International Journal of Science Education, 2015
http://dx.doi.org/10.1080/09500693.2015.1035357

Routledge
Taylor & Francis Group

Comparing Models of Nature of Science Dimensionality Based on the Next Generation Science Standards

George M. Harrison*, Kanesa Duncan Seraphin, Joanna Philippoff, Lisa M. Vallin and Paul R. Brandon
Curriculum Research & Development Group, College of Education, University of Hawai'i at Mānoa, Honolulu, HI, USA

Instruments measuring understanding of the nature of science (NOS) are required if educational institutions intend to use benchmarks or examine the effects of interventions targeting students' NOS development. Compared to other constructs, NOS understanding is complex, having been the subject of debate among scholars in both its substance and its dimensionality. This complexity invites challenges in defining what is to be measured. Drawing from the perspective that policy reform documents provide pragmatic consensus-based definitions of NOS, this study investigated how well the dimensionality described in the NOS component of the Next Generation Science Standards (NGSS) framework matched the empirical structure of data collected from a set of secondary-school students' responses to an NOS instrument comprising multiple-choice and Likert-scale items. Using multidimensional item response modeling to compare structures of NOS dimensionality, we found that treating NOS as comprising multiple dimensions—as defined by the themes in the NGSS NOS framework—resulted in a better fitting model than when treating NOS as a single dimension. The multidimensional model also had fewer poorly functioning items and revealed NOS profiles that otherwise would have been masked in a model treating NOS as a single dimension. These results provide support for the NOS NGSS framework and contribute to the ongoing discussion about the dimensionality of NOS.

Keywords: *Nature of science; Next Generation Science Standards; Item response modeling; Science assessment*

*Corresponding author. Curriculum Research & Development Group, College of Education, University of Hawai'i at Mānoa, 1776 University Avenue, Honolulu, HI 96822, USA. Email: georgeha@hawaii.edu

Introduction

Improving middle- and high-school students' understanding of the nature of science (NOS) has been an ongoing endeavor for more than 50 years (as evidenced by works cited in Lederman, 2007). The focus on NOS remains strong today, as demonstrated by the attention given to NOS in the USA's recently released Next Generation Science Standards (NGSS) framework (NGSS Lead States, 2013) and by the sheer number of publications on the subject in science education journals in the last decade (e.g. since 2000, more than 20 articles with 'nature of science' in the title have appeared in *Science Education*, *Journal of Research in Science Teaching*, and *International Journal of Science Education*). With this emphasis comes a need to gather data to inform instructional decisions in NOS, to research the nature of NOS and its relationship with other aspects of science, and to measure the effects of interventions intended to improve students' NOS understandings. Instruments measuring NOS understanding, therefore, have been and continue to be developed, as illustrated by the list of assessments and surveys in published reviews (Deng, Chen, Tsai, & Chai, 2011; Lederman, 2007).

Definitions of NOS

As with any instrument, validity arguments for NOS instruments depend on their definitions of the construct being measured. There have been disagreements among science education scholars in how NOS is to be defined in research and practice (Abd-El-Khalick, 2005; Allchin, 2011; Alters, 1997; Deng et al., 2011; Grandy & Duschl, 2007; Lederman, 2007; Schwartz, Lederman, & Abd-El-Khalick, 2012). Some scholars stress that NOS research should focus primarily on epistemological views of NOS knowledge (Abd-El-Khalick, 2012a; Lederman, 2007; Lederman, Abd-El-Khalick, Bell, & Schwartz, 2002). Others have the perspective that NOS research should address performance resulting from NOS understanding (Allchin, 2011). Still others argue that individuals' NOS understanding should not be measured on a continuum of epistemological values (such as from empiricist to constructivist views), but should instead be examined as a degree of appropriateness for a specific task in a scientific inquiry (Deng et al., 2011).

In the eyes of some researchers (e.g. Alters, 1997), these contrasting, and sometimes contentious, perspectives on what constitutes NOS present a burden to NOS instrument developers. In the eyes of others (e.g. Abd-El-Khalick, 2012a, 2012b), efforts to benchmark—and therefore to measure—NOS are best informed by NOS definitions provided in policy reform documents. In Abd-El-Khalick's (2012b) proposal for a synergistic framework of NOS, he recognizes that debates concerning nuanced definitions of NOS will likely carry on, and that this is consistent with the very notion that scientific knowledge is tentative. But, he advocates a pragmatic approach to defining NOS so that assessing NOS understanding is meaningful and productive. Abd-El-Khalick suggests that drawing from policy documents serves to identify central components of NOS that scholars have reached consensus on. In the USA,

a recently released policy document that researchers and practitioners can draw from is the NOS framework delineated in Appendix H of the NGSS (NGSS Lead States, 2013). It is this document's description of NOS that informed our study.

The Multidimensional Nature of NOS Understanding

In the literature on NOS research, some studies treat NOS as a unidimensional construct (e.g. Bell & Linn, 2000; Wenning, 2006), whereas others treat it as multidimensional (e.g. Chen, 2006; Lederman et al., 2002). In educational measurement, a construct is treated as unidimensional if the intent is to derive a score that serves as a measure of the student's knowledge in a single domain. This may be the case even if the domain is a composite of multiple subdomains. For example, an end-of-course exam in biology may include multiple components of biology knowledge, such as genetics, metabolic processes, and ecology. A single score would be interpreted as an estimate of the student's general biology knowledge across the three subdomains. If the intent is to measure subdomains of knowledge, the construct is typically assumed to be multidimensional, such that each subdomain is its own dimension (on a subscale) with its own score. For example, if the intent of the biology test were to identify content areas that the students need the most help on, a score on each subdomain would be informative. The validity of using subscales to measure students' understanding on a construct is strengthened when there is evidence that the pattern of item interrelationships, as observed in empirical response data, is consistent with the dimensionality proposed by the construct's theoretical framework (AERA, APA, & NCME, 2014).

