Diagnosing conceptions about the epistemology of science: Contributions of a quantitative assessment methodology

[1]Ángel VÁZQUEZ-ALONSO, [1]María-Antonia MANASSERO-MAS, [2*]Antonio GARCÍA-CARMONA and [3]Marisa MONTESANO DE TALAVERA

[1]University of the Balearic Islands, SPAIN
[2]Department of Science and Social Science Education, University of Seville, SPAIN
[3]National Secretary of Science, Technology and Innovation, PANAMA

[*]Corresponding author. E-mail: garcia-carmona@us.es

Received 19 Aug., 2015
Revised 3 Jun., 2016

Contents

Abstract

This study applies a new quantitative methodological approach to diagnose epistemology conceptions in a large sample. The analyses use seven multiple-rating items on the epistemology of science drawn from the item pool Views on Science-Technology-Society (VOSTS). The bases of the new methodological diagnostic approach are the empirical psychometric scaling of the item sentences in

accordance with experts' criteria, the implementation of a multiple-rating model of answering (the respondent appraises each item sentence), and a scoring procedure that computes standardized indices from the multiple-rating responses and the sentence's scaling. These standardized indices represent the extent to which the respondent's conceptions about the epistemology of science are informed (the higher the index, the more informed the conception). The approach is applied to evaluate the conceptions of a large and diverse national sample through its standardized indices, which provide statistical hypothesis-testing across individual sentences, items, groups, or studies, computation of reliability indices, the correlations between items, and an exploratory factor analysis that may complement qualitative analysis. Finally, the most relevant features of the new approach, its potential for applications to teacher training and curricular development in the science classroom and the method's power to make easy, quick, and economic evaluations of conceptions about the nature of science, are discussed.

Keywords: epistemology of science; evaluation methodology; hypothesis testing; instrument reliability; nature of science (NOS)

Introduction

The history, philosophy and sociology of science (which include epistemology as an important part) have been extensively advocated as central contents in science education in order to provide students with a clearer understanding—a more accurate image—of science and improved future decision making in personal and social settings (Aikenhead, 2006). Most of the recent science education research on these interdisciplinary issues (including epistemology issues) has been labelled "nature of science" (NOS), which embraces a variety of areas related to the nature of scientific knowledge (epistemology of science, science community, the relationships between science, technology and society, socio-scientific issues), and many other related topics concerning their effective teaching and learning, methods, NOS teaching materials, evaluation of students' and teachers' conceptions, theoretical matters, teacher training, etc. (Coll, 2012; Lederman, 2007).

This paper focuses on the evaluation of epistemological conceptions, crossing various concerns of NOS research, teaching and learning. For instance, the controversial features of most epistemological topics make it difficult to devise valid

methods and instruments for their evaluation. Furthermore, an underlying problem of epistemology and NOS research is the incommensurability of the studies, either because the methods and instruments are quite different, or because of the qualitative nature of results. Thus, accurate individual profiles are relatively incomparable, beyond broad stereotyped results on the poverty of students' and teachers' epistemological and NOS conceptions. When using the same qualitative instrument (e.g. Lederman and colleagues' VNOS), only a rough percent of informed conceptions, developed through researcher-based criteria, are devised.

The aim of this paper is twofold. On the one hand, it presents a new methodological approach to evaluate epistemological conceptions which advances NOS research allowing specific comparisons and hypothesis testing. The approach is extensive, flexible, functional, meaningful and standardized, and it easily allows adaptable applications, statistical hypothesis testing between groups, across treatments, or over time. Thus, research studies can be compared on the same scoring baseline; scaling up to larger samples is faster, easier, and cheaper; and, in practice, its use by teachers for curricular development or classroom evaluation is straightforward. On the other hand, this paper illustrates these properties through a real assessment of some epistemology issues in a large, nationwide sample of Panamanian students and teachers, whose presence is scarce in science education research. Hypothesis testing and correlation analyses of the epistemological conceptions are also taken into consideration.

The Nature of Science as the global framework for epistemology in Science Education

Today, science education literature usually considers the epistemology of science under the NOS umbrella or, even more precisely, the nature of scientific knowledge (Lederman, 2007). NOS refers to the values, suppositions, scientific practices, community, society, and technology, etc. involved in scientific practices, which depict science as a human activity aimed at gaining valid knowledge. Scholars do not agree on a precise definition or delimitation of the NOS field, which is acknowledged as complex, controversial, multifaceted, and changing over time, although these disagreements do not impede researching or teaching NOS issues (Erduran & Dagher, 2014; Matthews, 2012). A simple approach characterizes NOS as a human way of gaining valid knowledge that is practised by a special community of professionals called scientists, who work under certain values and epistemological assumptions.

The importance of NOS for science education stems from being considered a core content of scientific literacy. Besides the traditional contents "of" science (facts, laws, theories, processes, inquiry, etc.), NOS embodies knowledge "about" science (Osborne, Collins, Ratcliffe, Millar & Duschl, 2003). NOS issues have been adopted as curriculum content in the reforms of science education around the world, and consequently, NOS topics should also be a part of science teacher education (Eurydice, 2011; Next Generation Science Standard [NGSS], 2013). The aim of NOS teaching in pre-college science education is not to train students to become philosophers of science or to address particular philosophical standpoints. Instead, in responding to the crucial role of NOS in scientific literacy, students should be able to understand how science works, and hence, to have a more solid foundation on which to base their future decision making in personal and social settings. In a way, the NGSS provides an enhanced, streamlined and renewed vision of the curricular NOS along two strands: the features associated with scientific and engineering practices (scientific research, methods, empirical evidence, openness to revision, scientific models, laws, mechanisms and theories), and some global suppositions of scientific knowledge, which are considered as curricular cross-cutting concepts (human enterprise, assumption of order and consistency for natural systems and limited to the natural and material world).

There is some controversy about the most suitable NOS contents to include in the curriculum at the pre-college level, though different scholarly proposals do share some coincidences. However, closed lists of topics might run the risk of inducing sterile rote learning if pedagogical development is inappropriately applied (Abd-El-Khalick, 2012a; Deng, Chen, Tsai & Chai, 2011). All in all, the issues on the epistemology of science are widely agreed as a core component of NOS and are the centre of this paper.

Evaluation of epistemology conceptions within the Nature of Science framework

Recent years have seen a major growth in research on the evaluation of student and teachers' conceptions about NOS and epistemology. Many empirical studies using different methods and instruments have consistently found a broad collection of deficits in epistemology views. Neither students nor teachers understand the role that theories, laws, hypotheses, models, creativity, technology, tentativeness and scientific methods play in science. Furthermore, the results are broadly coherent

across methods, countries and age groups, confirming the importance of the problem (Lederman, 2007; García-Carmona, Vázquez & Manassero, 2012).

Science teachers' understanding of epistemology unfortunately reflects similar naïve patterns to those observed in students. They hold mythical conceptions about science, which reject the theory-laden, tentativeness and differences between scientific theories, laws, and hypotheses, and the status of scientific method(s), inference, observation, and empirical evidence (e.g., Abd-El-Khalick & Lederman, 2000; Celik & Bayrakçeken, 2006; García-Carmona, Vázquez & Manassero, 2011; Lederman, 2007).

