enrollment Decisions in Differing Learning Cycle Classrooms." *Journal of Research in Science Teaching* 38 (9): 1029–1062. doi:10.1002/tea.1046.

Chai, C. C., F. Deng, Y. Y. Qian, and B. Wong. 2010. "South China Education Major's Epistemological Beliefs and Their Conceptions of Nature of Science." *The Asia-Pacific Education Researcher* 19 (1): 111–125.

Chetcuti, D. 2009. "Identifying a Gender Inclusive Pedagogy from Maltese Teachers' Personal Practical Knowledge." *International Journal of Science Education* 31 (1): 81–99. doi:10.1080/09500690701647996.

De Pro Bueno, A., and A. Pérez Manzano. 2014. "Actitudes de los alumnos de Primaria y Secundaria ante la visión dicotómica de la ciencia." *Enseñanza de Las Ciencias* 32 (3): 111–132. doi:10.5565/rev/ensciencias.1015.

Demirel, M., and M. Dağyar. 2016. "Effects of Problem-Based Learning on Attitude: A Meta-Analysis Study." *Eurasia Journal of Mathematics, Science & Technology Education* 12 (8): 2115–2137. doi:10.12973/eurasia.2016.1293a.

Denessen, E., N. Vos, F. Hasselman, and M. Louws. 2015. "The Relationship between Primary School Teacher and Student Attitudes Towards Science and Technology." *Education Research International* 2015: 1–7. doi:10.1155/2015/534690.

Deng, F., D.-T. Chen, -C.-C. Tsai, and C. S. Chai. 2011. "Students' Views of the Nature of Science: A Critical Review of Research." *Science Education* 95 (6): 961–999. doi:10.1002/sce.20460.

DeWitt, J., and L. Archer. 2015. "Who Aspires to A Science Career? A Comparison of Survey Responses from Primary and Secondary School Students." *International Journal of Science Education* 37 (13): 2170–2192. doi:10.1080/09500693.2015.1071899.

EC (European Commission). 2008. *Green Paper. Migration & Mobility: Challenges and Opportunities for EU Education Systems*. Brussels: Commision of the European Communities.

EDUCACNIIE. 2014. "Boletín de Educación, Número 3." MECD. http://educalab.es/documents/10180/62610/Boletin3_Febrero2014.pdf/96a096d2-6b3d-4bab-805f-9465e9df6025

Elwood, J. 2005. "Gender and Achievement: What Have Exams Got to Do with It?" *Oxford Review of Education* 31 (3): 373–393. doi:10.1080/03054980500222031.

ERTF (European Roma and Travellers Forum). 2016. *Ficha Informativa Sobre La Situación De Los Romá/Gitanos En España* [Information Sheet on the Situation of Roma/Gypsies in Spain]. Accessed: http://presenciagitana.org/160131_SituacionRoma_ERTF_ES.pdf

FRA (Agency for Fundamental Rights). 2012. *The Situation of Roma in 11 EU Member States Survey Results at a Glance*. Luxembourg: Publications Office of the European Union.

Fraser, B. J. 1978. “Development of a Test of Science-Related Attitudes.” *Science Education* 62 (4): 509–515. doi:10.1002/sce.3730620411.

Fraser, B. J. 1981. *Test of Science-Related Attitudes*. Melbourne: Australian Council for Educational Research.

García-Carmona, A., and J. A. Acevedo Díaz. 2016. “Concepciones de estudiantes de profesorado de educación primaria sobre la Naturaleza de la Ciencia.” *Revista Mexicana De Investigación Educativa* 21 (69): 583–610.

George, R. 2006. “A Cross-Domain Analysis of Change in Students’ Attitudes toward Science and Attitudes about the Utility of Science.” *International Journal of Science Education* 28 (6): 571–589. doi:10.1080/09500690500338755.

Gil, F. 2014. “Metodologías didácticas empleadas en las clases de ciencias y su contribución a la explicación del rendimiento.” *Revista de Educación* 366: 190–214. doi:10.4438/1988-592X-RE-2014-366-271.

Gómez-Quintero, J. D., and C. Fernández-Romero. 2014. “Familias inmigrantes en España: Estructura socio demográfica, roles de género y pautas culturales de los hijos adolescentes.” *Papeles de Población* 20 (80): 87–118.

Hacıeminoğlu, E. 2016. “Elementary School Students’ Attitude toward Science and Related Variables.” *International Journal of Environmental and Science Education* 11 (2): 35–52. doi:10.12973/ijese.2016.288a.

Hacıeminoğlu, E., H. Ertepinar, Ö. Yilmaz-Tüzün, and H. Çakir. 2015. “Students and School Characteristics Related to Elementary School Students’ Views of the Nature of Science.” *Education* 3-13 (43): 700–721. doi:10.1080/03004279.2013.865655.

Hacıeminoğlu, E., Ö. Yılmaz-Tüzün, and H. Ertepınar. 2012. “Development and Validation of Nature of Science Instrument for Elementary School Students.” *Education 3-13* 42 (3): 258–283. doi:10.1080/03004279.2012.671840.

Hellgren, J. M., and S. Lindberg. 2017. “Motivating Students with Authentic Science Experiences: Changes in Motivation for School Science.” *Research in Science & Technological Education* 35 (4): 409–426. doi:10.1080/02635143.2017.1322572.

Hsieh, H.-F., and S. E. Shannon. 2005. “Three Approaches to Qualitative Content Analysis.” *Qualitative Health Research* 15 (9): 1277–1288. doi:10.1177/1049732305276687.

INE (Instituto Nacional de Estadística). (2017). *Principales series de población desde 1998* [Main population series since 1998]. http://www.ine.es/

Jones, M. G., L. Brader-Araje, L. W. Carboni, G. Carter, M. J. Rua, E. Banilower, and H. Hatch. 2000. “Tool Time: Gender and Student’s Use of Tools, Control, and Authority.” *Journal of Research in Science Teaching* 37 (8): 760–783. doi:10.1002/(ISSN)1098-2736.

Kalman, C. 2010. “Enabling Students to Develop a Scientific Mindset.” *Science & Education* 19: 147–163. doi:10.1007/s11191-009-9186-6.

Kang, S., L. C. Scharmann, and T. Noh. 2005. “Examining Students’ Views on the Nature of Science: Results from Korean 6th, 8th, and 10th Graders.” *Science Education* 89 (2): 314–334. doi:10.1002/sce.20053.

Khishfe, R. 2008. “The Development of Seventh Graders’ Views of Nature of Science.” *Journal of Research in Science Teaching* 45 (4): 470–496. doi:10.1002/tea.20230.

Khishfe, R., and F. Abd-El-Khalick. 2002. “Influence of Explicit and Reflective versus Implicit Inquiry-Oriented Instruction on Sixth Graders’ Views of Nature of Science.” *Journal of Research in Science Teaching* 39 (7): 551–578. doi:10.1002/tea.10036.

Khishfe, R., and S. Boujaoude. 2014. “Lebanese Students’ Conceptions of and Attitudes toward Science and Related Careers Based on Their Gender and Religious Affiliations.” *International Journal of Science and Mathematics Education* 14 (1): 145–167. doi:10.1007/s10763-014-9587-0.

Khishfe, R., and N. G. Lederman. 2007. “Relationship between Instructional Context and Views of Nature of Science.”.” *International Journal of Science Education* 29 (8): 939–961. doi:10.1080/09500690601110947.

Lederman, N. G., F. Abd-El-Khalick, R. L. Bell, and R. S. Schwartz. 2002. “Views of Nature of Science Questionnaire: Toward Valid and Meaningful Assessment of Learners’ Conceptions of Nature of Science.” *Journal of Research in Science Teaching* 39 (6): 497–521. doi:10.1002/tea.10034.