The multidimensional nature of NOS understanding has been an ongoing topic of discussion in the science education literature. Lederman (2007) and his colleagues (Lederman, Wade, & Bell, 1998), for example, evaluated 24 NOS instruments' merit based on whether they included multiple subscales. Deng et al. (2011) reviewed several NOS studies, categorizing them based on whether they treated NOS as unidimensional, as multidimensional, or as dependent on the science task at hand. Blalock et al. (2008) evaluated several NOS instruments based on whether dimensionality was addressed and whether the study compared empirical data with the study's intended, theoretically specified, dimensionality. Authors have described the dimensions of NOS as *aspects* (Abd-El-Khalick, 2005, 2012a; Chen, 2006; Lederman, 2007), *factors* (Chen et al., 2013), *themes* (NGSS Lead States, 2013), and *dimensions* (Neumann, Neumann, & Nehm, 2011).

Previous studies comparing groups of subjects in their NOS understanding have, in their analyses, specified NOS as comprising multiple dimensions. For example, Bayir, Cakici, and Ertas (2013) examined seven NOS aspects, similar to those described in Lederman (2007), and compared natural science and social science professors' views on each aspect. Tsai (1999) used five dimensions of NOS, each on an empiricist to constructivist continuum, to compare interview responses of treatment- and comparison-group students in an intervention in Grade 10 science classrooms. Measures of students' understandings on separate NOS dimensions are valuable for gaining an

understanding of the effects of interventions. If students in treatment groups, for instance, improve in their NOS understanding in particular dimensions but not in others, when compared to control-group students, the effects of the intervention can be better understood. The credibility of intervention studies examining effects on NOS dimensions, particularly those informed by policy documents, would be bolstered by empirical evidence supporting the presence of distinct, yet still related, NOS dimensions.

Compared to the amount of scholarly work done to delineate the sub-components of NOS, relatively few studies have examined how well dimensionality specifications hold up when compared with empirical data from students' responses (exceptions include Chen et al., 2013; Huang, Tsai, & Chang, 2005; Neumann et al., 2011). In other words, there is a need for more work that looks at how well the dimensionality on a given assessment reflects the observed pattern of students' responses on that assessment. Evaluators of the psychometric quality of science attitude instruments (Blalock et al., 2008) have recommended NOS instrument developers explicate their theory of the dimensionality of NOS and use factor-analysis methods to compare empirical response data with their theory.

In a study comparing empirical data with theoretical NOS dimensions, Huang, Tsai, and Chang (2005) used orthogonal factor analysis with Likert-scale data to argue for a three-dimensional NOS model. Their use of an orthogonal model, however, might not have been appropriate because this implies the NOS dimensions are completely unrelated to each other. In a similar study (Chen et al., 2013), Chen and colleagues conducted a confirmatory factor analysis of Likert-scale data. They argued, based on the good fit between the empirical data and the specified model, for a seven-dimension NOS structure. A limitation of both of these studies is that they did not conduct model comparisons to identify whether their proposed multidimensional models fit the data better than alternative models, such as one with NOS specified as a single dimension. Furthermore, a problem with using these conventional factor-analysis methods to examine dimensionality is that the statistical assumption of linearity is seldom met (Bond & Fox, 2007; Edwards, Wirth, Houts, & Xi, 2012).[1]

Fortunately, investigations into dimensionality are not limited to conventional types of factor analysis. Item response models, including those informed by item response theory (IRT) and Rasch measurement, can be utilized to examine the structure of empirical response data without violating the assumption of linearity (Edwards et al., 2012; Kamata & Bauer, 2008; Muthén & Muthén, 2010). One type of item–response model that can inform NOS researchers is the multidimensional random coefficients multinomial logit (MRCML) model (Adams, Wilson, & Wang, 1997; Adams & Wu, 2007), which has been described as an extension of the Rasch family of item–response models (Adams & Wu, 2007; Briggs & Wilson, 2003). In investigations into NOS dimensionality, the MRCML model has been recommended (Chen, 2006) and put to use (Neumann et al., 2011).

Neumann et al. (2011) conducted an empirical investigation into NOS dimensionality based on data from a Likert-scale survey (Lombrozo, Thanukos, & Weisberg, 2008) administered to 214 undergraduate university students. They used MRCML

modeling to compare three models: a one-dimension model that combined inquiry and NOS items in a single scale, a two-dimension model that specified inquiry and NOS as separate but related dimensions (thus, NOS was treated as being unidimensional), and a 13-dimension model which specified the one inquiry dimension and 12 NOS dimensions as they were defined in Lombrozo, Thanukos, and Weisberg (2008). They found that the 13-dimension model did not converge (i.e. stable subscale scores could not be computed), and that the two-dimension model (which specified NOS as unidimensional) fit better than the one-dimension model (which treated NOS and inquiry as a single dimension). In other words, Neumann et al. (2011) found support that NOS was distinct from inquiry, but they could not determine, based on the response data, whether NOS was more accurately described as comprising multiple dimensions or as a single, unidimensional, scale. The lack of convergence of the model that treated NOS as 12 dimensions was probably due to the small dataset ($N = 214$). Their inconclusive results on the dimensionality of NOS warrant further research with a larger dataset. Results from such a study will inform stakeholders wishing to determine whether to use subscale scores or a single NOS score.

Purpose

Following the call for studies examining NOS dimensionality, our purpose is to investigate how well the NOS sub-components described in the NGSS NOS framework (NGSS Lead States, 2013, Appendix H) match the empirical structure of the data collected from an NOS questionnaire given to secondary-school students. The NGSS NOS framework consists of two overarching components: NOS understandings associated with science practices, and NOS understandings associated with crosscutting concepts in science. Each of these broad components is made up of four NOS themes, or sub-components, of NOS understanding. Within the practices component, the sub-components are (a) scientific investigations use a variety of methods, (b) scientific knowledge is based on empirical evidence, (c) scientific knowledge is open to revision in light of new evidence, and (d) science models, laws, mechanisms, and theories explain natural phenomena. Within the crosscutting concepts component, the sub-components are (a) science is a way of knowing, (b) scientific knowledge assumes an order and consistency in natural systems, (c) science is a human endeavor, and (d) science addresses questions about the natural and material world.