Most diagnostic studies of conceptions have been performed using small, or convenience, samples of science participants. Recently however, some studies have started to use larger samples tied to applications of VOSTS-related instruments (e.g. Dogan & Abd-El-Khalick, 2008). Further, Holbrook et al. (2006) studied non-science students, and Liu and Tsai (2008) compared arts and science graduate students (including an initial teacher education group). In the latter, the two groups were generally not found to differ from each other, although the science students displayed less sophisticated beliefs (i.e., about the cultural dependency of scientific theories), and the science teacher education students scored lowest on all dimensions.

Educational evaluation has grown into a vast field of research. The numerous methods and instruments basically fall into two broad categories: qualitative (case studies, participant observation, interviews, open questionnaires, content analysis of lesson plans and classroom documents, concept maps, discourse analysis, etc.) and quantitative (ordinal Likert-type scales, multiple-choice and multiple-rating questionnaires, grids, etc.). Mainstream NOS research has drawn on this field to develop various methods of evaluating NOS conceptions (reviewed in Deng et al., 2011; Lederman, 2007; Liu, 2012).

The qualitative approach has its own unquestionable merits to penetrate the complex web of individuals' ideas; although it also suffers from semantic problems, the categories of analysis are often idiosyncratic, not very explicitly defined, and hardly ever equivalent among studies (Deng et al., 2011). Even though their results depict broad patterns of the NOS conceptions, they are hardly comparable, and have limited influence on the inclusion of NOS in school science assessments and on encouraging

schools to teach NOS in daily practice because they have targeted a readership by research specialists, far removed from school teachers (Chen, 2006). All in all, qualitative research provides valuable and unquestionable contributions, so that the previous criticisms are not intended to devalue it in any case, but only to frame some of its shortcomings.

On the other hand, reliance on quantitative scales and questionnaires has produced criticisms pointing to methodological shortcomings and poor validity or reliability. The researchers' perspective (philosophical preferences, biases and prejudices) of instrument construction may restrict its validity; for example, the adoption of cluster labels (relativist, constructivist, empiricist, etc.) to classify individuals (this is also a problem in qualitative research). The immaculate perception hypothesis (the implicit assumption that researcher and respondents perceive and understand the items in the same way) may severely affect the validity of investigator-designed instruments (Aikenhead, Fleming & Ryan, 1987; Lederman & O'Malley, 1990; Lederman, 2007). Other common criticisms refer to the scoring procedures, the underlying dimensionality of the models, the representativeness of the scores and the reliability statistics. Forced-choice instruments, in particular, limit the space of responses available to respondents (Lederman, Abd-El-Khalick, Bell & Schwartz, 2002).

The above criticisms about validity could also apply to qualitative research, as the qualitative processing of participants' open productions is often insufficiently detailed by researchers, thus preventing semantic and validity issues from becoming ostensible. This reflection intends to redress an apparent imbalance in research between qualitative and quantitative methods because the greater criticisms of the latter may be hindering them (Guerra-Ramos, 2012). The review of over one hundred research studies on students' NOS conceptions by Liu (2012) estimates the proportion of qualitative (two-thirds) to quantitative methods (one-third) used in research. Most studies (54%) combine several (two or more) methods to acquire their data, while the rest (46%) use just a single method. Overall, 81% of the data acquisition methods are qualitative, indicating the prevalence of the qualitative over the quantitative approach in current NOS research. Instead, we advocate trying to bridge the gap between the two methods because they can also complement each other in providing valuable information about the complex aspects of NOS conceptions through their different approaches to seeking evidence. Indeed, it is usually recommended to complement test scores with qualitative methods (interviews, observation, etc.) in order to better unveil the respondent's real

conceptions (Aikenhead & Ryan, 1992; Chen, 2006; Lederman et al., 2002). This complementary approach to the qualitative/quantitative evaluation instrumentation has also been initiated from the qualitative facet through the work of Brunner, Summers, Myers and Abd-El-Khalick (2016), who try to quantify the responses of the most widely used qualitative evaluation tool (VNOS).

The notion of authentic evaluation has been introduced into general education to aid in the evaluation of complex learning, such as performances or actions in real settings ("close to real"). In this framework, and in light of some criticisms of the VNOS questionnaire (Lederman et al., 2002), Allchin (2011) recently argued for applying the criteria of authentic evaluation to NOS conceptions, highlighting the complexities of teaching and evaluating NOS.

Recent evaluation instruments

Science education research needs standardized, valid and reliable instruments to evaluate NOS for diverse reasons: to provide trustworthy common grounds for research results and to foster NOS teaching, providing practical tools for teachers (Chen, 2006; Lederman, 2007). Partial accounts of quantitative evaluation instruments in the literature have contributed to the invisibility of some available instruments. Lederman (2007) displays a huge list of instruments for the period 1954-1992, although for recent years, he just refers to the five-form VNOS, the 114-item Views on Science-Technology-Society (VOSTS) (Aikenhead & Ryan, 1992) and the Critical Incidents Scale (Nott & Wellington, 1995). Liu's review (Liu, 2012) adds three different instruments: Views about Sciences Survey (VASS) (Halloun & Hestenes, 1998), Thinking about Science Survey Instrument (TSSI) (Cobern & Loving, 2002) and Views on Science and Education (VOSE) (Chen, 2006). Some additional questionnaires are listed in the table of Appendix A.

Though Lederman's VNOS may be the most influential qualitative instrument, Aikenhead's VOSTS item pool has also inspired a considerable amount of studies and some of the mentioned instruments. The teachers in Chen's (2006) study found that it was harder, more frustrating, and required a greater effort to answer the VNOS in the time allocated than to answer the VOSE (a VOSTS-based instrument). The standardized instrument devised in this paper, based on the VOSTS pool, is oriented towards coping with some current challenges of NOS research and teaching (assessment), which in turn provide reasons to choose quantitative instruments to assess NOS conceptions.

1. The instrument explicitly shows all the items, explains the method to obtain the scores, the interpretations of the scores, and its theoretical foundations. Such an explicit design allows straightforward and extensive use, replications, instrument improvements and associated data management through critical analysis of the results.
2. A standardized instrument usually involves procedures that are inexpensive, rapid and easy to apply. These features make large-scale evaluations feasible at a state or national level, thus making the monitoring process more robust and representative (Kind & Barmby, 2011; Chen et al., 2013). Open-ended instruments, on the contrary, require idiosyncratic, expensive, tedious and slow processes that are managed by scholars.
3. Standardized instruments would facilitate teachers' evaluation tasks in the classroom, and consequently, are likely to stimulate teachers to incorporate NOS teaching into curricula, as their reluctance to teach NOS explicitly is partially due to the lack of evaluation instruments (Lederman, 2007). Particularly, the item pool used here is large enough for different instruments to be tailored to different applications and objectives by choosing the appropriate items from the pool.
4. The standardized instrument provides researchers with a tool to delve deeper into the statistical analysis of data (group comparisons, time series, individual profiles, test-retest follow-up, correlation methods, strengths and weaknesses, inconsistencies and consistencies, etc.); this paper exemplifies this point. Furthermore, the relationships between NOS conceptions and other educational variables (learning, teaching, teacher attitudes, motivation, etc.) can be more readily examined (Deng et al., 2011).
5. Standardization is especially well suited to compare individual profiles of respondents' NOS conceptions, which facilitate researcher and teachers' evaluations of students, thus fostering the progress of NOS research and teaching.
6. The contrast among different research studies cannot currently be solved due to the multiplicity of approaches. Standardized instruments would provide researchers with common grounds, which could make it possible to compare research findings from different studies, groups and countries. Indeed, this reason does not try to unify the field though it advances NOS research, for instance, in some of the critical lines identified by Lederman (2007) for future NOS research (effectiveness of interventions, development over time, factors that affect development, change of teachers' conceptions during classroom practice, etc.).