LOE (Ley Orgánica de educación). Ley Orgánica 2/2006, de 3 de mayo. Boletín Oficial del Estado, núm. 106, 2006, 4 de mayo.
LOGSE (Ley Orgánica de Ordenación General del Sistema Educativo). Ley Orgánica 1/1990, de 3 de octubre. Boletín Oficial del Estado, núm. 238, 4 de octubre de 1990.
LOMCE (Ley Orgánica para la mejora de la calidad educativa). Ley Orgánica 8/2013, 9 de diciembre. Boletín Oficial del Estado, núm. 295, 2013, 10 de diciembre.
McComas, W. F. 2015. "The Nature of Science and the Next Generation of Biology Education." *The American Biology Teacher* 77 (7): 485–491. doi:10.1525/abt.2015.77.7.2.
McComas, W. F., and J. K. Olson. 1998. "The Nature of Science in International Science Education Standards Documents." In *The Nature of Science in Science Education: Rationales and Strategies*, edited by W. F. McComas, 41–52. Dordrecht: Kluwer.
McHugh, M. L. 2012. "Interrater Reliability: The Kappa Statistic." *Biochemia Medica* 22 (3): 276–282. doi:10.11613/issn.1846-7482.
MECD (Ministerio de Educación, Cultura y Deporte). 2017. *Datos Y Cifras. E5. El Alumnado Extranjero* [Facts and Figures. E5. Foreign Students]. http://www.mecd.gob.es/dms/mecd/servicios-al-ciudadano-mecd/estadisticas/educacion/indicadores-publicaciones-sintesis/cifras-educacion-espana/2014-15/e5-pdf.pdf
Metin, D., and G. Leblebicioğlu. 2015. "Development of Elementary 6th and 7th Grade Students' Views about Scientific Model and Modeling Throughout a Summer Science Camp."." *Education and Science* 40 (177): 1–18. doi:10.15390/EB.2015.1507.
Nunnally, J. C. 1978. *Psychometric Theory*. New York: McGraw-Hill.
Osborne, J., S. Simon, and S. Collins. 2003. "Attitudes Towards Science: A Review of the Literature and Its Implications." *International Journal of Science Education* 25 (9): 1049–1079. doi:10.1080/0950069032000032199.
Parker, L. H., and J. L. Rennie. 2002. "Teachers' Implementation of Gender-Inclusive Instructional Strategies in Single-Sex and Mixed-Sex Science Classrooms." *International Journal of Science Education* 24 (9): 881–897. doi:10.1080/09500690110078860.
Patton, M. Q. 2002. *Qualitative Research and Evaluation Methods*. Thousand Oaks, CA: Sage.
Porlán, R., R. Martín Del Pozo, A. Rivero, J. Harres, P. Azcárate, and M. Pizzato. 2010. "El cambio del profesorado de Ciencias I: Marco teórico y formativo." *Enseñanza de las ciencias* 28 (1): 31–46.
Portes, A., R. Aparicio Gómez, and W. Haller. 2013. *La Segunda Generación En Madrid: Un Estudio Longitudinal. Avance De Resultados [The Second Generation in Madrid: A Longitudinal Study. Preview of Results]*. Madrid: Real Instituto Elcano.
Quigley, C., K. Pongsanon, and V. L. Akerson. 2010. "If We Teach Them, They Can Learn: Young Students Views of Nature of Science Aspects to Early Elementary Students during an Informal Science Education Program." *Journal of Science Teacher Education* 21 (7): 887–907. doi:10.1007/s10972-009-9164-5.
RAE (Real Academia Española). 2014. *Diccionario de la lengua española* [Dictionary of the Spanish language]. http://dle.rae.es/
Rumbaut, R. G., and A. Portes. 2011. *Ethnicities.Children of Immigrants in America*. London and New York: University of California Press and Russell Sage Foundation.
Said, Z., R. Summers, F. Abd-El-Khalick, and S. Wang. 2016. "Attitudes toward Science among Grades 3 through 12 Arab Students in Qatar: Findings from a Cross-Sectional National Study." *International Journal of Science Education* 38 (4): 621–643. doi:10.1080/09500693.2016.1156184.
Scutt, H. I., S. K. Gilmartin, S. Sheppard, and S. Brunhaver. 2013. "Research-Informed Practices for Inclusive Science, Technology, Engineering, and Math (STEM) Classrooms: Strategies for Educators to Close the Gender Gap." Paper presented at the 120th ASEE Annual Conference & Exposition. https://web.stanford.edu/group/design_education/wikiupload/4/46/ASEE_2013_Scutt.pdf
Tai, R. H., C. Q. Liu, A. V. Maltese, and X. Fan. 2006. "Planning Early for Fareers in Science."." *Science* 312 (5777): 1143–1144. doi:10.1126/science.1128690.
Toma, R. B., and I. M. Greca. 2018. "The Effect of Integrative STEM Instruction on Elementary Students' Attitudes toward Science." *EURASIA Journal of Mathematics, Science & Technology Education* 14 (4): 1383–1395. doi:10.29333/ejmste/83676.

Toma, R. B., I. M. Greca, and J. Á. Meneses Villagrá. 2017. “Dificultades de maestros en formación inicial para diseñar unidades didácticas usando la metodología de indagación.” *Revista Eureka sobre Enseñanza y Divulgación de las Ciencias* 14 (2): 442–457. doi:10.25267/Rev_Eureka_ensen_divulg_cienc.2017.v14.i2.11.

Vázquez Alonso, A., and A. Manassero-Mas. 2008. “El declive de las actitudes hacia la ciencia de los estudiantes: Un indicador inquietante para la educación científica.” *Revista Eureka Sobre Enseñanza Y Divulgación de Las Ciencias* 5 (3): 274–292. doi:10.25267/Rev_Eureka_ensen_divulg_cienc.2008.v5.i3.03.

Vázquez-Alonso, Á., M. A. Manassero-Mas, A. García-Carmona, and A. Bennássar-Roig. 2014. “Teachers’ Beliefs on Science-Technology-Society (STS) and Nature of Science (NOS): Strengths, Weaknesses, and Teaching Practices.” In *Topics and Trends in Current Science Education: 9th ESERA Conference Selected Contributions*, edited by C. Bruguiére, A. Tiberghien, and P. Clément, 117–135. Dordrecht: Springer.

Walls, L. 2012. “Third Grade African American Students’ Views of the Nature of Science.” *Journal of Research in Science Teaching* 49 (1): 1–37. doi:10.1002/tea.20450.

Walls, L. 2016. “Awakening A Dialogue: A Critical Race Theory Analysis of U. S. Nature of Science Research from 1967 to 2013.” *Journal of Research in Science Teaching* 53 (10): 1546–1570. doi:10.1002/tea.21266.

Yoon, J., K. J. Kim, and L. A. Martin. 2016. “Culturally Inclusive Science Teaching (CIST) Mode for Teachers of Culturally and Linguistically Diverse Students.” *Journal for Multicultural Education* 10 (3): 322–338. doi:10.1108/JME-01-2016-0012.

Yoon, S. Y., J. K. Suh, and S. Park. 2014. “Korean Students’ Perceptions of Scientific Practices and Understanding of Nature of Science.” *International Journal of Science Education* 36 (16): 2666–2693. doi:10.1080/09500693.2014.928834.

Supplemental material 3: Preservice teachers study

This supplemental material includes the full-text manuscript of second preliminary study about in-service teachers' difficulties when developing inquiry-based teaching units for science education:

Toma, R. B., Greca, I. M., Meneses-Villagrá, J. A. (2017). Dificultades de maestros en formación inicial para diseñar unidades didácticas usando la metodología de indagación. *Revista Eureka sobre Enseñanza y Divulgación de las Ciencias, 14*(2), 442-457.
doi: 0.25267/Rev_Eureka_ensen_divulg_cienc.2017.v14.i2.11

Dificultades de maestros en formación inicial para diseñar unidades didácticas usando la metodología de indagación

Radu Bogdan Toma[1a], Ileana María Greca[1b], Jesús Ángel Meneses-Villagrá[1c]

[1] *Departamento de Didácticas Específicas, Universidad de Burgos. Burgos. España*
[a] *rtx0001@alu.ubu.es,* [b] *imgreca@ubu.es,* [c] *meneses@ubu.es*

[Recibido en mayo de 2016, aceptado en enero de 2017]

El propósito de este estudio fue identificar las dificultades que presentan maestros en formación para elaborar programaciones didácticas de ciencias naturales para educación primaria utilizando una metodología de indagación. Se analizan 157 unidades didácticas, primero clasificándolas atendiendo al grado de indagación que proponen. En aquellas con tipología más indagatoria, se identifican los componentes principales del proceso indagatorio que contienen, siguiendo una rúbrica elaborada a partir del protocolo 'Reformed Teaching Observation Protocol' (RTOP). Los resultados muestran que solamente un tercio de las unidades didácticas permitirían desarrollar una indagación, y parecen indicar que muchos estudiantes tienden a concebir la indagación como un quehacer práctico en el laboratorio, descontextualizado en la mayoría de las ocasiones, existiendo una dicotomía entre lo que recoge el marco teórico y la praxis que plantean los futuros docentes.

Palabras clave: Metodología de indagación, formación inicial de maestros, ciencias en primaria, unidades didácticas.

Elementary pre-service teachers' difficulties for designing science-teaching units by inquiry

The purpose of this study was to identify the difficulties that pre-service teachers show to develop science teaching units for primary education, using a methodology of inquiry. 157 didactic units were analysed, first classifying them according to the degree of investigation proposed. In those showing more inquisitive traits, the main components of the investigative process, according to a rubric developed from the Reformed Teaching Observation Protocol (RTOP) Protocol were identified. The results obtained established that only a third of the didactic units prepared by the pre-service teachers would be useful for developing an inquiry, and they seem to indicate that many students tend to conceived inquiry only as practical work in the laboratory, decontextualized in the majority of occasions.

Keywords: Teaching by inquiry, pre-service teachers, science in primary school, didactic units.

Para citar este artículo: Toma R. B., Greca I. M., Meneses-Villagrá, J. A. (2017) Dificultades de maestros en formación inicial para diseñar unidades didácticas usando la metodología de indagación. *Revista Eureka sobre Enseñanza y Divulgación de las Ciencias* 14 (2), 442–457. Recuperado de: http://hdl.handle.net/10498/19228

Introducción

La metodología didáctica basada en la indagación, cuyo rasgo característico es el papel activo de los estudiantes involucrados en actividades enfocadas a la resolución de problemas para aprender sobre ciencias y sobre la naturaleza de las ciencias, se propone como un elemento central para la reforma de la enseñanza de las ciencias naturales (American Association for the Advancement of Science 1993; Artigue, Baptist, Dillon, Harlen y Léna 2010; National Research Council 2000; Rocard, Csermely, Jorde, Lenzen, Walwerg-Henriksson y Hemmo 2007). Esta metodología permitiría abandonar las concepciones repetitivas, fragmentadas y memorísticas, fruto del enfoque tradicional basado en el libro de texto como principal material didáctico en las aulas de primaria (Cañal, Criado, García-Carmona y Muñoz 2013; García y Martínez 2003). Sin embargo, la realidad muestra una muy discreta - casi inexistente- presencia de la indagación en las aulas de ciencias (Confederación de Sociedades Científicas de España 2011; Escobar y Vílchez 2006; Osborne y Dillon 2008).

Además, las no poco erróneas percepciones sobre la indagación por parte del profesorado favorecen la dicotomía entre su práctica docente y la correcta aplicación de esta estrategia didáctica (Crawford 1999; Roth, Boutonneâ, Mcrobbie, y Lucas 1999; Windschitl, Thompson y Braaten 2008). Así, por ejemplo, se considera el debate de ideas como el principal elemento de la indagación (Carnes 1997), desviarse de la planificación curricular como símil de indagar (Keys y Kennedy 1999) o confundir estar involucrado y activo físicamente con estar motivado y activo intelectualmente (Couso 2014).