There are, therefore, at least three possible measurement models to consider with instruments addressing the NGSS NOS framework: a model specifying NOS as a single dimension, a model specifying NOS as the two overarching dimensions, and a model specifying NOS as sub-dimensions corresponding to the NOS themes. We address these models in our study, comparing them to identify which best fits the patterns in a set of empirical data. This strategy to comparing measurement models that differ in the way they divide up the construct of NOS is consistent with measurement theory (Wilson, 2004) in the sense that the competing models represent different mappings of NOS understanding. Results will contribute to the literature on NOS

dimensionality and will add to discussions about how well the NGSS NOS theme framework holds up with empirical data.

Methods

The instrument used to collect the response data was the Student Science Questionnaire (SSQ). It was developed to measure secondary-school students' NOS understanding and science content knowledge after their teachers completed a professional-development (PD) series in science education. The final draft of the SSQ consisted of 45 items, 28 of which measured NOS and 17 of which measured aquatic science. Of the NOS items, 18 were multiple-choice (MC) and 10 were Likert-scale in format. The content items were intended to measure the science content provided in the PD. In the instrument's MC section, which had NOS and content items interspersed, the instructions asked students to select the best answer. On the Likert-scale section (adapted from Ayala, 2005), students were asked to rate how true they thought each statement was on a six-point scale with the boundaries labeled as *not true at all* and *very true*.[2] Half of the Likert-scale statements were worded with negative valence and were reverse-coded before quantifying students' responses. Examples of the MC and Likert-scale items are presented in Figure 1. The appendix lists the instrument development procedures and provides the NOS items.

Empirical Data for the Dimension Analysis

The SSQ was administered to a total of 1,437 middle- and high-school students in intact classes in the US state of Hawai'i. Of these, 472 students (33%) completed an earlier version of the SSQ that consisted of 19 NOS items and none of the content items; thus, 23% of the total dataset comprised empty item–response cells, by design. This type of missing data is not a problem in item–response models because students' performance is estimated based on the items they were presented (de Ayala, 2009). Missing responses due to students' skipping items made up less than 1% of the dataset.

NOS Theme Assignment Procedures

To identify which, if any, of the eight NGSS NOS themes (Appendix H of the NGSS framework) each NOS item addressed, we had five panelists participate in a Delphi-method item–judgment procedure. The panelists were three science education experts (two associate professors in science education and one doctoral student with three years' experience teaching inquiry-based PD courses) and two educational assessment experts (one assistant professor and one full professor). There were two stages of this procedure.

In the first stage, the five participants assigned each NOS item to one or more NGSS NOS themes. The panelists were told that some items might not align with any theme and that they could withhold theme assignment on any item, though none of the items

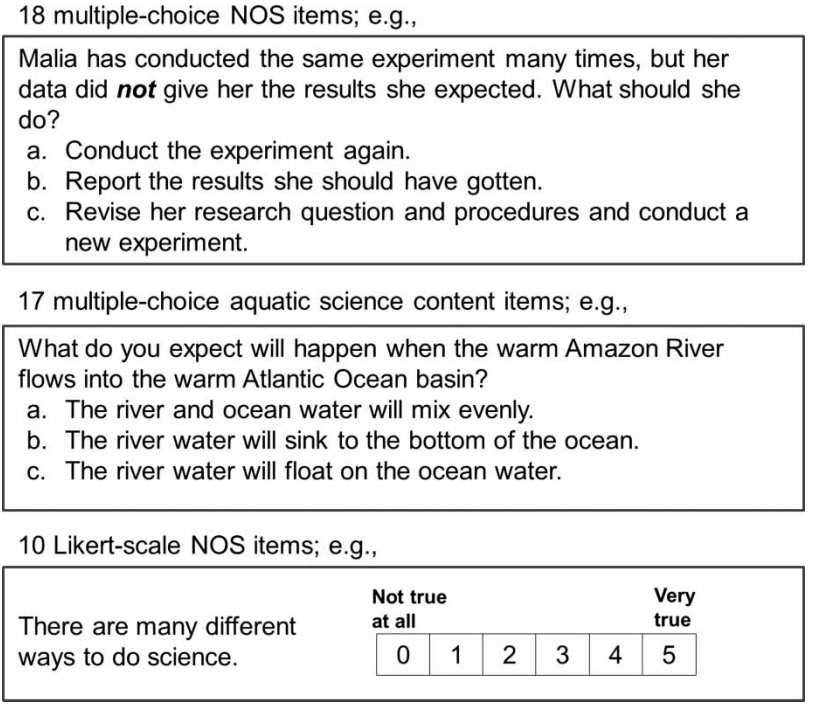

Figure 1. Example items from the Student Science Questionnaire

was skipped by any panelist. Panelists were also asked to provide a written account of their rationale for each of their theme choices. Every panelist assigned all 28 NOS items to at least one theme, but there were discrepancies among panelists on 11 items (9 MC and 2 Likert-scale) in this stage. These items were marked for later discussion among panelists.

In the second stage, the three science education experts discussed the marked discrepancies after they (a) reviewed their fellow panelists' ratings and rationales and (b) re-examined the items and the NGSS NOS framework. The panelists reached agreement on the 11 earlier discrepant items. Of these, 10 were judged to be addressing two NGSS NOS themes rather than one. Table 1 displays the final item–theme assignments along with their broader NGSS component categories. Among the eight NGSS NOS themes, two were represented by fewer than three items. Because there were so few items in these two themes, they were eliminated from the analysis, which in turn resulted in a set of 43 items for the subsequent model-comparison procedure (items MC16 and Lik9 were excluded).