In response to the criticisms of quantitative instruments to cope with the challenges of assessment in NOS research and to faithfully represent the respondents' conceptions, the present study uses an instrument that is empirically developed (largely avoiding the immaculate perception) and based on a multiple-rating response model, which eludes forced choice (Vázquez & Manassero, 1999). The research questions of this paper refer to its twofold aims: does the evaluation instrument and its associated methodology allow for a new, valid, reliable, fast, effective, inexpensive and standardized evaluation of NOS conceptions on epistemology of science? Second, what are the Panamanian student and teachers' conceptions on the epistemology of science? In particular, the development of statistical hypothesis testing (e.g., comparative group analyses, time series, profiles, etc.) and correlation analyses between variables (e.g., teaching factors, factor analysis, etc.) are more straightforward to perform with this instrument than with other instruments and methodological approaches.

Method

Participants and context

Science education in Panama emphasises environmental education to strengthen the younger generations' awareness of their responsibilities in the careful management of their country's rich natural and environmental resources, though it currently intends to take part of the worldwide trend of teaching epistemology and NOS issues at all levels of the country's educational system by promoting in-service training teachers in NOS issues.

Within this framework, the present study was possible thanks to the voluntary collaboration across the whole country of many Panamanian students and teachers, who were invited to participate by SENACYT (the governmental agency for science and technology). The participants randomly and anonymously responded to one of two different questionnaires (Form 1 and Form 2) that had been constructed by a research team to cover the entire theoretical framework of VOSTS (see the following subsection) to develop a research project across several countries and to avoid respondents' fatigue through the adequate length of each Form. The present study accounts for the responses obtained for the seven multiple-rating items on

epistemology of science included in both forms; Form 1 and Form 2 were validly answered by 1887 and 1885 participants, respectively.

The study targeted four groups of the population: high-school students preparing to start university or first-year undergraduates (typically average age 17-18 years, labelled here as "young students" – 61%), final-year undergraduates or new graduates ("veteran students" – 24%), teachers in their initial training ("pre-service teachers" – 7%), and practising teachers ("in-service teachers" – 24%). The above percentages sum higher than 100% because many individuals belonged to two groups that were not incompatible (e.g., an undergraduate student preparing to be a teacher was included within student groups and within the pre-service teacher group).

The age ranges from 16 years (youngest student) to 70 (oldest teacher) years, and the distribution by gender was roughly even (men: 52%; women: 48%). The in-service teachers had a mean experience of about 15 years, and their distribution over the levels was primary (8%), secondary (30%), and university (62%). There were both science (66%) and non-science – humanities – (34%) specialities in every group. The large sample is representative of Panamanian student and teacher groups (alpha = 95.5%; $e < \pm 3\%$; $p = q = 50\%$).

Research instrument

Seven NOS items on epistemology were drawn from the pool "Views on Science-Technology-Society" – VOSTS – (Aikenhead, Ryan & Fleming, 1989) after its faithful translation and careful adaptation to the Spanish language and cultural context. The original VOSTS items were empirically developed from student and teachers' open responses and interviews to questions, which were then systematized into a multiple-choice format by researchers (Aikenhead & Ryan, 1992) who consider that the empirical process overcomes the immaculate perception objection and is equivalent to a pilot testing, thus endowing the instrument with intrinsic construct validity. Further, Lederman, Wade and Bell (1998) consider VOSTS to be a valid and reliable instrument to investigate standpoints concerning the NOS. Additionally, empirical reliability was first established by Botton and Brown (1998) using a unique response model (respondents make a forced choice for one sentence in each item).

The seven items used in this study have a multiple-rating format (see Appendix B). The item stem presents an epistemology issue, which is followed by multiple sentences, each one explaining a reason that develops a particular position (belief) on the stem issue and sorted with a label A, B, C, D, E…. Both the stem and the sentences use a simple, common, non-technical language style, as they were empirically developed from students' answers. Further, the set of sentences does not reflect any particular philosophical standpoint, instead the whole set of sentences tries to cover a wide range of different positions, which supports a balanced evaluation of respondents' conceptions for each item. The set of the respondent's indices on the sentences yields his/her evaluation profile on the item issue. A total of 36 epistemological sentences (16 for the three items of Form 1, and 20 for Form 2) were rated by the respondents (Appendix B).

The items are labelled with a five-digit number, and each sentence within an item is identified by the item number plus the letter that labels the sentence position within the item (e.g., 90521D means Sentence D within Item 90521). Some sentences have an additional coding _C_ prepended to the tag number, indicating that the sentence represents an idea about which a group of expert judges strongly agreed on the category assigned to that sentence, whose details are given elsewhere (Vázquez, Manassero & Acevedo, 2006).

Procedures

The new model of multiple-rating responses and its methodological approach that enable its application to evaluate NOS conceptions were developed through a series of prior stages, whose complexity do not try to emulate any naïve step-model of scientific method:

- The translation and adaptation of items through back-translation processes into Spanish ("Questionnaire of Views on Science, Technology, and Society", Spanish acronym COCTS).
- The limitations of the single response model, which had been extensively applied by VOSTS users, on validity and information scores (Vázquez & Manassero, 1999).
- The scaling of sentences into one of a 3-category scheme (Adequate, Plausible, or Naïve) by a panel of expert judges (Vázquez, Manassero & Acevedo, 2006), according to the concomitance between the sentence's content and contemporary

scholars in history, philosophy, and sociology of science and technology (HPSST). The scheme is similar to those suggested by other researchers (Rubba, Schoneweg-Bradford & Harkness, 1993; Tedman & Keeves, 2001): Adequate (A) – the sentence expresses a fully acceptable NOS conception; Plausible (P) – although not totally adequate, the sentence expresses some acceptable aspects; and Naïve (N) – the sentence expresses a conception that is neither adequate nor plausible. The final letter of each sentence codes its assigned category (e.g., 90521D_A_ means that Sentence D of Item 90521 belongs to the Adequate category).

- The design of a new multiple response model (MRM) in which the respondents rate their degree of agreement with all the sentences within the item. The MRM avoids the disadvantages of the "forced" choice and maximizes the information about the respondent's thinking about the question issue (Vázquez & Manassero, 1999).
- The construction of a metric that provides a normalized, homogeneous and invariant index [-1, +1] for each sentence, whose value is computed taking into account the respondent's rating and the sentence's category. These indices are averaged on the basis of their sentence indices to compute item indices that efficiently summarize the respondent's conception of an item, (Manassero, Vázquez & Acevedo, 2003a, 2003b).