Para que los maestros en formación puedan aplicar correctamente esta metodología deben disponer de conocimientos apropiados, es decir, que respondan a las cinco características esenciales de la indagación, descritas en el informe de referencia del National Research Council (NRC 2012). Por tanto, los maestros en formación poseerán los conocimientos necesarios para utilizar esta metodología en sus clases en la medida en que consigan plantear preguntas científicamente orientadas, que permitan a sus alumnos desarrollar explicaciones a partir de evidencias empíricas y ser capaces de comunicarlas y justificarlas. Además, las explicaciones de los alumnos deberán poder ser conectadas con el conocimiento científico y, a su vez, con el currículo de ciencias. En la literatura, ya encontramos varias propuestas para la formación de maestros en la metodología de la indagación, tanto a nivel internacional como en España (Cortés y Gándara 2006; Gil, Martínez, De la Gándara, Calvo y Cortés 2008; Gil y Martínez, 2005; Martínez Chico, López Gay y Jiménez 2013; Martínez Chico, Jiménez Liso y López Gay 2015; Vílchez González y Bravo Torija 2015). En general, los maestros en formación valoran positivamente estas propuestas, a pesar de sus déficits iniciales. Sin embargo, aún son escasos los estudios que muestren las dificultades que presentan estos mismos maestros en formación para implementar esta metodología en las aulas de ciencias de la escuela primaria.

Un primer paso en este sentido consiste en estudiar si son capaces de diseñar programaciones didácticas acordes con esta metodología didáctica. Aunque del diseño a la efectiva implementación hay un largo camino, es el primer paso. Por ello, el objetivo de este trabajo es identificar el dominio que tienen los estudiantes del Grado de Maestro en Educación Primaria sobre la metodología que nos ocupa, a partir del análisis de las Unidades Didácticas (U.D.) que elaboraron durante el curso académico 2013-2014, en la asignatura 'Investigación e Innovación en el aprendizaje del Conocimiento del Medio' en la que se trabaja explícitamente la metodología de la indagación. Para ello, primero presentaremos lo que entendemos por indagación, así como el contexto de formación en didáctica de las ciencias experimentales de los estudiantes analizados.

Fundamentación teórica

Para los maestros, dada su limitada formación en y sobre ciencias (Abell 2007) cualquier reforma metodológica en esta área, que vaya más allá del uso del libro de texto, implica muchos desafíos. Además, para el caso particular de la metodología de la indagación, la mayoría de los maestros no han tenido la oportunidad de aprender ciencias a través de su uso o realizar investigaciones científicas y no disponen del conocimiento y las habilidades necesarios para utilizarla de forma reflexiva y adecuada en el aula. De hecho, los resultados de la investigación (entre otros, Cañal 2000; Greca 2016; Murphy, Neil y Beggs 2007; Porlán, Martín, Rivero, Harres, Azcárate y Pizzato 2010) muestran que, en general, los maestros tienen puntos de vista de y sobre las ciencias a menudo incompatibles con las concepciones que aparecen en la investigación y en las propuestas de reforma: un conocimiento científico fragmentado, superficial y poco sólido, tanto conceptual como procedimental y una concepción empírico-inductivista de la naturaleza de la ciencia, que les lleva a enseñar ciencias en primaria mayoritariamente de forma tradicional, con el libro de texto como principal recurso.

Antes de avanzar, es necesario definir qué es lo que entendemos por metodología de indagación en la enseñanza de las ciencias, dado existen diferentes interpretaciones. Nosotros hemos adoptado la que plantea el National Research Council (NRC 2000, 2012), con los siguientes componentes:

a) Los saberes, eventos o fenómenos abordados deben poseer un interés científico que conecte con los conocimientos propios de los alumnos, y que les cree un conflicto cognitivo que impulse el desarrollo de sus ideas previas.

b) Los alumnos deben involucrarse en experiencias manipulativas que entrañen la formulación y comprobación de hipótesis, la resolución de problemas reales, y la generación de explicaciones del fenómeno estudiado a partir del análisis y la interpretación de datos y la síntesis de ideas propias.

c) Durante el proceso de indagación es necesario que los alumnos construyan modelos, aclaren conceptos y amplíen conocimientos y habilidades aplicables a otras situaciones, además de ir aprendiendo también algunos elementos centrales sobre la forma de hacer ciencia.

d) Debe haber una revisión conjunta entre el alumnado y el profesorado para evaluar el qué y el cómo se ha aprendido.

Desde esta perspectiva, el uso de la metodología de la indagación no implica solamente potenciar el desarrollo de "destrezas indagativas", sino también contenidos conceptuales de ciencia (leyes, teorías y modelos) y contenidos sobre la naturaleza de la ciencia. En relación a los modelos, cabe resaltar que están, salvo algunas excepciones, ausentes de la enseñanza en la escuela primaria, tanto porque persisten las ideas de que a los niños sólo debe enseñárseles los componentes más simples (Schauble, Glaser, Duschl, Schulze, y John 1995) así como a las propias dificultades de los maestros (Anderson 2002; Hayes 2002).

Los maestros en formación, igual que los que están en activo, presentan semejantes concepciones inadecuadas sobre la naturaleza de la ciencia (Campbell y Bohn 2008; Guisasola y Morentin 2007; Windschitl 2002;): una visión positivista y un énfasis en la ciencia como producto. Sin embargo, algunas investigaciones muestran que pueden llegar a puntos de vista congruentes con los articulados por los especialistas, si participan en auténticas investigaciones científicas como parte de su formación (Haefner y Zembal-Saul 2004; Seung, Park y Jung 2014; Windschitl 2002).

No obstante, desarrollar mejores comprensiones sobre contenidos científicos o sobre el quehacer científico no garantiza que los maestros en formación puedan ser capaces de diseñar U.D. basadas en esta metodología sin percepciones erróneas acerca de su uso en el aula de primaria. Recordemos que el diseño de unidades didácticas se constituye en el elemento central de los docentes para la implementación efectiva del currículo (Brown 2009; Remillard 2005), cuando, al usar sus recursos personales (conocimientos, creencias, identidades, entre otros) interactúan con el material curricular disponible, interpretándolo, criticándolo, seleccionándolo y adaptándolo. En el caso del diseño de U.D. por indagación los futuros maestros parten en general de recursos curriculares (libros, experiencias en internet, entre otros) insuficientes o de nivel muy dispar que deberían modificar usando sus recursos personales, que incluyen en general un dominio limitado de la materia y sus propias concepciones de lo que entienden por indagación. De hecho, investigaciones previas muestran que los docentes en formación tienen dificultades para articular diseños que estén en consonancia con una metodología de indagación, lo que les impide luego ponerlos en práctica (Abell 2007; Blumenfeld y Krajcik 2000; Davis, Petish y Smithey 2006; Seung, Park y Jung 2014; Zembal-Saul, Blumenfeld y Krajcik 2000).

Contexto y antecedentes

Al planificar la programación de las asignaturas obligatorias del área de didáctica de las ciencias experimentales del Grado en Maestro de Educación Primaria de la Universidad de Burgos, partimos del supuesto que si los estudiantes pueden experimentar de forma positiva indagaciones en su aprendizaje en ciencias es más probable que las repliquen en su docencia. Nuestra propuesta de formación consta de tres asignaturas. Los principales objetivos de las dos primeras, de segundo y tercer curso respectivamente, son conseguir que los estudiantes adquieran un mínimo de conocimientos científicos estructurantes, una visión más actual sobre la naturaleza de las ciencias y, sobre todo, una forma diferente de entender su enseñanza, a partir del uso de diferentes metodologías innovadoras propuestas en la literatura (Greca, Díez Ojeda y Meneses Villagrá 2016). En la última disciplina, realizada durante el primer semestre del cuarto curso, se trabaja y utiliza únicamente la metodología de la indagación. En esta asignatura se proponen a los estudiantes situaciones problemáticas abiertas que deben abordar siguiendo los procesos que caracterizan una indagación científica, trabajando en forma grupal. Las problemáticas, que contemplan diversos contenidos científicos, están pensadas para profundizar en un nivel correspondiente al inicio del Bachillerato. Durante cinco semanas, los estudiantes analizan la situación problemática e identifican dos o tres problemas más concretos que puedan ser resueltos de forma experimental, emiten hipótesis y hacen predicciones, planifican los diseños experimentales, utilizan instrumentos y analizan los datos obtenidos, relacionándolos con los modelos científicos correspondientes, establecen conclusiones y buscan nuevas aplicaciones relacionadas con el conocimiento construido. Una vez realizado el trabajo experimental, deben dar respuesta a los problemas abordados elaborando un informe escrito, basándose en la evidencia empírica obtenida.

Nuestra pretensión principal, con esta formación, es ayudar a que nuestros estudiantes, futuros maestros, aprendan a pensar científicamente (NRC 2012), para lo cual consideramos necesario el trabajo indagatorio que realizan en el aula y su comunicación, pues a través del mismo pueden llegar a ser capaces de construir, evaluar y utilizar explicaciones científicas, modelizando y argumentando (Couso 2014), así como participar en prácticas y discursos de la ciencia. Paralelamente a este trabajo experimental, en las clases teóricas se reflexiona junto con los estudiantes sobre los diferentes componentes de esta metodología, que ellos están utilizando, así como la forma de su aplicación en clases de primaria. Durante las tres últimas semanas de la asignatura, los estudiantes diseñan, de forma individual, una secuencia didáctica para la escuela primaria, usando la metodología de indagación sobre algún tema que esté incluido dentro del currículum de este nivel.