Model-Comparison Procedures

To examine the match between the NGSS NOS framework and the empirical data, we compared five models that differed in the way they treated NOS dimensionality. This entailed identifying the best fitting model, estimating the reliability of the subscales

Table 1. Results of the NOS theme assignment procedure

Construct element[a]	NGSS NOS theme[b]	Overarching NGSS component	Item numbers[c]	*N* of items
Scientific investigations use a variety of methods	1	Practices	*MC7*, MC11, *MC17*, *MC24*, Lik3, *Lik6*, Lik7	7
Scientific knowledge is based on empirical evidence	2	Practices	*MC2*, *MC5*, MC19, *MC20*, *MC22*, *MC23*	6
Scientific knowledge is open to revision in light of new evidence	3	Practices	MC3, MC10, MC14, Lik5, Lik8, *Lik10*	6
Science models, laws, mechanisms, and theories explain natural phenomena	4	Practices	MC16	1
Science is a way of knowing	5	Crosscutting concepts	*MC7*, MC13, *MC17*, *MC22*, *MC23*	5
Scientific knowledge assumes an order and consistency in natural systems	6	Crosscutting concepts	MC1, *MC5*, *MC20*, *MC24*, *Lik6*	5
Science is a human endeavor	7	Crosscutting concepts	*MC2*, MC6, MC9, Lik1, Lik2, Lik4	6
Science addresses questions about the natural and material world	8	Crosscutting concepts	Lik9, *Lik10*	2
Aquatic science content		—	MC4, MC8, MC12, MC15, MC18, MC21, MC25–MC35	17

[a]The first eight construct elements are the NGSS NOS themes; the last construct element represents the science content specific to the PD in which this instrument was used.

[b]These are numbered according to their order of appearance in the NGSS NOS framework.

[c]Item numbers include the prefix MC or Lik to indicate whether they were multiple-choice or Likert-scale in format, respectively. Items specified to load on two NGSS NOS themes are in italics and are listed in each of the two respective construct elements.

defined by the dimensions in each model, examining the item functioning in two competing models, and examining the impact of a multidimensional NOS model on students' NOS profiles.

Model 1 treated all 43 items as being on a single dimension, regardless of whether they measured content or NOS. This model is not desirable because it would fail to distinguish content knowledge from NOS understanding, placing the credibility of the instrument at stake.

Model 2 treated NOS as a single dimension and content as a separate, yet related, dimension. Evidence that this model fits best would support using a single NOS score, across all NGSS NOS themes, to investigate students' NOS understanding.

Model 3 treated NOS as two dimensions and content as its own dimension. In this model, the two NOS dimensions represented the two overarching components in the NGSS NOS framework; that is, processes and crosscutting concepts. Evidence that this model fits best would support using subscales corresponding to these two overarching components when investigating NOS understanding.

Model 4 was a seven-dimension model, with six dimensions representing their corresponding NGSS NOS themes and the remaining dimension measuring content. Evidence that Model 4 fits best would support investigating NOS understanding as multiple subscales that correspond to the themes in the NGSS NOS framework.

Models 3 and 4 included within-item multidimensionality. That is, some items were treated as measuring more than one dimension. For example, Item MC7 in Model 3 was modeled to measure both NOS Theme 1 and NOS Theme 5. There were 10 NOS items with this type of complexity, all of which corresponded to the item–theme assignment results reported in the preceding section (these items are presented in italics in Table 1).

Model 5 was of two dimensions: a Likert-scale dimension, with the 10 Likert-scale items, and a multiple-choice dimension, with the remaining 33 MC items. The purpose of including this model was to investigate whether an item-format effect was stronger than the best of the first four models.

With each of the models under comparison, we used an MRCML model framework (Adams et al., 1997). In the same way that a unidimensional Rasch measurement model operates (Wright & Stone, 1979), MRCML models simultaneously include the respondent's ability and the item's difficulty when estimating the probability that a particular respondent will answer an item correctly. MRCML modeling has been used in science education research to compare models differing in how a construct is partitioned into dimensions (e.g. Bao, Gotwals, Songer, & Mislevy, 2006; Briggs & Wilson, 2003; Neumann et al., 2011; Wei, Liu, & Jia, 2014). Because the students' SSQ responses were on MC and Likert-scale items, we used a partial-credit model (Masters, 1982), which handles both types of data in a single run. With Likert-scale response data, the probability of a respondent's tendency to select one higher number on an item's scale is estimated (for examples of the partial-credit model in science education research, see Bond & Fox, 2007; Boone, Townsend, & Staver, 2011; Neumann et al., 2011; Sjaastad, 2013). We used ConQuest 3 software (Wu, Adams, Wilson, & Haldane, 2007) to perform all estimation procedures. With each model under comparison, students' subscale scores on each dimension were estimated.

To compare the models, we used the Akaike information criterion (AIC), which is an index of model fit that adjusts for the complexity of a model. This criterion has been used in other MRCML model-comparison studies (e.g. Bao et al., 2006; Briggs & Wilson, 2003; Neumann et al., 2011). Each model's AIC estimate was calculated as the deviance (the -2 log-likelihood) of the model plus two times the degrees of freedom of the model. Essentially, this criterion is an estimate of how poorly the pattern of students' responses corresponds with the dimensionality structure imposed by the model. If one model has a lower AIC than another, it fits the data better.

To compare models in their precision of the subscales, we examined the expected a posterior (EAP) reliability (Adams, 2005). Unlike classical-test-theory types of reliability, such as Cronbach's alpha, EAP reliability permits estimation of precision when there are missing item-level responses among observations in the data.

Additionally, we examined NOS item functioning in Models 2 and 4 to see whether the instrument suffered a loss in quality when treating NOS as multiple dimensions rather than as a single dimension. This included an analysis of item discrimination indexes, item fit indexes, and person–item patterns as represented in person–item maps. The item discrimination index indicates how strongly the item discriminates among respondents at different levels of NOS understanding, as defined by the items on the respective dimension (Kelley, Ebel, & Linacre, 2004). For this, we examined the correlations between the item responses and the subscale scores (i.e. the item–measure correlations). A negative item–measure correlation would indicate the item is functioning differently than the set of items specified to measure that dimension. Item fit indexes indicate how well an item fits with its corresponding dimension. To compare the models' item fit indexes, we examined the weighted and unweighted fit indexes reported by ConQuest. The unweighted index (sometimes referred to as *outfit*, Smith, 2004) is the traditional mean-square-error estimate of item misfit, and is sensitive to outliers. The weighted index (sometimes referred to as *infit*) is adjusted by the item's variance, and is therefore less prone to bias (see Smith, 2004). Bond and Fox (2007) present several criteria for evaluating whether an item is misfitting; however, they point out that with a large sample size, it is probable that all items will appear to fit well (also see Linacre, 2003). To account for this, we set the criterion for unstandardized misfit at $0.90 >$ fit > 1.10 and additionally examined the standardized fit of each item, with the criterion set at $-1.96 >$ standardized fit $> +1.96$ (to correspond with a 95% confidence interval in a unit-normal distribution). Standardized fit is also sensitive to sample size, but in the opposite direction to how sample size affects the unstandardized fit indexes (Bond & Fox, 2007); even a small degree of misfit can yield a value greater in magnitude than ± 1.96. Given this sensitivity, we expected more items to be flagged as misfitting using the standardized index. The person–item maps (also called *Wright maps*; Wilson, 2011) provide an indication of how well the items in each dimension represent the various levels of student understanding on the specific dimension (Linacre, 2004). Ideally, the items will be dispersed across the range of person levels, and there will be no large gaps between items.