Response and metric

The MRM asks participants to express their agreement/disagreement with each item sentence on a nine-point scale (1 to 9, disagreement to agreement). If a respondent does not wish to answer, he/she may choose one of two reasons ("I do not understand the issue" or "I do not have sufficient knowledge about the issue") or leave it blank (opening the MRM possibilities to avoid forced choice).

Each raw rating score (1-9) is transformed into a homogeneous invariant normalized response index in the interval [-1, +1] through a scaling procedure that takes into account the category of the sentence (Adequate, Plausible, Naïve) as previously assigned by a panel of expert judges (further details in Vázquez, Manassero & Acevedo, 2006). For instance, an adequate sentence expresses an appropriate view on the issue, so that the scaling procedure assigns the top index score +1 to total agreement (9) and the bottom score -1 to total disagreement (1), and proportionally for the in-between scores. A naïve sentence expresses a view that is neither

adequate nor plausible, so that the scaling assigns the inverse scoring index to those of the adequate sentences. A plausible sentence assigns the +1 scoring index to the middle raw score (5) and -1 to the two extremes (1 and 9), and proportionally for the in-between scores (see Table 1). This three-category scaling for Likert multi-rating responses not only avoids forcing choices, but also guessing the "right" response patterns (Eagly & Chaiken, 1993) and is similar to the three-point rubric recently applied in other papers (e.g., Akerson & Donnelly, 2010).

Table 1: Correspondence between the respondent's direct score on each sentence and its scaling transformation, according to the category of each sentence (Adequate, Plausible, or Naïve), into the normalized index for the sentence, which represents the standardized value of the respondent's belief about each sentence.

Direct Score Scale [1-9] (Respondent's Degree of Agreement)									
	Total	Near Total	High	Partial High	Partial	Partial Low	Low	Near Null	Null
	9	8	7	6	5	4	3	2	1
Category of the sentence	***Scaling Transformation into Normalized Indices [-1, +1] (Standardized Value of Belief)***								
Adequate	1	0.75	0.50	0.25	0	-0.25	-0.50	-0.75	-1
Plausible	-1	-0.50	0	0.50	1	0.50	0	-0.50	-1
Naïve	-1	-0.75	-0.50	-0.25	0	0.25	0.50	0.75	1

The value of the index represents the degree of match between the respondent's opinion, as originally expressed through the raw agreement scores, and the current views of HPSST experts. The higher (lower) the index is, the stronger (weaker) the match, regardless of the category of the sentence (this is the invariance property of indices). Thus, the closer an index is to the maximum positive value (+1), the better informed (closer to the experts' views on NOS) is the respondent's conception, whereas the closer it is to the negative value (-1), the more misinformed (distant from current NOS conceptions) is the respondent's conception. As misinformed conceptions are associated with the lowest negative values of the index, and informed conceptions are represented by the highest positive values of the index, for brevity they will simply be referred to as "negative" or "positive", without implying any bias in the meaning of these words.

The sentence's indices form the basis for further computations and statistical analyses. For instance, the average of the item sentence indices yields the global weighted item index, which constitutes a quantitative evaluation of the overall conception of the item issue.

Table 2. Item labels and issues about the nature of science displayed across the two questionnaire forms (Form 1 and Form 2) together with their reliability parameters.

Form 1 (F1) Items	Reliability[a]	Form 2 (F2) Items	Reliability[a]
F1_90211 scientific models (inference)	0.611	F2_90111 observations (theory-laden)	0.477
F1_90411 tentativeness	0.520	F2_90311 classification schemes (inference)	0.661
F1_90621 scientific method	0.586	F2_90521 role of assumptions (hypothesis, theories, laws)	0.649
		F2_91011 epistemological status (creativity)	0.684

Note: a (superscript) Cronbach's Aplha

The reliability coefficients (Cronbach's alpha) computed from the raw scores of all sentences is excellent (0.922 for Form 1, and 0.925 for Form 2). The reliability coefficients of the seven epistemological items (computed from the few sentence scores belonging to each item) are obviously lower (Table 2), due to the mechanical effect of the sharp diminution of the number of sentences that contribute to single item reliability coefficient.

Statistical analysis

The indices provide a homogeneous, invariant, and normalized meaning for the scores across all sentences and items, e.g., they provide a measure of the magnitude of the correctness of a conception. The index scores allow various further computations to be made – averaging and relating different variables and applying inferential statistics for hypothesis testing, group comparisons, or to establish cut-off points defining different achievement levels (Vázquez, Manassero & Acevedo, 2006). Inferential statistics are usually analysed in terms of probabilistic measurements (p-values) of the significance of any apparent differences found.

Given that p-values carry no information about the relevance of the magnitude of the differences, the effect size parameter (the difference between means expressed in standard deviation units) is often used to this end. The resulting values are usually

interpreted on the basis of some simple criterion (Cohen, 1988), classifying differences into intervals labelled as trivial (d < .10), small (d < .20), medium (d < .50), large (d < .80), etc. For a large sample, an effect size over 0.30 usually corresponds to statistically significant differences (p < .01). Therefore, we shall henceforth use the term "relevant" to refer to differences that satisfy both an effect size over 0.30 and statistical significance (p < .01), which means they represent some practical educational value. Differences that do not pass this threshold will be considered "irrelevant", even though they still might be statistically significant or interesting from other perspectives (e.g., personnel evaluation).

Results

The participants' overall viewpoints on the seven NOS (epistemology of science) issues are represented by the mean item indices (Table 3). The grand mean for all the items was modest, e.g., close to zero, and none of them were far enough from zero to meet the effect size relevance criterion (d > .30) to consider them relevantly adequate. The most positive mean item indices pertained to observations and tentativeness, and the most negative to the scientific method and the role of assumptions in science.

Overall, the sample presents insufficient or misinformed conceptions on the seven epistemological issues that were inquired into, as the grand mean indices for all issues were around zero. Within this overall poor profile, the scientific method and the role of assumptions in scientific knowledge (laws, theories, etc.) presented relatively lower negative indices, while the views on observation and the tentative nature of science scored slightly positively. The remaining epistemological issues (scientific models, classification, and epistemological status) had intermediate mean indices (scores close to zero).

Table 3. Descriptive statistics of the epistemology of science items for the whole sample (total) and for the groups of science and humanities; the last column displays the effect size of the differences between the two groups (difference expressed in mean standard deviation units).

		Science			Humanities			Total			Effect Size (*d*)
		n	***M***	***SD***	***n***	***M***	***SD***	***n***	***M***	***SD***	
F1_90211	Scientific	1236	-0.005	0.285	651	-0.004	0.294	1887	-0.005	0.288	-0.003

Models										
F1_90411 Tentativeness	1236	0.031	0.267	642	0.036	0.280	1878	0.033	0.271	-0.018
F1_90621 Scientific Method	1235	-0.063	0.245	648	-0.078	0.259	1883	-0.068	0.250	0.059
F2_90111 Observations	1308	0.046	0.322	577	0.078	0.313	1885	0.055	0.319	-0.101
F2_90311 Classification Schemes	1304	0.005	0.249	573	-0.008	0.260	1877	0.001	0.252	0.051
F2_90521 Role of Assumptions	1298	-0.060	0.306	573	-0.080	0.294	1871	-0.066	0.302	0.067
F2_91011 Epistemological Status	1301	-0.030	0.266	572	-0.034	0.256	1873	-0.031	0.263	0.015

Note: Positive effect size means the science group scores higher than the humanities group and vice-versa (Differences are usually deemed relevant if $d > .30$).