En una encuesta contestada por el grupo que realizó esta asignatura por primera vez (ver Greca, Meneses y Díez Ojeda 2016) los estudiantes declararon, de forma casi unánime, que después de la indagación mejoraron sus conocimientos procedimentales y conceptuales, y la mitad manifestó haber mejorado su imagen de ciencia, que pasó a ser percibida como más cercana e interesante. En relación al uso de la metodología de indagación en educación primaria, es significativo el número de aspectos positivos que destacaron, desde el aprendizaje de ciencia y sobre la ciencia pasando al desarrollo de autonomía por parte de los alumnos y la cooperación. Muchos de los estudiantes indicaron que esta metodología era altamente motivadora (66%) y que podría permitir que los niños aprendiesen ciencias 'haciendo ciencias', mostrando indicios de mejora en su percepción de la importancia de enseñar ciencias de forma significativa. Además, parecieron comenzar a aceptar las preguntas y el interés de los alumnos como base para el diseño curricular de la enseñanza de las ciencias, factor que influye en la utilización de la indagación en el aula (Howes 2002).

Preguntas de investigación

La alta motivación que despierta la indagación en los maestros en formación no tiene por qué ir necesariamente acompañada de una comprensión adecuada de lo que es una indagación, lo que impediría que, al ejercer como maestros, la aplicasen con cierta corrección. Por ello, decidimos estudiar en profundidad las secuencias didácticas que diseñaron los estudiantes que respondieron la encuesta indicada en el párrafo anterior. Las preguntas de investigación planteadas fueron:

1. ¿Qué es lo que entienden los estudiantes en formación por «indagación»?
2. ¿Cuáles son las principales dificultades que poseen los maestros en formación para incluir la metodología de indagación en el diseño de unidades didácticas?
3. ¿En qué medida difieren las propuestas didácticas de los maestros en formación con respecto al corpus teórico de la indagación?

Metodología

Este estudio descriptivo se ha realizado según una metodología cuantitativa no experimental, basada en la aplicación de una categorización y una rúbrica. La categorización permitió clasificar las U.D. de acuerdo a la presencia o no de elementos relacionados con la indagación, y la rúbrica, identificar las principales dificultades para elaborar programaciones didácticas de naturaleza indagadora. Ambos instrumentos fueron confeccionados a partir de la adaptación de otros.

En el caso de la categorización, se utilizaron los resultados obtenidos por Schwarz y Gwekwerere (2007) para establecer seis tipos de U.D. que poseen una menor o mayor presencia de indagación: 1) práctica incoherente, 2) práctica coherente, 3) académica, 4) proyecto, 5) investigadora y 6) indagadora (en el anexo 1 aparecen las orientaciones de estos tipos de unidades didácticas). Con esta categorización fue posible realizar una primera selección de las U.D. que mostraban indicios de la metodología de indagación.

La rúbrica es una adecuación del protocolo 'Reformed Teaching Observation Protocol' (RTOP), desarrollado por Sawada y Piburn (2000), que se ha utilizado para profundizar en el análisis de las programaciones didácticas una vez que fueron clasificadas mediante el instrumento indicado en el párrafo anterior. Este protocolo fue desarrollado originalmente para la observación de clases teniendo en cuenta las características de la metodología de la indagación tal como nosotros la concebimos y trabajamos con los estudiantes. Las modificaciones que realizamos en este caso fueron dos. Por una parte, seleccionamos del instrumento original los ítems más significativos para definir una indagación y que pudieran ser detectados en el diseño de una unidad didáctica y los agrupamos teniendo en cuenta los cuatro componentes de una indagación definidos anteriormente en la fundamentación teórica. La otra modificación fue añadir, de forma explícita, características a cada ítem para valorar de forma más objetiva las U.D. Así, en cada ítem se describen diferentes características del mismo, en una escala de 0 a 3, los valores superiores que corresponden a aquellos que capturan más elementos cruciales de la indagación (ver ejemplos en las tablas 1, 2, 3, y 4). La asignación de los valores de cada ítem se realizó después de una "codificación" de cada U.D. Así, por ejemplo, todas las actividades propuestas (ya fueran a ser realizadas por el maestro o por el alumno) eran identificadas por su tipo, característica, finalidad según el contexto, tipo de participación del alumnado y recursos. Con esto, se podía identificar fácilmente si había tres, dos, una o ninguna de las actividades que, por ejemplo, son indicativas para asignar valores en el item7 del componente B, que aparece en la Tabla 2.

Con esta rúbrica se pueden detectar aspectos que resultan más difíciles de integrar en una secuencia didáctica, así como el grado de coherencia y de apertura de las indagaciones planteadas. Antes del análisis definitivo, se realizó una prueba piloto con el 20% de las U.D., que permitió definir mejor los ítems de la rúbrica. Para la prueba piloto, los dos autores primeros de este trabajo seleccionaron cada uno diez U.D. al azar, aplicando los dos instrumentos. Finalmente, estos autores analizaron de forma independiente todas las unidades didácticas usando los dos instrumentos en sus versiones finales. Las U.D. con discrepancias mayores del 10% fueron nuevamente analizadas por el tercer autor del trabajo tanto en la categorización como en su posterior análisis, resolviéndose las diferencias por consenso. Los principales cambios introducidos en la rúbrica después de este estudio piloto estuvieron referidos a la claridad en la definición de cada uno de los ítems, de forma a evitar ambigüedades.

La muestra está constituida por las U.D. diseñadas individualmente por 157 estudiantes del 4º curso del Grado en Maestro de Educación Primaria de la Universidad de Burgos. elaboradas durante el curso académico 2013-2014 en la asignatura titulada Investigación e Innovación en el aprendizaje del Conocimiento del Medio, tras haber cursado con anterioridad otras dos asignaturas obligatorias que figuran como escalones para su formación en Ciencia, impartidas en cursos 2º y 3º. Cada una de estas U.D. consta de 10-12 páginas y fueron analizadas de forma completa.

Resultados y discusión

En la figura 1 se plasman los resultados de la clasificación de las U.D atendiendo al primer instrumento.

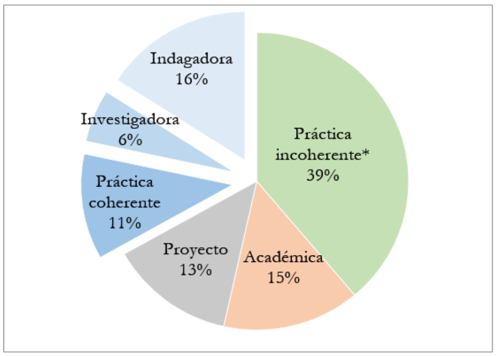


Figura 1. Clasificación de las U.D. en seis tipos

Un 15% del total lo conforman las U.D con un planteamiento académico en el que el problema planteado se aborda de una forma teórica y tradicional a través de clases expositivas. A la clasificación de "Proyecto" pertenecen un 13% de las programaciones didácticas analizadas, siendo caracterizadas por guiar el proceso de enseñanza según un artefacto que se persigue construir, sin abordarse conceptos claves para su construcción ni conectar con las ideas previas del alumno. Aproximadamente la mitad de las U.D (ver figura 1: práctica incoherente 39%, práctica coherente 11%) poseen un enfoque práctico en el que se desarrollan actividades manipulativas guiadas que el alumno ha de realizar siguiendo unas instrucciones determinadas. Sin embargo, una parte mayoritaria de las mismas tiene una estructura incoherente, presentándose conceptos que guardan poca o nula relación con el problema inicial, los experimentos que se realizan no abordan la temática objetivo de estudio, y/o la práctica está descontextualizada de la teoría, por lo que los alumnos no pueden generar explicaciones sobre el fenómeno observado. Por ello, decidimos separar estas unidades categorizándolas como "prácticas incoherentes". Las unidades restantes categorizadas en este enfoque práctico (que llamamos de prácticas coherentes), aunque inciden en demasía en el hacer manipulativo, plantean actividades pertinentes al problema estudiado y existe una cierta progresión en la dificultad y abstracción de las mismas.

Por último, apenas el 22% de las programaciones pueden ser catalogadas como investigaciones. No obstante, cabe diferenciar entre dos tipos. Por un lado, las investigaciones abiertas, en la que los alumnos prácticamente solos deberían llegar a un resultado (6% del total). Por el otro, las indagaciones más guiadas, que responden a las características planteadas, pero en las que el profesor escalona las actividades para que los alumnos adquieran progresivamente mayores niveles de autonomía. Esto es, el 16% de las U.D. parten del análisis de un problema real que se resuelve a través de la experimentación, que conecta y compara teorías científicas, plantea y comprueba hipótesis y se manipulan variables, lo que produce como resultado la posible construcción de modelos, con componentes explicativas, a partir de la observación de los patrones en los datos y el fomento del desarrollo de conceptos y habilidades en el alumnado. En todos estos procesos, docente y alumnos trabajan de forma conjunta. En resumen, casi el 67% de los alumnos, a pesar del programa de formación específico y de su motivación, no consiguen diseñar U.D. que contengan elementos relacionados con la metodología de la indagación, ya sea porque relacionan la indagación exclusivamente con actividades manipulativas (actividades prácticas, diseño de prototipos), porque sus dificultades conceptuales les impiden plantear una secuencia coherente entre problemas y actividades o porque adoptan un modelo tradicional de clase expositiva.

A las unidades clasificadas dentro de las categorías 'práctica coherente', 'investigadora', e 'indagadora' (N=52) se les aplicó la rúbrica RTOP. En la Tabla 1 se presenta los resultados obtenidos en los ítems del componente "Los saberes, eventos o fenómenos abordados poseen un interés científico que conecta con los conocimientos propios de los alumnos y les provoca un conflicto cognitivo que impulsa el desarrollo de las ideas previas".

Tabla 1. Resultado final del primer componente de la rúbrica RTOP.