To examine whether a multidimensional NOS model would reveal different profiles of students' NOS understanding than would a composite NOS score, we investigated the degree to which the NOS dimensions in Model 4 differed from each other in their rank-ordering of the respondents. Following the MRCML example in Briggs and Wilson (2003; also see Linacre, 2004), we examined the correlations among the dimensions on the six NOS subscales and calculated, for each student, the standard deviation of their standardized subscale scores (further explained in the online supplemental research materials). This information is valuable in determining whether there is any practical value in treating the NOS themes as separate

dimensions. If there are high positive correlations between the dimensions (or small deviations among the subscale scores), there would be little benefit to providing subscale scores on these dimensions.

Results

The AIC fit statistics of each model are shown in Table 2. The best fitting model was Model 4, which treated NOS as multidimensional according to the NGSS NOS themes. Its improvement over Model 2, which treated NOS as being a single dimension, indicated that, for these data, the NGSS NOS themes described the students' patterns of responses better than if a composite NOS score were used. Models 1 and 3 fit poorly. The poor fit of Model 1 indicated that the empirical data were a poor match with a model that treated content and NOS as being a single dimension. Similarly, the poor fit of Model 3 indicated that partitioning NOS understanding into practices and crosscutting concepts did not explain the patterns of students' responses. Model 5 fit moderately well, indicating that the pattern of students' scores depended, to some extent, on the format of the item.

The EAP reliability estimates of the dimensions in each of the models are reported in Table 3. The reliabilities of the NOS subscales in the best fitting model (Model 4) were generally low compared to their counterparts in the other models. Also reflected in Table 3, however, is that Model 4 had many fewer items per NOS dimension than the other models.

Model 4 performed moderately better than Model 2 on the five item-quality indexes listed in Table 4, indicating that item discrimination and fit improved. The person–item maps (provided in the online supplemental research materials) did not reveal drastic changes in the construct representation, however. For both Models 2 and 4, the NOS dimensions are underrepresented at the high end of the scale; that is, the

Table 2. Fit statistics of the models under comparison

Model number	Model	Deviance	*N* of free parameters	AIC
1	One-dimension model (NOS and content as a single dimension)	77,392	71	77,534
2	Two-dimension model (NOS and content as separate dimensions)	77,326	73	77,472
3	Three-dimension model (NOS practices, NOS crosscutting concepts, and content as separate dimensions)	77,422	76	77,574
4	Seven-dimension model (six NGSS NOS themes, and content as separate dimensions)	77,099	98	77,295
5	Two-dimension item-format model (MC items and Likert-scale items as separate dimensions)	77,164	73	77,310

Notes: A lower AIC index suggests a better fitting model. AIC = Akaike information criterion.

Table 3. EAP reliability estimates of the dimensions within each model

Model	Dimension	*N* of items	EAP reliability
1	Combined NOS and content	43	.77
2	NOS	26	.78
	Content	17	.53
3	NOS practices	19	.69
	NOS crosscutting concepts	17	.71
	Content	17	.52
4	NOS Theme 1: Scientific investigations use a variety of methods	7	.58
	NOS Theme 2: Scientific knowledge is based on empirical evidence	6	.57
	NOS Theme 3: Scientific knowledge is open to revision in light of new evidence	6	.71
	NOS Theme 5: Science is a way of knowing	5	.50
	NOS Theme 6: Scientific knowledge assumes an order and consistency in natural systems	5	.53
	NOS Theme 7: Science is a human endeavor	6	.70
	Content	17	.58
5	Multiple-choice format	34	.72
	Likert-scale format	9	.72

Note: EAP reliability = expected a posterior (EAP) reliability (Adams, 2005).

instrument is not functioning well with students who have a high degree of NOS understanding. In both models, there were also regions in the scale that were underrepresented by the items, though the NOS items in Model 4, when examined across dimensions, appear to be slightly more evenly distributed.

The rank-ordering of students' estimated subscale scores in Model 4 showed that the multidimensional model revealed information about students' NOS understanding that would have been unavailable in a unidimensional NOS model. These

Table 4. Comparison of Models 2 and 4 in their item-quality indexes

	N of items	
Item-quality index	Model 2	Model 4
Weak item–measure correlation (<.30)	10 (38%)	8 (31%)
Unweighted misfit ($0.90 > \text{MnSq}_{\text{outfit}} > 1.10$)	9 (35%)	6 (23%)
Standardized unweighted misfit ($-1.96 > \text{ZStd}_{\text{outfit}} > +1.96$)	12 (46%)	10 (38%)
Weighted misfit ($0.90 > \text{MnSq}_{\text{infit}} > 1.10$)	5 (19%)	1 (4%)
Standardized weighted misfit ($-1.96 > \text{ZStd}_{\text{infit}} > +1.96$)	10 (38%)	8 (31%)

Notes: There was a total of 26 NOS items. Model 2 treated NOS as a single dimension; Model 4 treated NOS as comprising six NGSS NOS themes. The criteria for poor item quality are in brackets. $\text{MnSq}_{\text{outfit}}$ = unweighted fit index; $\text{ZStd}_{\text{outfit}}$ = standardized unweighted fit index; $\text{MnSq}_{\text{infit}}$ = weighted fit index; $\text{ZStd}_{\text{infit}}$ = standardized weighted fit index.