The former empirical analysis could be extended to the item sentences for identifying the specific strengths and weaknesses on each question to deepen our comprehension about how a group actually understands each epistemological aspect. Due to space limitations, just an example for the group of science teachers on the three issues of Form 1 is developed. Panamanian science teachers hold informed conceptions, as they strongly support some key sentences about the change of scientific knowledge (90411C …because the interpretation or the application of the old facts can change) and scientific method (90621C… scientific method is useful in many instances, but it does not ensure results. Thus, the best scientists will also use originality and creativity). Besides, Panamanian science teachers also hold misinformed conceptions about the same issues (as they support naïve sentences) on change of scientific knowledge (90411D … because new knowledge is added on to old knowledge, the old knowledge doesn't change) and scientific method (90621A … the scientific method ensures valid, clear, logical and accurate results. Thus, most scientists will follow the steps of the scientific method; and 90621B … the scientific method should work well for most scientists, based on what we learned in school).

Validity issues

As previously mentioned, the language and sentence wording of the items is simple and non-technical, as they arose from their original empirical development. Both forms have been previously applied to similar samples in some Latin neighbouring countries of Panama, where no problems in wording or understanding were found. Nonetheless, prior to the large application, both forms were also piloted for

comprehension and cultural adequacy with small samples of Panamanian students and teachers during a science methods workshop in Panama, without any significant remarks. After the implementation to the large Panamanian sample, the rates of blank answers were also compared with those of the neighbouring countries and were found to be similar. All in all, this set of indicators supports the cross-cultural and content validity of the instruments.

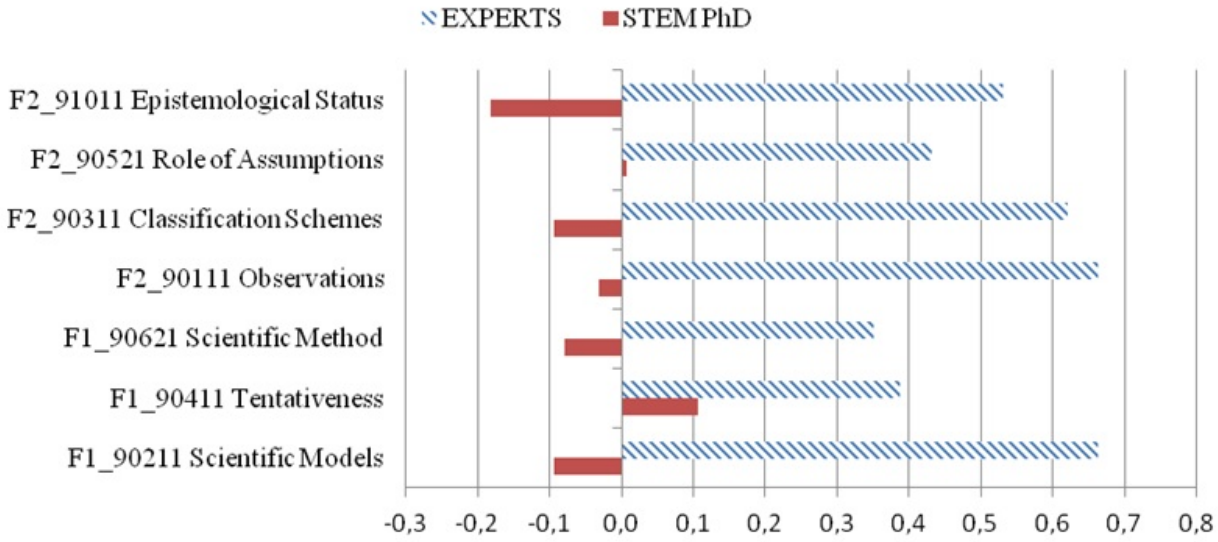

Figure 1. Average standardized indices of the seven epistemological items across the two groups compared for testing the discriminant validity of the items (a group of experts in epistemology of science and the group of Panamanian experts, with the same academic level of experts).

Further, the concept or discriminant validity of the epistemological items was tested by comparing the responses of a group of eleven experts in epistemology of science against a group mined from the large Panamanian sample under the condition of being academically equivalent to the expert group. Thus, twenty-five Panamanians with a Ph.D. in science and engineering (eleven for Form 1 and fourteen for Form 2) were drawn. The average standardized indices in Figure 1 for the seven epistemological items in both groups (experts and STEM) show that the indices of the experts scored much higher than the Panamanian STEM Ph.D. group across all seven epistemological items. This result, which clearly distinguishes experts from non-experts by controlling for academic level, supports the construct and discriminant validity of the items.

Correlation analyses

The interrelatedness between diverse issues of epistemology of science is a commonplace in NOS research and a feature of its complexity. However, it has rarely been investigated due in part to the dominance of qualitative methodologies. The relationships between the epistemology items and sentences and the various conceptions are examined by means of correlation analyses that are constrained to the two independent samples (Form1 and 2).

The correlations between almost all the epistemology item indices were positive and small, though statistically significant. For instance, the Pearson coefficient between scientific method and scientific models ($r = 0.253$, $n = 1871$, $p < 0.000$) means that higher levels of understanding the scientific method are associated with higher levels of understanding scientific models. The same positive correlation pattern applies to the remaining pairs of items (Table 4).

However, the item on observations is the one exception to this positive pattern, as this item has insubstantial correlations with the other epistemology items. This suggests that views on observations are poorly related to the other epistemological conceptions, namely, classification schemes, the role of assumptions, and epistemological status.

Table 4. *Correlation coefficients between the epistemology of science items. Coefficients of the three items of Form 1 are shown under the diagonal, and those of the four items of Form 2 are shown above the diagonal.*

		F2_90311 classification schemes	**F2_90521 role of assumptions**	**F2_91011 epistemological status**	
		0.001	-0.038	0.021	F2_90111 observations
F1_90411 tentativeness	0.152*		0.247*	0.269*	F2_90311 classification schemes
F1_90621 scientific method	0.253*	0.187*		0.240*	F2_90521 role of assumptions
	F1_90211 scientific models	F1_90411 tentativeness			

Note: * correlation is significant at $p < .01$ (bilateral).

Many of the correlations between the NOS sentence indices (not displayed here due to lack of space) were large to moderate and positive (e.g., the strongest correlation was between sentences E and F within item F1_90211 about scientific models: $r = 0.618$, $n = 1806$, $p < .0000$). There were also a few small and negative NOS sentence index correlations (e.g., $r = -0.235$, $n = 1821$, $p < .0000$ between sentences 90621B on the scientific method and 90411B on tentativeness), which means that better conceptions on sentence B of scientific method are associated with worse conceptions on sentence B of tentativeness).