Componente A (NRC, 2000): Los saberes, eventos o fenómenos abordados poseen un interés científico que conecta con los conocimientos propios de los alumnos y les provoca un conflicto cognitivo que impulsa el desarrollo de las ideas previas.				
ITEMS RTOP	**Escalas para clasificación de las Unidades Didácticas**			
	3	**2**	**1**	**0**
Ítem 1. *Las estrategias de enseñanza y las actividades planteadas respetan los conocimientos previos y las pre-concepciones de los estudiantes*	El maestro identifica y reflexiona acerca de las ideas previas del alumnado, las compara con los modelos científicos y las retoma al final de la U.D para analizar su desarrollo.	El maestro realiza 2 de las 3 actividades siguientes: Identifica y reflexiona acerca de las ideas previas del alumnado, las compara con los modelos científicos, las retoma al final de la U.D para analizar su desarrollo.	El maestro realiza 1 de las 3 actividades siguientes: Identifica y reflexiona acerca de las ideas previas del alumnado, las compara con los modelos científicos, las retoma al final de la U.D para analizar su desarrollo.	Ausencia del tratamiento de las ideas previas del alumnado.
Porcentaje de respuestas	***1.9%***	***38.4%***	***51.9%***	***7.8%***
Ítem 6. *La lección abordaba conceptos fundamentales de la materia.*	El contenido de la U.D está estructurado de forma clara y coherente, con niveles progresivos de abstracción, y atendiendo a lo recogido en la LOE.	El contenido de la U.D presenta niveles progresivos de abstracción, y atiende a lo recogido en la LOE.	El contenido de la U.D está desestructurado; se abordan conceptos abstractos sin actividades previas, pero se atiende a lo recogido en la LOE.	El contenido de la U.D no posee una estructura. Se abordan conceptos descontextualizados, sin atender a lo recogido en la LOE y sin que haya una progresión en la abstracción de los contenidos.
Porcentaje de respuestas	***25%***	***57.7%***	***17.3%***	***0%***
Ítem 17. *Los comentarios y las dudas planteadas por los estudiantes marcan el desarrollo de la clase.*	La lección formula preguntas investigables, significativas, contextualizadas y éticas y abordables. Las preguntas se contestan a partir de las actividades propuestas.	La lección formula preguntas que cumplen al menos varios criterios necesarios para una investigación efectiva. Las preguntas giran en torno al "¿por qué?" y "¿cómo?, y se contestan en parte a partir de las actividades propuestas.	Escasa presencia de preguntas científicas o mala elaboración de las mismas. Se confunde el "¿por qué?" con el "¿cómo?" y también viceversa. Las preguntas no pueden ser contestadas a partir de las actividades planteadas pues precisan de información adicional.	Ninguna evidencia de preguntas científicas.
Porcentaje de respuestas	***9.6%***	***65.4%***	***25%***	***0%***

Del primer ítem, se deduce que en más de la mitad de las U.D. (51,9%) solo se identifican las ideas previas y tan solo en un 1,9% se comparan dichas ideas con los modelos científicos para retomarlas al final de la programación y analizar su desarrollo. El siguiente ítem indica la gran dificultad de elaborar una programación en la que las actividades planteadas presenten una estructura clara y coherente, y atendiendo a niveles progresivos de abstracción (25%), aunque todas aborden conceptos recogidos en la legislación vigente (Ley Orgánica para la mejora de la calidad educativa 2013). En el tercer y último ítem se observa que todas las programaciones didácticas formulan preguntas que se contestan a partir de las actividades planteadas y que además poseen los criterios necesarios para una investigación efectiva, girando la mayoría de ellas (65,4%) en torno al "¿por qué?" y al "¿cómo?" del funcionamiento de un determinado fenómeno.

La tabla 2 recoge los resultados obtenidos en el componente "Los alumnos son involucrados en experiencias manipulativas que entrañan la formulación y comprobación de hipótesis, la resolución de problemas reales, y la generación de explicaciones del fenómeno estudiado a partir del análisis y la interpretación de datos, y la síntesis de ideas propias".

Tabla 2. Resultado final del segundo componente de la rúbrica RTOP.

Componente B (NRC, 2000): Los alumnos son involucrados en experiencias manipulativas que entrañan la formulación y comprobación de hipótesis, la resolución de problemas reales, y la generación de explicaciones del fenómeno estudiado a partir del análisis y la interpretación de datos y la síntesis de ideas propias.				
ÍTEMS RTOP	**Escalas para clasificación de las Unidades Didácticas**			
	3	**2**	**1**	**0**
Ítem 5*. El hilo con-ductor de la clase estaba a menudo determina-do por las ideas originadas por los estudiantes.*	El maestro fomenta el planteamiento de preguntas, las discusiones grupales y el desarrollo de ideas del alumnado.	El maestro fomenta al menos 2 de las siguientes 3 actividades: planteamiento de preguntas; discusiones grupales; desarrollo de ideas del alumnado.	El maestro fomenta al menos 1 de las siguientes 3 actividades: planteamiento de preguntas; discusiones grupales; desarrollo de ideas del alumnado.	Las inquietudes e ideas de los estudiantes no influyen en el desarrollo de la clase.
Porcentaje de respuestas	17.3%	57.8%	21.1%	3.8 %
Ítem 7. *La lección promovió la comprensión de los conceptos estudiados*	En la U.D se retoma el problema inicial para plantear soluciones alternativas, se realiza una exposición o presentación del estudio realiza do por el alumnado, y se explica o se plantea al alumnado la aplicación de los resultados obtenidos.	En la U.D se realizan 2 de las siguientes 3 actividades: Plantear soluciones alternativas al problema inicial; realizar exposición o presentación didáctica; explicar o plantear al alumnado la aplicación de los resultados obtenidos.	En la U.D se realizan 1 de las siguientes 3 actividades: Plantear soluciones alternativas al problema inicial; realizar exposición o presentación didáctica; explicar o plantear al alumnado la aplicación de los resultados obtenidos.	No hay actividades que permitan identificar el grado de comprensión de los conceptos estudiados.
Porcentaje de respuestas	1.9 %	21.1%	69.8%	7.2 %
Ítem 10. *Se exploró la conexión del contenido de una forma interdisciplinar y/o con fenómenos del mundo real.*	El problema planteado tiene relación con el mundo cercano el alumno. Los resultados obtenidos le ayudan a entender mejor su entorno, y le permite aplicarlos a otros fenómenos o áreas.	Tan solo se cumplen 2 de los siguientes 3 fenómenos: El problema planteado tiene relación con el mundo cerca-no al alumno. Los resultados obtenidos le ayudan a entender mejor su entorno, y le permite aplicarlos a otros fenómenos o áreas.	Tan solo se cumplen 1 de los siguientes 3 fenómenos: El problema planteado tiene relación con el mundo cerca-no el alumno. Los resultados obtenidos le ayudan a entender mejor su entorno, y le permite aplicarlos a otros fenómenos o áreas.	Ninguna evidencia de tratamiento interdisciplinar o aplicación de resultados obtenidos. El problema planteado es demasia-do complejo o ambiguo para las capacidades del alumnado.
Porcentaje de respuestas	19.2%	55.8%	25%	0%
Ítem 12. *Los estudiantes han formulado estimaciones, predicciones y/o hipótesis, así como experimentos y medios para su comprobación.*	La U.D posibilita a los estudiantes formular hipótesis, diseñar experimentos, y generar explicaciones que tienen relación con el problema planteado y los datos obtenidos.	La U.D posibilita a los estudiantes 2 de las 3 acciones siguientes: formular hipótesis, diseñar experimen-tos, y generar explicaciones que tienen relación con el problema planteado y los datos obtenidos.	La U.D posibilita a los estudiantes 1 de las 3 acciones siguientes: formular hipótesis, diseñar experimentos, y generar explicaciones que tienen relación con el problema planteado y los datos obtenidos.	Ninguna evidencia de que el alumnado sea protagonista de su aprendizaje.
Porcentaje de respuestas	3.8 %	57.8%	34.6%	3.8 %

Los ítems 5, 7, 10 y 12 muestran que en más de la mitad de las U.D (57,8%), se fomenta el planteamiento de preguntas y discusiones grupales; sin embargo, en apenas un 21,1% de los casos se plantean actividades que busquen el desarrollo de ideas propias del alumnado. Además, la aplicación de los resultados obtenidos a otras áreas o fenómenos se da en el 1,9% de las programaciones, predominando los planteamientos en los que el alumnado se limita a exponer los resultados obtenidos (69,8%). El tercer ítem de esta categoría indica que pese a que los problemas del que parten las U.D tienen relación con el mundo cercano del alumno (55,8%), apenas un 19,2% plantean actividades que le ayuden realmente a entender mejor su entorno, potenciando un aprendizaje significativo. En cuanto a la formulación de estimaciones, predicciones y/o hipótesis, cabe destacar que en el 57,8% de los casos es el alumno el encargado de realizarlas y, sin embargo, en muy pocas ocasiones tienen la posibilidad de plantear experimentos (3,8%).

En la Tabla 3 se muestran los resultados que atañen al tercer componente denominado "Se construyen modelos, aclaran conceptos y amplían conocimientos y habilidades aplicables a otras situaciones".

Tabla 3. Resultado final del tercer componente de la rúbrica RTOP.