Table 5. Correlations of the dimensions in Model 4

Dimension	Theme 1	Theme 2	Theme 3	Theme 5	Theme 6	Theme 7
Theme 1	—					
Theme 2	.59	—				
Theme 3	.76	.65	—			
Theme 5	.18	.60	.53	—		
Theme 6	.44	.62	.70	.61	—	
Theme 7	.65	.66	.85	.54	.68	—
Content	.36	.67	.64	.78	.64	.67

Note: All correlations were estimated in ConQuest.

results should be prefaced, however, with the caution that students' subscale scores are estimates and included measurement error. Of the 15 correlations among NOS themes (Table 5), only two (Themes 1 & 3 and Themes 3 & 7) were above .70, suggesting the NOS sub-components are related but distinct. Figure 2 provides an example of how students' subscale scores on two dimensions, in this case Themes 1 and 2, differed from each other. For a large portion of the students, their scores on one subscale differed by over one standard-deviation unit from the other subscale score. This deviation is also reflected in the standard deviations of the six standardized NOS subscales for each student: The average standard deviation was around 0.50 standardized units ($M = 0.55$, median $= 0.52$); that is, on average, students' subscale scores on the six themes differed by half a standard-deviation unit.

Discussion

The results illustrate the multidimensional nature of NOS and contribute to arguments made by proponents of NOS subscale scores (Blalock et al., 2008; Deng et al., 2011; Lederman, 2007). More specifically, the findings provide support for treating NOS as comprising the NGSS NOS themes: Model 4 fit the empirical data

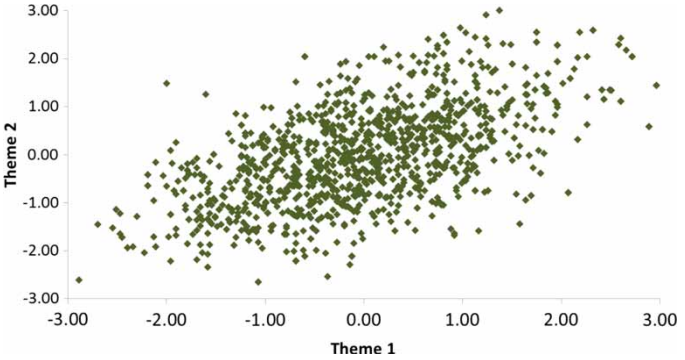

Figure 2. Plot of students' standardized subscale scores on Themes 1 and 2 based on Model 4

better than the other models, had reasonable reliability (given the small number of items per dimension), showed improved item functioning, and yielded subscale scores that provided information on students' NOS theme profiles. For studies examining students' NOS understanding, these subscale scores can provide greater fidelity than a global NOS score would.

The model-comparison results also support the validity of the instrument's use. Because Model 1 was poorly fitting, there is evidence in favor of the instrument developers' intention to measure NOS and content as distinct dimensions; that is, the distinction between NOS understanding and content knowledge appears to hold with the empirical data. This pattern is similar to the one reported in Neumann et al. (2011), though they examined inquiry rather than content as a distinct dimension.

The relative poor fit of Model 3, which treated NOS as two dimensions (practices and crosscutting concepts), is somewhat surprising given the emphasis on practices and crosscutting concepts as overarching dimensions in the NGSS framework. The Pearson correlation between these two dimensions in Model 3 was .72, suggesting that the two dimensions are moderately distinct. More research is needed to determine whether this partitioning of NOS is warranted.

The model defined by the item-formats (Model 5) was only slightly worse than the model treating the NOS themes as dimensions (Model 4). One can interpret the strength of Model 5 as evidence of a method effect; that is, the instrument was doing a good job of distinguishing among students based on their patterns of responses to item types. Another interpretation, which requires further investigation, is that this pattern reflects two separate aspects of NOS understanding: The Likert-scale items solicit self-reports of understanding of NOS using decontextualized statements, whereas the MC items solicit NOS understanding in the context of a scientific inquiry or other scenario. This can be illustrated with one pair of items addressing the same theme: Lik7, a Likert-scale item in Theme 1, asked for degree of endorsement with the statement, '*There are many different ways to do science*'; MC11, an MC item on the same theme, asked the student:

> Imagine that you are reading about a scientific study. You need to decide how well the study was conducted. Select the best question to ask about the study. (a) Did the study's methods include an experiment? (b) Were the study's methods appropriate for the study? (c) Was the study conducted by well-known scientists?

In this example, the multiple-choice question prompts thinking about NOS in a specific scenario whereas the Likert-scale question asks about science in general. This distinction between NOS understanding in a given task versus understanding in a decontextualized self-report is consistent with concerns raised in Deng et al. (2011). More research is needed to determine whether these two different aspects of NOS are responsible for what appears in this study to be a method effect.

As with any study, there are limitations. We recognize that because the NGSS NOS framework was being published at the same time we were completing a final draft of the instrument, there is room for improving the content aspect of our validity argument. We should point out, however, that the item developers drew heavily from

the National Research Council's *Framework for K–12 Science Education* (2012), which informed the development of the NGSS. Still, it is important to note that the best fitting model had considerable item misfit and gaps in its representation of the NOS themes, which indicates a need for further item development. An additional limitation may be that although the MC items were randomly ordered, the instrument was administered as a single form, which might have invited an ordering effect. One concern identified in the development, though, was the need to place the Likert-scale items at the end of the instrument to avoid prompting students to think about NOS in the general context of science prior to thinking about NOS in the specific contexts represented in the MC items. So, to some extent, this ordering effect was inevitable. Ideally, the MC items would have been presented in a random order, as is often done in computer-based assessment.

This study contributes to ongoing research into how NOS dimensionality should be treated, and it provides evidence supporting the NGSS NOS framework. As with any scientific inquiry, these results on their own are tentative, particularly given the room for improvement in item functioning. It is our hope that studies continue to examine the structure of NOS understanding, as such research would contribute to better assessments and more informed decisions based on those assessments.

Acknowledgments

The opinions expressed are those of the authors and do not represent views of the Institute of Education Sciences, the US Department of Education, or NOAA. This research was approved by the University of Hawaii (UH) committee on Human Subjects CHS # 15657. The authors thank their colleagues at the UH Curriculum Research & Development Group, UH Sea Grant, the University of Western Oregon, and the University of Kentucky for their intellectual and evaluative contributions to the professional development and NOS studies, including Dr Erin Baumgartner, Dr Thomas Guskey, Dr Frank Pottenger, Dr Lauren Kaupp, and Brian Lawton.