Given the large number of sentence variables, correlation techniques aimed at reducing the number of these variables can help one to understand the complexity of the relationships between the single sentences about epistemology. To this end, the indices of the 16 Form-1 and 20 Form-2 sentences were studied separately through a principal component analysis (PCA), using SPSS Version 18. The prior evaluation of data suitability for PCA supported the factorability of the correlation matrix. Many correlation coefficients were greater than 0.30, the values of the Kaiser-Meyer-Olkin measure of sampling adequacy exceeded the recommended threshold of 0.6 (0.765 for Form 1 and 0.802 for Form 2), and Bartlett's sphericity test reached statistical significance ($p < .000$).

The PCA yielded five (Form 1) and six (Form 2) components with eigenvalues exceeding 1, explaining, respectively 14.3%, 12.3%, 11.0%, 10.9%, and 8.1% of the common variance for Form 1 and 10.4%, 9.7%, 9.3%, 9.0%, 8.5%, and 7.9% for Form 2. Inspection of the scree plot showed notable breaks after the first and third components. Using Cattell's scree test, we decided to retain three components for further investigation. The three-factor solutions for the two forms following oblimin rotation explained 43% (Form 1) and 39% (Form 2) of the common variance of the sentence indices (see table of Appendix C).

The interpretation of the factors of Form 1 suggested that each of them was mainly associated with one of the three categories used to scale the sentences according to their content. In particular, the first factor contains all the sentences categorized as naïve, the second factor groups the plausible sentences, and the third one the adequate sentences. The exceptions to this pattern are two adequate sentences that appear within the second factor, although their low loadings suggest that this assignation would require further analysis.

The interpretation of the factors of Form 2 is quite conditioned by the singular role played by item 90111 (observations), whose sentences load onto a single factor that is unique for this item, with the remaining sentences distributed between the other two factors. The second factor contains mainly the sentences categorized as naïve (except for one with a low loading), and the first factor contains mainly the plausible sentences (except for one with a low loading) and the adequate sentences.

When separate similar PCA analyses were performed for each of the four groups of the sample (young students, veteran student, pre-service teachers and teachers), the above structures were repeated across groups with only very small variations. This result suggests that the factor structures are stable in describing the participants' epistemology conceptions, regardless of the kind of respondents being considered.

Overall, the loadings of the adequate sentences in the factor structure were generally opposite in sign to those of the naïve and plausible sentences – a surprising negative correlation. This finding implies that high levels of comprehension of adequate sentences are associated with low levels of comprehension of plausible or naïve sentences, and, vice versa, that low levels of comprehension of adequate sentences are associated with high levels of comprehension of plausible or naïve sentences. The simple correlation between the plausible and naïve sentences is positive, meaning that high levels of comprehension of plausible sentences are associated with high levels of comprehension of naïve sentences.

From a logical perspective, the former is an apparently anomalous result because it indicates that the clearer a respondent's identification of the adequate sentences is, the harder he or she finds it to clarify other sentences that are just plausible or naïve. In other words, what would seem to be the logical implication of the recognition of adequate sentences, e.g., the rejection of naïve sentences or the partial recognition of plausible sentences, does not emerge from the empirical results of the principal component analysis; indeed, quite the contrary is the case. This anomalous correlation suggests some superficiality in the respondents' comprehension of NOS, in the sense that they make no use of simple logical reasoning when assessing the sentences, as a high valuation of an adequate sentence should logically imply a low valuation of the naïve sentences. A possible explanation is that the complexity involved in the comprehension of NOS leads to the respondents' difficulties to distinguish the opposite sentences within a given item that are expressions of contrary epistemological positions and, therefore, to value them differently.

All in all, correlation analysis quantitatively contributes to shed light on the complexity of the relationships among different NOS ideas, unanimously held by the researchers, although scarcely supported by data.

Discussion and Conclusions

This paper evaluates the conceptions that Panamanian students and teachers hold about the epistemology of science, which constitute an essential component of scientific literacy and an indicator of the comprehension of NOS, as a means to present the applicative potential of the new quantitative assessment methodology. One objective of the work was to serve as an example of the application and partial validation of the instrument and its associated methodological approach to diagnose epistemological conceptions (for this exemplification reason, only seven items were taken into account). Summing up, the present quantitative methodological approach to evaluating NOS conceptions based on the MRM model provides fuller (based on ratings along a wide spectrum of positions), sounder (contextualized on a specific frame) and more accurate (measured by sensitive indices) information about the respondent's views of a NOS issue than would a single response model. Furthermore, as the evaluation of the item is constructed from the scores on all of its sentences, the set of invariant multiple indices (sentence and average item indices) constitute global, valid, and reliable quantitative data that allow the application of statistical hypothesis testing procedures. Thus, the method guarantees the comparability and straightforwardness of the results, from which their qualitative analysis and subsequent discussion flow naturally.

Overall, the large sample exhibited inadequate or misinformed conceptions about the seven epistemological issues that were investigated because the average scores for all these issues were close to zero. Within this overall low profile, the lowest specific profiles were those of method and of the role of assumptions in scientific knowledge (laws, theories, etc.), whereas conceptions concerning observations and the tentative nature of science were slightly better. The remaining epistemological issues (scientific models, classification and epistemological status) had intermediate index scores, closer to zero than the other four.

Of course, the overall low item indices represent a general estimate of the comprehension of NOS because they are calculated as averages across all

participants and all sentences within the items. However, the distribution of the personal mean item indices among the sample of respondents was more variable, with some participants presenting well-informed conceptions while others presented clearly misinformed conceptions. It is important to note that indices are quasi-continuum assessment parameters that go beyond the usual simplistic right/wrong or informed/misinformed classifications of NOS conceptions leaving room for controversy and complexity. In addition, the mean indices of the sentences of the items showed an even greater variation among the respondents, again with some participants presenting well-founded beliefs and others poorly informed beliefs; the qualitative analysis of individual profile answers may raise the “why” of personal understandings (explanations) and cultural interpretations and idiosyncrasies, as qualitative research does. These results therefore mean that the generally negative picture of teachers' and students' epistemological thinking that is transmitted by most of the studies mentioned in the introduction should be much more nuanced. In particular, within the different items and among the different respondents, both well- and ill-informed views about the epistemology of science can coexist, a general fact about NOS conceptions that is developed in depth elsewhere (Vázquez, García-Carmona, Manassero, & Bennàssar, 2013).

The multiple-rating instrument used in the present work and the methodological approach offer an economical, fast, and effective form of inquiring into people's conceptions about NOS, and can have various useful applications. The information obtained from the respondents by means of the instrument does not come from generic or abstract questions (for example, "Are scientific models copies of reality?"), and the respondents are not obliged to choose one sentence and ignore the others. Instead, each question is straightforward and specific, and is presented in a context. The respondents are thus being asked to evaluate simple, non-technical sentences expressing different positions on the issue. The set of index scores they assign to all the sentences portray their personal position, which naturally translates into a NOS personal profile. Aikenhead and Ryan (1992) claimed that the original empirical qualitative construction of the pool (the sentences were created from participants' open responses, and are expressed in plain language) warrants an inherent validity of the item pool; further, the new method and the factor analysis lend some empirical support to its validity. The excellent reliability coefficients of the whole set of items as well as the moderate coefficients for the items underpin evaluation through the instrument. All in all, it must be underlined that the items’

empirical construction and methodology does not reflect the researchers' opinions, but that of experts, and then the simple language contributes to overcome the flaws associated with the instruments (e.g. immaculate perception, etc.).