Componente C (NRC, 2000): Se construyen modelos, aclaran conceptos y amplían conocimientos y habilidades aplicables a otras situaciones.				
ÍTEMS RTOP	**Escalas para clasificación de las Unidades Didácticas**			
	3	**2**	**1**	**0**
Ítem 4. *Esta lección animó a los estudiantes a buscar y valorar modos de investigación alternativos o de resolución de problemas.*	El alumnado tiene la oportunidad de buscar y analizar, plantear y diseñar, y llevar a la práctica experimentos alternativos para su comprobación.	El alumnado tiene la oportunidad de realizar 2 de las siguientes 3 actividades: Buscar y analizar; plantear y diseñar; llevar a la práctica experimentos alternativos.	El alumnado tiene la oportunidad de realizar 1 de las siguientes 3 actividades: Buscar y analizar; plantear y diseñar; llevar a la práctica experimentos alternativos.	Ninguna evidencia de tratamiento de modos de investigación alternativos.
Porcentaje de respuestas	***3.8%***	***3.8%***	***15.3%***	***77.1%***
Ítem 19. *Se anima a los estudiantes a elaborar conjeturas, estrategias de resolución alternativas y formas varias de interpretar los resultados obtenidos.*	El maestro fomenta el desarrollo de diferentes explicaciones que guardan relación con el problema planteado y el fenómeno objeto de estudio. Las explicaciones elaboradas se defienden a partir de los datos obtenidos.	El maestro fomenta el desarrollo de diferentes explicaciones que guardan o no relación con el problema planteado y el fenómeno objeto de estudio. Las explicaciones elaboradas se defienden a partir de los datos obtenidos.	El maestro fomenta el desarrollo de explicaciones que guardan poca relación con el problema planteado y el fenómeno objeto de estudio. Las explicaciones elaboradas no se defienden a partir de los datos obtenidos.	El maestro no fomenta la elaboración de explicaciones por parte de los estudiantes.
Porcentaje de respuestas	**13.4%**	**69.4%**	**13.4%**	**3.8 %**
Ítem 20. *El maestro ofrece recursos para apoyar y mejorar las propuestas del alumnado.*	El maestro ofrece diferentes tipos de recursos que varían en el grado de dificultad y abstracción, presentan diferentes materiales y soportes, y su naturaleza es distinta (conceptual y procedimental)	El maestro ofrece 2 de los siguientes 3 tipos de recursos: que varían en grado de dificultad y abstracción; diferente material y soporte; naturaleza distinta (conceptual y procedimental)	El maestro ofrece 1 de los siguientes 3 tipos de recursos: que varían en grado de dificultad y abstracción; diferente material y soporte; naturaleza distinta (conceptual y procedimental)	El maestro no ofrece recursos que pueda orientar el quehacer del alumnado.
Porcentaje de respuestas	**23%**	**36.5%**	**40.5%**	**0%**

El 77,1% de las U.D no muestran evidencia de abordar modos de investigación o resolución de problemas alternativos. Pese a ello, y a que casi la mitad de las actividades tan solo varían en el material o el soporte utilizado (40,5%), en el 69,4% de los casos las actividades fomentan que el alumnado intente explicar los resultados obtenidos con el problema inicial desde algún marco teórico. En cuanto a los recursos ofrecidos por el maestro, apenas un 23% de las U.D. estaban acompañadas de diferentes materiales y soportes enfocados a mejorar la resolución del problema inicial por parte del alumnado. Por el contrario, la gran mayoría apenas ofrecían apoyo para esclarecer las dudas conceptuales (40,5%) que pudieran surgir a partir del análisis de los nuevos contenidos curriculares a enseñar por el maestro y aprender por el alumnado.

El cuarto y último componente permite analizar el planteamiento de la evaluación (Tabla 4) para determinar en qué medida "Hay una revisión conjunta entre el alumnado y el profesorado para evaluar el qué y el cómo se ha aprendido".

Ninguna de las U.D. analizadas ha considerado una evaluación conjunta entre el profesorado y el alumnado. En el 82,8% de los casos se ha planteado una evaluación tradicional, basada en actividades escritas (44,4%) y en exposiciones de los alumnos (38,4%), mostrándose incoherentes con su planteamiento general. Es significativa también la falta de actividades que evalúen, al final de la unidad, la presencia de ideas o modelos alternativos a los planteados por el docente (3,8%).

Tabla 4. Resultado final del cuarto componente de la rúbrica RTOP.

Componente D (NRC, 2000): Hay una revisión conjunta entre el alumnado y el profesorado para evaluar el qué y el cómo se ha aprendido.				
ÍTEMS RTOP	**Escalas para clasificación de las Unidades Didácticas**			
	3	**2**	**1**	**0**
Ítem 13. *Los estudiantes han participado de forma activa en actividades de reflexión que implicaban la evaluación crítica de los procedimientos empleados.*	La U.D genera diálogo y debate a partir de la comparación de los resultados obtenidos. Se realiza una evaluación conjunta y se estudia la viabilidad y aplicación de las soluciones propuestas por el alumnado.	La U.D genera 2 de las 3 actividades: Genera diálogo y debate a partir de la comparación de los resultados obtenidos; se realiza una evaluación conjunta; se estudia la viabilidad y aplicación de las soluciones propuestas por el alumnado.	La U.D genera 1 de las 3 actividades: Genera diálogo y debate a partir de la comparación de los resultados obtenidos; se realiza una evaluación conjunta; se estudia la viabilidad y aplicación de las soluciones propuestas por el alumnado.	La U.D concibe al alumnado como un sujeto pasivo: se limita a realizar actividades. No hay reflexión ni actitud crítica.
Porcentaje de respuestas	***0%***	***17.3%***	***75.1%***	***7,6%***
Ítem 14. *El rigor intelectual, la crítica constructiva y la valoración de ideas y modelos alternativas fueron valoradas.*	El maestro ha utilizado hojas de registro, exposiciones y actividades escritas para determinar el grado de consistencia de las explicaciones de los alumnos y la presencia de ideas y modelos alternativos para explicar el mismo conjunto de datos.	El maestro utiliza 2 de las siguientes 3 actividades para determinar el grado de consistencia de las explicaciones de los alumnos y la presencia de ideas alternativas: hojas de registro; exposiciones; actividades escritas.	El maestro utiliza 1 de las siguientes 3 actividades para determinar el grado de consistencia de las explicaciones de los alumnos y la presencia de ideas alternativas: hojas de registro; exposiciones; actividades escritas.	No hay evidencias de la valoración de las ideas alternativas.
Porcentaje de respuestas	***3.8%***	***38.4%***	***44.4%***	***13.4 %***

En resumen, las principales carencias que parecen haber tenido este grupo de estudiantes (el 33% de los alumnos totales) para diseñar secuencias didácticas que incluyan los cuatro componentes de una indagación para la enseñanza de las ciencias en la escuela primaria, tal como nosotros la hemos definido, son:

a) Dificultades para incluir en el diseño el trabajo con (y a partir de) las ideas previas de los alumnos. Esto está relacionado con las dificultades que tienen los maestros en formación para evaluar y modificar sus clases de acuerdo con el conocimiento de los niños (Buck, Trauth y Kaftan 2010; Zangori y Forbes 2013).

b) Dificultades para dar más autonomía a los alumnos, tanto en el desarrollo de sus propias ideas como en el desarrollo de potenciales experimentos, posiblemente relacionado con su propia inseguridad con los conceptos científicos, así como con su falta de confianza en las capacidades de los alumnos (Wallace y Kang 2004).

c) Dificultades para proponer evaluaciones coherentes con el planteamiento didáctico general. Es importante destacar que este problema no es exclusivo de la metodología de la indagación, sino que las evaluaciones con bajas demandas cognitivas, parecen ser la norma, independientemente de la metodología, en la enseñanza de las ciencias en primaria en toda Europa (Osborne y Dillon 2008).

Por el contrario, los aspectos más favorables del diseño de las U.D. reside en establecer secuencias de actividades con progresivos niveles de abstracción que permitan a los alumnos aprender significativamente los conceptos deseados, resultado en consonancia al obtenido por Vilchez González y Bravo Torija (2015), en el que los maestros en formación son capaces de reconocer las diferentes etapas de la indagación científica.

Sumamos, para cada U.D., los valores totales obtenidos según la rúbrica RTOP. Luego, fueron establecidos puntos de cortes con percentiles iguales basados en los casos explorados, lo que ha conformado tres categorías diferentes –indagaciones con 4 componentes NRC, indagaciones con 2 y 3 componentes NRC, indagaciones con 1 componente NRC Así, en las U:D. que sumaban entre 24 a 36 puntos se puede apreciar la presencia de los 4 componentes; las que sumaban entre 12 a 23 puntos, es posible apreciar que sólo dos o tres de las componentes son abordadas y finalmente, aquellas que sumaban de 0 a 11 puntos sólo conseguían plasmar un componente. Los resultados de esta categorización aparecen recogidos en la figura 2.

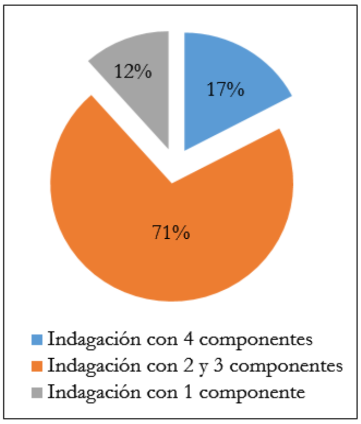


Figura 2. Resultados al aplicar la rúbrica RTOP según la presencia de componentes comunes en la metodología por indagación

Se observa que la mayoría de las U.D pertenecen a la segunda categoría, y una minoría, apenas un 17% de las 52 analizadas con esta rúbrica (si tenemos en cuenta las 157 unidades iniciales, sólo el 5,7%) se consolidan como indagaciones que recogen las diversas pautas sobre esta metodología.