Disclosure statement

No potential conflict of interest was reported by the authors.

Funding

The research reported here was supported by the Institute of Education Sciences, US Department of Education [grant number R305A100091] and by National Oceanic and Atmospheric Administration (NOAA) grants to the University of Hawai'i (UH) at Mānoa.

Supplemental research materials

Supplemental research materials for this article can be accessed at doi:10.1080/09500693.2015.1035357.

Notes

1. Conventional factor-analysis methods assume there is a linear relationship between the factor and each observed variable. Typically, test items are dichotomous (scored as correct or incorrect) or ordinal (scored as partial credit or as steps in a Likert scale), which results in a non-linear relationship between each item and the factor and, therefore, a misspecification of the model in these traditional methods (Bond & Fox, 2007; Edwards et al., 2012). Item-response models, such as the MRCML model, are essentially non-linear factor-analysis models that can handle ordinal and dichotomous item-response data (Edwards et al., 2012).
2. Although the Likert-scale items were taken from an earlier instrument that had been examined for validity (Ayala, 2005), a limitation of these items is that only the boundaries of the scales were labeled.

ORCID


George M. Harrison http://orcid.org/0000-0002-6011-0063

References

Abd-El-Khalick, F. (2005). Developing deeper understandings of nature of science: The impact of a philosophy of science course on preservice science teachers' views and instructional planning. *International Journal of Science Education*, *27*, 15–42. doi:10.1080/09500690410001673810

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*, 353–374. doi:10.1080/09500693.2011.629013

Abd-El-Khalick, F. (2012b). Nature of science in science education: Toward a coherent framework for synergistic research and development. In B. J. Fraser, K. Tobin, & C. J. McRobbie (Eds.), *Second international handbook of science education* (Vol. 2, pp. 1041–1060). Dordrecht: Springer.

Adams, R. J. (2005). Reliability as a measurement design effect. *Studies in Educational Evaluation*, *31*, 162–172. doi:10.1016/j.stueduc.2005.05.008

Adams, R. J., Wilson, M., & Wang, W. C. (1997). The multidimensional random coefficients multinomial logit model. *Applied Psychological Measurement*, *21*, 1–23. doi:10.1177/0146621697211001

Adams, R. J., & Wu, M. L. (2007). The mixed-coefficients multinomial logit model: A generalized form of the Rasch model. In C. H. Carstensen (Ed.), *Multivariate and mixture distribution Rasch models* (pp. 57–75). New York, NY: Springer.

Allchin, D. (2011). Evaluating knowledge of the nature of (whole) science. *Science Education*, *95*, 518–542. doi:10.1002/sce.20432

Alters, B. J. (1997). Whose nature of science? *Journal of Research in Science Teaching*, *34*, 39–55. doi:10.1002/(SICI)1098-2736(199701)34:1<39::AID-TEA4>3.0.CO;2-P

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.

Ayala, C. C. (2005). *Development and validation of an inquiry science student achievement*. Paper presented at the annual meeting of the American Evaluation Association, Toronto, Canada.

de Ayala, R. J. (2009). *The theory and practice of item response theory*. New York, NY: Guilford Press.

Bao, H., Gotwals, A. W., Songer, N. B., & Mislevy, R. J. (2006). Using structured item response theory models to analyze content and inquiry reasoning skills in BioKIDS. In X. Liu & W. J. Boone (Eds.), *Applications of Rasch measurement in science education* (pp. 188–211). Maple Grove, MN: JAM Press.

Bayir, E., Cakici, Y., & Ertas, O. (2013). Exploring natural and social scientists' views of nature of science. *International Journal of Science Education, 36*, 1286–1312. doi:10.1080/09500693.2013.860496

Bell, P., & Linn, M. C. (2000). Scientific arguments as learning artifacts: Designing for learning from the web with KIE. *International Journal of Science Education*, *22*, 797–817. doi:10.1080/095006900412284

Blalock, C., Lichtenstein, M., Owen, S., Pruski, L., Marshall, C., & Toepperwein, M. (2008). In pursuit of validity: A comprehensive review of science attitude instruments 1935–2005. *International Journal of Science Education*, *30*, 961–977. doi:10.1080/09500690701344578

Bond, T. G., & Fox, C. M. (2007). *Applying the Rasch model: Fundamental measurement in the human sciences* (2nd ed.). Mahwah, NJ: Lawrence Erlbaum.

Boone, W. J., Townsend, J. S., & Staver, J. (2011). Using Rasch theory to guide the practice of survey development and survey data analysis in science education and to inform science reform efforts: An exemplar utilizing STEBI self-efficacy data. *Science Education*, *95*, 258–280. doi:10.1002/sce.20413

Briggs, D. C., & Wilson, M. (2003). An introduction to multidimensional measurement using Rasch models. *Journal of Applied Measurement*, *4*, 87–100.

Chen, S. (2006). Development of an instrument to assess views on nature of science and attitudes toward teaching science. *Science Education*, *90*, 803–819. doi:10.1002/sce.20147

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*, 408–430. doi:10.1002/tea.21079

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. doi:10.1002/sce.20460

Edwards, M. C., Wirth, R. J., Houts, C. R., & Xi, N. (2012). Categorical data in the structural equation modeling framework. In R. H. Hoyle (Ed.), *Handbook of structural equation modeling* (pp. 195–208). New York, NY: Guilford.

Grandy, R., & Duschl, R. A. (2007). Reconsidering the character and role of inquiry in school science: Analysis of a conference. *Science & Education*, *16*, 141–166. doi:10.1007/s11191-005-2865-z

Huang, C.-M., Tsai, C.-C., & Chang, C.-Y. (2005). An investigation of Taiwanese early adolescents' views about the nature of science. *Adolescence*, *40*, 645–654.

Kamata, A., & Bauer, D. J. (2008). A note on the relation between factor analytic and item response theory models. *Structural Equation Modeling*, *15*, 136–153. doi:10.1080/10705510701758406

Kelley, T., Ebel, R., & Linacre, J. M. (2004). Item discrimination indices. *Rasch Measurement Transactions*, *16*, 883–884.