Furthermore, the standardization of precise, homogeneous, and invariant indices with which to accurately evaluate NOS conceptions constitutes a common measure, allowing the use of statistical hypothesis testing, comparison of results of different research studies, and correlation analyses. This set of features seems to represent major advantages for future research because they provide absolute (index scores) and relative (group comparison) criteria with which to evaluate NOS conceptions. In particular, the indices allow statistical tests of hypotheses (e.g., comparisons between groups and between researchers), and the explicit and standardized interpretation of the indices facilitates contrasting the results with other methods of evaluation (e.g., qualitative instruments) as well as showing whether there is real improvement of the validity of the evaluation. For instance, the overall comparison between science and humanities groups does not display significant differences (an unreasonable expectation), which points to the inefficacy of science education to improve NOS understanding (Liu & Tsai, 2008; Vázquez, García-Carmona, Manassero, & Bennàssar, 2014), in this case of Panamanian education.

The method and tools are applicable to large samples without major increases in cost and time. They provide standardized numerical indices reflecting how much (or how little) and how well (or how poorly) each person knows and thinks about the different characteristics of NOS. As has been shown by other researchers (e.g., Dogan & Abd-El-Khalick, 2008; Chen et al., 2013), the above considerations are crucial for the feasibility and planning of diagnostic evaluations of large samples without requiring a major investment in time and resources. The large sample of participants in this study is representative of Panamanian students and teachers, thereby demonstrating the instrument's capacity to perform representative studies with a minimal investment of time and resources, making it suitable for comparing different groups or different researchers, or for tracking participants' conceptions over time (Kind & Barmby, 2011).

The controversy between Allchin (2011) and Schwartz et al. (2012) on authentic assessment of complex NOS comprehension has shed light on the strengths and weaknesses of the different NOS evaluation proposals as well as on the objectives, contents, and methods of NOS teaching. The approach to NOS assessment offered in

the present study also satisfies some of the features that Allchin (2011) assigns to the functional and authentic evaluation of NOS knowledge. It is authentic, as the stem of the items provides a contextual framework for the respondents, which situates a specific socio-scientific NOS issue close to reality; and it is functional, as it could be easily applied and tailored by researchers and teachers. Although the participants do not strictly compose their own written responses, the soundness of the analysis arises from the complete information provided by all the sentences of each item and the multiple-rating response to profile; of course, qualitative data from post-test interviews or freely drafted responses to the items constitute a natural complement of this method yet to be explored. Adaptability to diagnosis, to teacher training, or to general contexts of performance evaluation at school are ensured by the capacity for items to be tailored from the total pool and by the construction of individual NOS knowledge profiles based on the set of indices shown in more detail elsewhere (Vázquez, Manassero & Acevedo, 2006). In addition, the instrument's adaptability to various comparative uses (between individuals, groups, or locations), and the different stakeholders is quite evident because the quantitative method and the normalization of the indices allows the application of statistical hypothesis testing for comparative purposes; applications with large scale samples are rapid, low cost, and highly flexible and adaptable.

Furthermore, the use of this standardized method and instrument enables different research studies to be compared. To date, most studies have used non-equivalent methods which only allow coarse-grained comparison of the key ideas or results concerning NOS. The present standardized instrument and method can therefore contribute to NOS research by encouraging its synergic development, as suggested by Abd-El-Khalick (2012b) because they permit the results of different researchers to be compared, the establishment of benchmarks for comparative NOS evaluations, and the assessment of the quality and achievements of different competing NOS teaching methods (implicit, explicit, and reflective), etc.

In many countries such as Panama, whose curricula have never before included NOS contents, teaching NOS in school demands a special innovative effort, as many teachers lack training to teach these topics. Indeed, there is evidence that even when expert teachers are faced with teaching a new topic with which they are unfamiliar (as may be the case with NOS), they may be unable to transfer the expert behaviour that characterizes their teaching to this new, relatively uncomfortable, context. As a result, they may take unexpectedly incompetent approaches to this teaching that

would be more typical of a novice teacher, and may find it harder to include NOS topics in their practice than might have been expected (Sanders, Borko & Lockard, 1993). Hence, teachers primarily require appropriate teaching materials to help them design and implement NOS activities in their classes, but they also need to engage in explicit and reflexive analysis of the issues of NOS. Indeed, this should become the core of both initial and ongoing science teacher education programs in order to stimulate authentic NOS teaching in schools. Although the non-expert science teachers' obstacles to teach NOS stem mainly from diverse negative perceptions about NOS (Höttecke & Silva, 2011), the instrument and the standardized method presented herein should be of assistance to teachers in their educational evaluation tasks. However, the teachers' knowledge and analysis of sentence and item content also constitute a guide to fostering the development of explicit and reflexive analyses of issues concerning the NOS curriculum, as each sentence's category may orientate about its actual value and increase teachers' training on NOS (Acevedo, 2009; Hanuscin, Lee & Akerson, 2011).

Summing up, the illustration of the assessment of some conceptions on epistemology of science in a large sample contributes to the field of NOS research by developing a significant method that provides quantitative, quick, valid and informative results on tailored NOS topics.

Acknowledgments

This study corresponds to grant EDU2010-16553 funded by the national R+D+i 2010 programme of the Ministry of Science & Innovation (Spain) with the support of the Organization of Ibero-American States (OEI).

References

Abd-El-Khalick, F. (2012a). Examining the sources for our understandings about science: Enduring conflations and critical issues in research on nature of science in science education. *International Journal of Science Education, 34*(3), 353-374.

Abd-El-Khalick, F. (2012b). Nature of science in science education: Toward a coherent framework for synergistic research and development. In B. J. Fraser et al. (Eds.), Second International Handbook of Science Education (pp. 1041-1060). New York, NY: Springer-Verlag.

Abd-El-Khalick, F., & Lederman, N. G. (2000). The influence of history of science courses on students' view of nature of science. *Journal of Research in Science Teaching, 37*(10), 1057-1095.

Acevedo, J. A. (2009). Conocimiento didáctico del contenido para la enseñanza de la naturaleza de la ciencia (II): una perspectiva. *Revista Eureka sobre Enseñanza y Divulgación de las Ciencias, 6*(2), 164–189.

Aikenhead, G. S. (2006). *Science education for everyday life: Evidence-based practice*. New York, NY: Columbia University.

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.

Aikenhead, G. S., Fleming, R. G., & Ryan, A. G. (1987). High school graduates' beliefs about science-technology-society. I. Methods and issues in monitoring students' views. *Science Education, 71*, 145-161.

Aikenhead, G. S., Ryan, A. G., & Fleming, R. W. (1989). *Views on Science-Technology-Society* (form CDN.mc.5). Saskatoon, Canada: University of Saskatchewan. Retrieved from http://www.usask.ca/education/people/aikenhead/vosts.pdf.

Akerson, V., & Donnelly, L. A. (2010) Teaching nature of science to K-2 students: What understandings can they attain? *International Journal of Science Education, 32*, 97-124.

Allchin, D. (2011). Evaluating knowledge of the nature of (whole) science. *Science Education, 95*, 518–542.