Resulta significativa la puntuación obtenida en cada componente. El puntaje más alto corresponde a los ítems que evalúan el grado de implicación del alumnado en experiencias manipulativas. En contraste, el componente relacionado con la evaluación del qué y cómo se ha aprendido posee la puntuación más baja. Esto evidencia una comprensión no apropiada de la metodología de la indagación, que no reside solamente en el desarrollo de experimentos o la adecuación del método científico, sino que es imprescindible la presencia de la reflexión inter e intrapersonal si se pretende fomentar un aprendizaje significativo. De lo contrario, se estaría desvirtuando la idiosincrasia de esta metodología, y el proceso de enseñanza y aprendizaje se limitaría a transmitir ciencia en lugar de hacer ciencia en la escuela (Couso 2014).

Por otra parte, analizamos posibles relaciones entre esta última categorización y las opiniones de los estudiantes sobre la metodología de la indagación según una encuesta que respondieron al final de la asignatura (y que comentamos en el apartado de Contexto). Este aspecto es interesante para observar si existe relación entre la percepción de los estudiantes sobre su aprendizaje de la metodología de la indagación y la forma en que son capaces de plasmar (y, eventualmente, usar) esta metodología. Se ha encontrado que existen correlaciones significativas, aunque bajas, entre esta categorización de las U.D. y el número de aspectos positivos que los estudiantes señalan sobre la metodología de la indagación (Rho de Spearman 0,24, $p<0,01$) así como con el número de mejoras percibidas en su aprendizaje mediante el uso de esta metodología (Rho de Spearman 0,18, $p<0,01$). Es posible que este hecho esté relacionado con la capacidad de reflexión de los maestros en formación sobre sus propios procesos de aprendizaje, pero esto debe estudiarse con mayor profundidad.

Conclusiones e implicaciones didácticas

Del presente estudio se deducen varias conclusiones. En cuanto a la primera pregunta de investigación, "¿Qué es lo que entienden los estudiantes en formación por «indagación» ?, los resultados parecen mostrar que los maestros en formación aun consiguiendo desarrollar una actitud positiva para implementar la metodología de la indagación a partir de una formación específica, les es difícil traducir ese conocimiento en diseños apropiados. En particular, la confunden con actividades manipulativas, semejantes a las tradicionales prácticas de laboratorio que, a pesar de aumentar el interés de los alumnos por la ciencia y ayudarles a clarificar sus ideas (Gil y Martínez 2005; Osborne y Dillon 2008; Rocard 2007;), no son suficientes para que los alumnos construyan significativamente modelos científicos e ideas sobre la ciencia.

En relación con la segunda pregunta de investigación, "¿Cuáles son las principales dificultades que poseen los maestros en formación para incluir la metodología de indagación en el diseño de unidades didácticas?", la rúbrica RTOP ha permitido identificar varias dificultades. Entre ellas destaca la dificultad en plantear U.D. que permitan retomar las ideas previas del alumnado para modificarlas, ausencia de una estructuración clara y coherente tanto en las actividades propuestas como en su temporalización, escasa aplicación de los resultados obtenidos a través de las indagaciones a otros contextos o problemáticas, y poca flexibilidad en el desarrollo de la propuesta. Es decir, el alumnado sigue siendo, en estas U.D. un actor pasivo, limitado a realizar o contestar las actividades y experimentos diseñados por los maestros.

Por último, la tercera pregunta de investigación "¿En qué medida difieren las propuestas didácticas de los maestros en formación con respecto al corpus teórico de la indagación?", ha puesto de manifiesto que los planteamientos didácticos de los estudiantes se encuentran aún alejados del marco teórico tomado como referencia, bien por escasez de conocimientos científicos, poca congruencia entre objetivos perseguidos y experimentos planteados o evaluaciones no coherentes con la metodología, o bien por una mezcla de los tres elementos. Esto desemboca en la elaboración de U.D. que se encuentran a caballo entre la teoría tradicional y la práctica de aula descontextualizada.

Muchos de los problemas parecen estar asociados a un deficiente bagaje de conocimientos científicos, como lo demuestra el casi 40% de las unidades, con estructuras incoherentes, resultado que coincide con diferentes estudios que indican que este es el mayor hándicap (Davis y Petish 2005; Kim y Tan 2011; Yoon, Joung y Kim 2012). Sin embargo, los resultados muestran también que la propia comprensión de la metodología y su posible trasposición didáctica podrían estar dificultando su uso apropiado, por lo menos a la hora de diseñar unidades didácticas. De hecho, en un estudio de caso (Greca 2016) con estudiantes de este grupo que aplicaron secuencias didácticas por indagación durante su Practicum se observó que el estudiante con mejor formación en ciencias fue el que más dificultades presentó a la hora de diseñar e implementar la indagación por su forma de entenderla. Por ello es posible que, aun siendo cierto que el bajo nivel de formación y de confianza en ciencias influye en la capacidad de los maestros para diseñar secuencias didácticas indagatorias, existan otros factores (entre otros, cómo, después de vivenciarla, entienden la indagación) que deben tenerse en cuenta. Los resultados de este estudio parecen indicar la necesidad que los cursos de formación propicien un aumento del uso de esta metodología durante la formación, utilizándola en varias asignaturas (un único curso parece no ser suficiente), y fomentando su puesta en práctica por parte de los maestros en formación en los períodos de prácticas escolares o en propuestas extraescolares.

Cabe resaltar, finalmente, que la rúbrica para el análisis de U.D. que ha sido usada en esta investigación, construida a partir del protocolo RTOP y con valores claramente definidos, puede ser usada tanto como instrumento de evaluación de U.D. que usen la metodología de la indagación (a nivel de escuela primaria o secundaria) cuanto como instrumento didáctico para el análisis de U.D. En este sentido, puede ayudar a los futuros docentes a comprender la forma en que la metodología de la indagación debe plasmarse en una planificación didáctica.

Agradecimientos

Este trabajo ha sido financiado por MINECO Edu2013-46167-R.
Queremos agradecer las contribuciones realizadas por los revisores, que han servido para mejorar sustancialmente este trabajo.

Referencias bibliográficas

American Association for the Advancement of Science. (1993). Benchmarks for science literacy. New York: Oxford University Press.

Abell S.K. (2007) Research on science teacher knowledge. En S.K. Abell y N.G. Lederman (Eds.), *Handbook of Research on Science Education* (pp.1105-1149). Mahwah. Mahwah, Nueva Jersey. Lawrence Erlbaum.

Anderson R.D. (2002) Reforming Science Teaching: What Research Says About Inquiry. *Journal of Science Teacher Education* 13(1), 1-12.

Artigue M., Baptist P., Dillon J., Harlen W., Léna, P. (2010) *The Fibonacci Project. Scientific Background*. Recuperado de: http://www.ntaskolutveckling.se/globalassets/starting_package_scientific_background1.pdf

Brown M. W. (2009) The teacher-tool relationship: theorizing the design and use of curriculum materials. En J.T Remillard, B.A. Herbel-Eisenmann, G.M. Lloyd (Eds.) *Mathematics teachers at work: connecting curriculum materials and classroom instruction* (pp. 17-36). New York: Routledge.

Buck G.A., Trauth-Nare,A., Kaftan J. (2010) Making formative assessment discernable to Pre-Service Teachers of Science. *Journal of Research in Science teaching*, 47(4), 402-421.

Campbell T., Bohn C. (2008) Science laboratory experiences of high school students across one state in the U.S.: Descriptive research from the classroom. *Science Educator* 17, 36-48.

Cañal P. (2000) El conocimiento profesional sobre las ciencias y la alfabetización científica en primaria. *Alambique. Didáctica de las Ciencias Experimentales* 24, 46-56.

Cañal P., Criado A., García-Carmona A., Muñoz G. (2013) La enseñanza relativa al medio en las aulas españolas de educación infantil y primaria: concepciones didácticas y práctica docente. *Investigación en la Escuela* 81, 21-42.

Carnes G.N. (1997) Teacher conceptions of inquiry and related teaching practices. *The annual meeting for the national association of research in science teaching*, Chicago.

Cortés A. L., Gándara M. (2006*)* La construcción de problemas en el laboratorio durante la formación del profesorado: una experiencia didáctica. *Enseñanza de las Ciencias* 25(3), 435-450.

Confederación de Sociedades Científicas de España (2011) *Informe enciende. Análisis, reflexiones y propuestas para un acercamiento de la ciencia al mundo escolar que promueva en los niños el interés por la ciencia, el aprendizaje científico y una visión no estereotipada de la empresa científica y sus protagonistas*. Editorial Rubes: Barcelona.

Couso D. (2014) De la moda de 'aprender indagando' a la indagación para modelizar: una reflexión crítica. *Conferencia inaugural 26 encuentros de didáctica de las ciencias experimentales*. Universidad De Huelva.

Crawford B.A. (1999) Is it realistic to expect a preservice teacher to create an inquiry based classroom? *Journal of Science Teacher Education* 10, 175-194.

Davis E. A., Petish D., Smithey J. (2006) Challenges new science teachers face. *Review of Educational Research* 76 (4), 607-651.

Davis E. A., Petish D. (2005) Real-world applications and instructional representations among prospective elementary science teachers. *Journal of Science Teacher Education* 16, 263-286.

Escobar T., Vílchez J.E. (2006) Uso del laboratorio escolar en educación primaria: la visión de los estudiantes de magisterio durante el prácticum. En A.L. Cortés y M. D. Sánchez (Eds.), XXII Encuentros de Didáctica de las Ciencias Experimentales. Zaragoza: Universidad de Zaragoza.

García Barros S. y Martínez Losada C. (2003) Las actividades de primaria y eso incluidas en los libros de texto. ¿qué objetivo persiguen? ¿qué procedimientos enseñan? *Enseñanza de las Ciencias* 21, 243-264.

Gil Quilez M. J., Martínez, M. B., de la Gándara M., Calvo J. M., Cortés, A.(2008) De la universidad a la escuela: no es fácil la indagación científica.*Revista Interuniversitaria de Formación del Profesorado* 63 (22,3), 81-100.