Botton, C., & Brown, C. (1998). The reliability of some VOSTS items when used with preservice secondary science teachers in England. *Journal of Research in Science Teaching, 35*, 53-71.

Brunner, J., Summers, R., Myers, J. Y., & Abd-El-Khalick, F. (2016). Toward quantifying responses to the Views of Nature of Science Questionnaire: Empirically investigating qualitative coding. NARST 2016 Annual International Conference Abstracts (pp. 126-127). Retrieved from https://www.narst.org/annualconference/2016conference.cfm

Chen, S., Chang, W-H., Lieu, S.-C., Kao, H.-L., Huang, M.-T., & Lin, S.-F., (2013). Development of an empirically based questionnaire to investigate young students' ideas about nature of science. *Journal of Research in Science Teaching, 50*(4), 408–430.

Cobern, W. W., & Loving, C. C. (2002). Investigation of preservice elementary teachers' thinking about science. *Journal of Research in Science Teaching, 39*(10), 1016-1031.

Coll, R. K. (2012). Foreword. In M. S. Khine (Ed.), *Advances in nature of science research* (p. v). Dordrecht, The Netherlands: Springer.

Deng, F., Chen, D.-T., Tsai, C-C, & Chai, C.-S. (2011). Students' views of the nature of science: A critical review of research. *Science Education, 95*, 961–999.

Dogan, N., & Abd-El-Khalick, F. (2008). Turkish grade 10 students' and science teachers' conceptions of nature of science: A national study. *Journal of Research in Science Teaching, 45*(10), 1083–1112.

Eagly, A.H., & Chaiken, S. (1993). The psychology of attitudes. Orlando, FL: Harcourt Brace Jovanovich College Publishers.

Erduran, S., & Dagher, R. F. (2014). *Reconceptualizing the nature of science for science education*. Dordrecht, The Netherlands: Springer.

Eurydice (2011). *Science education in Europe: National policies, practices and research*. Retrieved from http://eacea.ec.europa.eu/education/eurydice

García-Carmona, A., Vázquez, A., & Manassero, M. A. (2011). Estado actual y perspectivas de la enseñanza de la naturaleza de la ciencia: una revisión de las creencias y obstáculos del profesorado. *Enseñanza de las Ciencias, 29*(3), 403-412.

García-Carmona, A., Vázquez, A., & Manassero, M. A. (2012). Comprensión de los estudiantes sobre naturaleza de la ciencia: análisis del estado actual de la cuestión y perspectivas. *Enseñanza de las Ciencias, 30*(1), 23-34.

Halloun, I., & Hestenes, D. (1998). Interpreting VASS dimensions and profiles for physics students. Science & Education, 7(6), 553-577.

Hanuscin, D. L., Lee, M. H., & Akerson, V. L. (2011). Elementary teachers' pedagogical content knowledge for teaching nature of science. *Science Education, 95*(1), 145–167.

Kind, P., & Barmby, P. (2011). Defending attitude scales. In I. M. Saleh & M. S. Khine (Eds.), *Attitude research in science education: Classic and contemporary measurements* (pp. 117-135). Charlotte, NC: Information Age Publishing.

Lederman, N. G. (2007). Nature of science: Past, present, and future. In S. K. Abell & N. G. Lederman (Eds.), *Handbook of research on science education* (pp. 831-879). Mahwah, NJ: Erlbaum.

Lederman, N. G., Wade, P. D., & Bell, R. L. (1998). Assessing understanding of the nature of science: A historical perspective. In W. F. McComas (Ed.), *The nature of science in science education: Rationales and strategies* (pp. 331-350). Dordrecht, The Netherlands: Kluwer.

Lederman, N.G., & O'Malley, M. (1990). Students' perceptions of tentativeness in science: Development, use, and sources of change. *Science Education, 74*, 225–239.

Lederman, N. G., Abd-El-Khalick, F., Bell, R. L., & Schwartz, R. (2002). Views of Nature of Science Questionnaire (VNOS): Toward valid and meaningful assessment of learners' conceptions of nature of science. *Journal of Research in Science Teaching, 39*, 497–521.

Liu, S.-Y., & Tsai, C.-C. (2008). Differences in the scientific epistemological views of undergraduate students. *International Journal of Science Education, 30*(8), 1055 – 1073.

Liu, X. (2012). Developing measurement instruments for science education research. In B.J. Fraser, K. Tobin, & C. J. McRobbie (Eds.), *Second international handbook of science education* (pp. 651-666), Dordrecht, The Netherlands: Springer.

Manassero, M. A., Vázquez, A., & Acevedo, J. A. (2003a). *Cuestionario de opiniones sobre ciencia, tecnologia i societat (COCTS)*. Princeton, NJ: Educational Testing Service.

Manassero, M. A., Vázquez, A., & Acevedo, J. A. (2003b). Views on science-technology-society questionnaire: Categories and applications. Paper presented at the 4th Conference of the *European Science Education Research Association (ESERA) on the Research and the Quality of Science Education*, Noordwijkerhout, The Netherlands.

Matthews, M. R. (2012). Changing the focus: From nature of science (NOS) to features of science (FOS). In M. S. Khine (Ed.), *Advances in nature of science research. Concepts and methodologies* (pp. 3-26). Dordrecht, The Netherlands: Springer.

Next Generation Science Standards [NGSS] (2013). *The next generation science standards: For states, by states*. Washington, DC: National Academies Press.

Osborne, J., Collins, S., Ratcliffe, M., Millar, R., & Duschl, R. (2003). What "ideas-about-science" should be taught in school science? A Delphi study of the expert community. *Journal of Research in Science Teaching, 40*(7), 692 – 720.

Rubba, P. A., Schoneweg-Bradford, C., & Harkness, W. L. (1996). A new scoring procedure for the Views on Science-Technology-Society instrument. *International Journal of Science Education, 18*(4), 387-400.

Ryan, A. G., & Aikenhead, G. S. (1992). Students' preconceptions about the epistemology of science. *Science Education, 76*, 559–580.

Schwartz, R. S., Lederman, N. G., & Abd-El-Khalick, F. (2012). A series of misrepresentations: A response to Allchin's whole approach to assessing nature of science understandings. *Science Education, 96*, 685–692.

Tedman, D. K., & Keeves, J. P. (2001). The development of scales to measure students' teachers' and scientists' views on STS. *International Education Journal, 2*, 20-48.

Vázquez, A., & Manassero, M. A. (1999). Response and scoring models for the 'Views on Science-Technology-Society' Instrument. *International Journal of Science Education, 21*(3), 231-247.

Vázquez, A., García-Carmona, A., Manassero, M. A., & Bennàssar, A. (2013). Spanish Secondary-School Science Teachers' Beliefs about Science-Technology-Society (STS) Issues. *Science & Education, 22*(5), 1191-1218.

Vázquez, A., García-Carmona, A., Manassero, M. A., & Bennàssar, A. (2014). Spanish students' conceptions about NOS and STS issues: A diagnostic study. *Eurasia Journal of Mathematics, Science & Technology Education, 10*(1), 33-45.

Vázquez, A., Manassero, M. A. & Acevedo, J. A. (2006). An Analysis of Complex Multiple Choice Science-Technology-Society Items: Methodological Development and Preliminary Results. *Science Education, 90*(4), 681-706.