Gil Quilez M. J., Martínez de la Gándara G. (2005). Evolución de la didáctica de la Biología: ¿es posible una teoría de síntesis? *Educación Abierta*, 171, 11-28. ICE, Universidad de Zaragoza, ISBN: 84-7791-219-10.

Guisasola J. y Morentin M. (2007) Comprenden la naturaleza de la ciencia los futuros maestros y maestras de educación primaria? *Revista Electrónica de Enseñanza de las Ciencias* 6(2), 246-262.

Greca I. M. (2016) Supporting pre-service elementary teachers in their understanding of inquiry teaching through the construction of a third discursive space. *International Journal of Science Education* 38(5), 791-813.

Greca I. M., Meneses Villagrá, J.A., Díez Ojeda, M. (2016) La formación en ciencias de los alumnos del Grado en Maestro de Educación Primaria. *Revista Electrónica de Enseñanza de las Ciencias.* En prensa

Haefner L. A., Zembal-Saul C. (2004) Learning by doing? Prospective elementary teachers' developing understandings of scientific inquiry and science teaching and learning. *International Journal of Science Education* 26(13), 1653-1674.

Hayes M. (2002) Elementary preservice teachers' struggles to define inquiry-based science teaching. *Journal of Science Teacher Education* 13(2), 147-165.

Howes V.E. (2002) Learning to teach science for all in the elementary grades: what do pre-service teachers bring? *Journal of Research in Science Teaching* 39, 845-869.

Keys C.W., Kennedy, V. (1999) Understanding inquiry science teaching in context: a case study of an elementary teacher. *Journal of Science Teacher Education* 10, 315-333.

Kim M., Tan A.L. (2011) Rethinking difficulties of teaching inquiry-based practical work: stories from elementary pre-service teachers. *International Journal of Science Education* 33, 465-486.

Ley orgánica para la mejora de la calidad educativa (LOMCE) (Ley Orgánica 8/2013, 9 de diciembre). Boletín Oficial del Estado, nº 295, 2013, 10 diciembre.

Martínez-Chico M., López-Gay R., Jiménez-Liso M.R. (2013) Propuesta de formación inicial de maestros fundamentada en la enseñanza por indagación centrada en el modelo de sol-tierra. *Enseñanza de las Ciencias*. Nº Extra, 2173-2178.

Martínez-Chico M., Jiménez-Liso M.R., López-Gay R. (2015). Efecto de un programa formativo para enseñar ciencias por indagación basada en modelos, en las concepciones didácticas de los futuros maestros. *Revista Eureka sobre Enseñanza y Divulgación de las Ciencias* 12(1), 149-166. Recuperado de: http://hdl.handle.net/10498/16929

Murphy C., Neil P., Beggs J. (2007) Primary science teacher confidence revisited: ten years. *Educational Research* 49 (4), 415-430.

National Research Council (2000) *Inquiry and the national science education standards.* Washington: National Academy Press.

National Research Council (2012) *A framework for K-12 science education: practices, crosscutting concepts, and core ideas.* Washinton, DC: The National Academies Press.

Osborne J., Dillon J. (2008) *Science Education in Europe: critical reflections. A report to the Nuffield Foundation.* King's College London.

Porlán R., Martín R., Rivero A., Harres J., Azcárate P., Pizzato M. (2010) El cambio del profesorado de ciencias: marco teórico y formativo. *Enseñanza de las Ciencias* 28(1), 31-46.

Remillard J. T. (2005) Examining key concepts in research on teachers' use of mathematics curricula. *Review of Educational Research* 75(2), 211-246.

Rocard M., Csermely P., Jorde D., Lenzen D., Walwerg-Henriksson H., Hemmo V. (2007) *Science Education Now: A Renewed Pedagogy for the Future of Europe. European Commission. Community Research.* Recuperado de: http://ec.europa.eu/research/science-society/document_library/pdf_06/report-rocard-onscience-education_en.pdf

Roth W., Boutonneâ S., Mcrobbie C., Lucas K. (1999) One class, many worlds. *International Journal of Science Education* 21, 59 -77.

Sawada D., Piburn M. (2000) *Reformed teaching observation protocol (RTOP) Training Guide*, Arizona State University: 10.

Schauble L., Glaser R., Dusch R., Schulze S., John J. (1995) Students'understanding of the objectives and procedures of experimentation in the science classroom. *Journal of The Learning Sciences* 4 (2), 131-166.

Schwarz C. V., Gwekwerere,Y. N. (2007) Using a guided inquiry and modeling instructional framework (EIMA) to support preservice K-8 science teaching. *Science Education* 91(1), 158-186.

Seung E., Park S., Jung J. (2014) Exploring preservice elementary teachers' understanding of the essential features of inquiry-based science teaching using evidence-based reflection. *Research in Science Education* 44(4), 507-525.

Vílchez González J.M., Bravo Torija B. (2015) Percepción del profesorado de ciencias de educación primaria en formación acerca de las etapas y acciones necesarias para realizar una indagación escolar. *Enseñanza de las Ciencias* 33(1), 185-202.

Wallace C. S., Kang N. H. (2004) An investigation of experienced secondary science teachers' beliefs about inquiry: an examination of competing belief sets. *Journal of Research in Science Teaching* 41, 936-960.

Windschitl M. (2002) Inquiry projects in science teacher education: what can investigative experiences reveal about teacher thinking and eventual classroom practice? *Science Education* 87(1), 112-143.

Windschitl M., Thompson J., Braaten M. (2008) Beyond the scientific method: model-based inquiry as a new paradigm of preference for school science investigations. *Science Educatio*n, 92, 941–967.

Zangori L., Forbes C. T. (2013) Preservice elementary teachers and explanation construction: knowledge-for-practice and knowledge-in-practice. *Science Education* 97(2), 310-330.

Zembal-Saul C., Blumenfeld P., Krajcik J. (2000) Influence of guided cycles of planning, teaching, and reflection on prospective elementary teachers' science content representations. *Journal of Research in Science Teaching* 37 (4), 318-339.

Yoon H.G., Joung Y.J., Kim M. (2012) The challenges of Science Inquiry teaching for Pre-service teachers in Elementary classrooms: difficulties on and under the scene. *Research in Science Education* 42: 589-608.

Anexo 1: Tipos de unidades didácticas según el grado de indagación que presentan

Práctica incoherente: Se trata de un tipo de unidad didáctica (U.D.) cuya principal característica es que el tratamiento de conceptos guarda poca o nula relación con el problema inicial que se plantea, haciendo demasiado énfasis en actividades manipulativas que tienden a carecer de fundamento científico. Los experimentos se plantean de forma descontextualizada y con poca relación con el conocimiento científico que se pretende enseñar.

Práctica coherente: Es una U.D. que plantea un problema inicial que se resuelve de una forma totalmente práctica, prescindiendo de la teoría y otorgando todo el protagonismo al quehacer práctico. Puede posibilitar que el alumnado plantee hipótesis respecto al problema inicial, pero las actividades planteadas tienen una secuencia muy guiada para que el alumno se limite a realizarlas siguiendo unas pautas determinadas sin pedir reflexiones teóricas. Se trata de planteamientos que poseen un gran potencial de convertirse en indagaciones, dado que los experimentos planteados son adecuados y las posibilidades de generar conocimientos y aprendizajes significativos en el alumnado son elevadas.

Académica: Es una U.D. donde se realiza un esfuerzo por plantear una situación problemática, aunque el resultado final dista de ser coherente. Los conceptos objeto de estudio se abordan de forma teórica y tradicional, a través de clases expositivas centradas en la figura del maestro. Puede proponer alguna actividad experimental, incluso con realización de mediciones, pero magistral o totalmente dirigida por el profesor. No contiene actividades prácticas en las que el papel del alumno sea relevante.

Proyecto: Se trata de una U.D. que a priori conecta la teoría con la práctica y en la que el alumnado tiende a tener un papel protagonista. No obstante, las actividades manipulativas están encaminadas hacia la elaboración de un artefacto determinado, siguiendo unos pasos que el profesor determina. La situación problemática no se retoma al final de la unidad puesto que se incide demasiado en las actividades tipo receta.

Investigadora: Son U.D. que parten de un problema inicial basado en el interés del alumnado. Éstos plantean hipótesis en función de sus ideas previas, diseñan experimentos y los llevan a cabo. Las actividades propuestas permitirían conectar la teoría con la práctica, así como generar conclusiones que den respuesta al problema inicial planteado. Finalmente, se retoma el problema inicial y se ofrece una posible respuesta. Sin embargo, a diferencia de una U.D. 'indagadora', el modelo de programación no acompaña al alumno en una gradación de dificultades hasta que alcance un trabajo independiente, pudiendo el alumno, sobre todo en primaria, perderse en este proceso.

Indagadora: En este tipo de U.D. se trabajan las ideas previas para introducir el problema objeto de estudio, que siempre están relacionados con aspectos del entorno e interés de los alumnos. Se realizan experimentos clave y secuenciados partiendo de conceptos básicos para el desarrollo del problema inicial. Asimismo, se discuten y comparan los resultados obtenidos, tratando de generar modelos similares a los aceptados por la comunidad científica. Destaca la presencia de preguntas que guían la temática antes, durante y después de la realización de los experimentos, y por último se retoman las ideas previas y las hipótesis planteadas para evidenciar su desarrollo y potenciar el aprendizaje significativo de los nuevos conceptos. En todo el proceso, el profesor escalona las actividades para que los alumnos adquieran progresivamente mayores niveles de autonomía. En definitiva, se trata de un modelo de programación que analiza las ideas previas del alumnado y su desarrollo proponiendo una situación problemática de interés entre los alumnos y que permite el análisis del entorno que les rodea a través del planteamiento de hipótesis, diseño y realización de experimentos, discusión de los resultados obtenidos y finalmente, la redacción y exposición de un informe.