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: Lawrence Erlbaum.

Lederman, N. G., Abd-El-Khalick, F., Bell, R. L., & Schwartz, R. S. (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*, 497–521. doi:10.1002/tea.10034

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: Kluwer Academic.

Linacre, J. M. (2003). Rasch power analysis: Size vs. significance: Infit and outift mean-square and standardized chi-square fit statistic. *Rasch Measurement Transactions*, *17*, 918.

Linacre, J. M. (2004). Mapping multidimensionality. *Rasch Measurement Transactions*, *18*, 990–991.

Lombrozo, T., Thanukos, A., & Weisberg, M. (2008). The importance of understanding the nature of science for accepting evolution. *Evolution: Education and Outreach*, *1*, 290–298. doi:10.1007/s12052-008-0061-8

Masters, G. N. (1982). A Rasch model for partial credit scoring. *Psychometrika*, *47*, 149–174. doi:10.1007/BF02296272

Muthén, L. K., & Muthén, B. O. (2010). *MPlus user's guide: Statistical analysis with latent variables*. Los Angeles, CA: Muthén & Muthén.

National Research Council. (2012). *A framework for K-12 science education: Practices, crosscutting concepts, and core ideas*. Committee on a Conceptual Framework for New K-12 Science Education Standards. Board on Science Education, Division of Behavioral and Social Sciences and Education. Washington, DC: The National Academies Press.

Neumann, I., Neumann, K., & Nehm, R. (2011). Evaluating instrument quality in science education: Rasch-based analyses of a nature of science test. *International Journal of Science Education*, *33*, 1373–1405. doi:10.1080/09500693.2010.511297

NGSS Lead States. (2013). *Next generation science standards: For states, by states*. Washington, DC: Achieve, Inc.

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. doi:10.1002/sce.21013

Sjaastad, J. (2013). Measuring the ways significant persons influence attitudes towards science and mathematics. *International Journal of Science Education*, *35*, 192–212. doi:10.1080/09500693.2012.672775

Smith, R. M. (2004). Fit analysis in latent trait models. In E. V. Smith & R. M. Smith (Eds.), *Introduction to Rasch measurement* (pp. 73–92). Maple Grove, MN: JAM Press.

Tsai, C. C. (1999). The progression toward constructivist epistemological views of science: A case study of the STS instruction of Taiwanese high school female students. *International Journal of Science Education*, *21*, 1201–1222. doi:10.1080/095006999290156

Wei, S., Liu, X., & Jia, Y. (2014). Using Rasch measurement to validate the instrument of students' understanding of models in science (SUMS). *International Journal of Science and Mathematics Education*, *12*, 1067–1082. doi:10.1007/s10763-013-9459-z

Wenning, C. J. (2006). Assessing nature-of-science literacy as one component of scientific literacy. *Journal of Physics Teacher Education Online*, *3*(4), 3–14.

Wilson, M. (2004). On choosing a model for measuring. In E. V. Smith & R. M. Smith (Eds.), *Introduction to Rasch Measurement* (pp. 123–142). Maple Grove, MN: JAM Press.

Wilson, M. (2011). Some notes on the term: 'Wright Map'. *Rasch Measurement Transactions*, *25*, 1331.

Wright, B. D., & Stone, M. H. (1979). *Best test design*. Chicago, IL: MESA Press.

Wu, M. L., Adams, R. J., Wilson, M., & Haldane, S. A. (2007). *ACER ConQuest: Generalized item response modelling software*. Camberwell, Victoria: ACER Press.

Appendix

Student Science Questionnaire Development

The Student Science Questionnaire development panel was staffed by four evaluators, including two university graduate research assistants and two university faculty with assessment development experience. To establish content validity, members of the panel did the following activities: (a) consulted the PD program staff in developing a test blueprint, (b) reviewed the literature on NOS assessment, (c) identified the feasibility constraints in administering and quantifying (hereafter referred to as scoring) the responses to the instrument, (d) drafted items, (e) solicited item-content review from

five science education experts, (f) revised items and developed new ones to address blueprint gaps and ameliorate construct irrelevant variance, (g) conducted two field tests—each with over 300 Grades 7–12 students, (h) solicited feedback from students and teachers on their experiences during the first field testing, (i) identified each NOS item's alignment to theme categories in the NGSS NOS framework (which had been released just after completion of the field testing) using a combination of online and in-person Delphi-method revisions, and (j) developed a final draft of the instrument. To minimize respondent fatigue and to ensure the instrument could be administered during regular classroom hours, the panel designed the instrument to take less than one hour for students to complete.

NOS Items in the Student Science Questionnaire

MC1. Two students independently tested the boiling point of a liquid. They used liquid from the same source and followed the same procedure, but they got different results. Which of the following do you think is the ***best*** explanation for the difference in their results?

(a) They have different results because they did not conduct the experiment at the same time.
(b) The results of scientific tests are sometimes different because of human or instrument error.
(c) One of the students probably has more experience conducting scientific experiments than the other.

MC2. You are unsure whether one of the scales you used to weigh clams was accurate. You cannot repeat the measurements because the clams have been put back in the ocean. What is the best way to deal with the data from the questionable scale?

(a) Remove it from your analysis and report only the part of your data that you are sure is accurate.
(b) Report it with the rest of your data, but explain which values might be inaccurate and why.
(c) Include it in your analysis because the inaccurate values will average out.

MC3. Accepted scientific explanations change over time because:

(a) scientists conduct experiments more carefully.
(b) scientists share new findings and new information.
(c) scientists vote on the correct explanations.

MC5. When scientists make careful measurements many times, they expect that:

(a) they will have to repeat the measurements until the results are identical each time.
(b) most of the measurements will be identical.
(c) most of the measurements will be close but not exactly the same.

MC6. Scientists today are able to study small parts of living cells with electron microscopes. The invention of the electron microscope is an example of:
(a) how scientists are getting more intelligent over the years.
(b) how technology helps to advance scientific knowledge.
(c) how scientific findings are becoming more accurate.

MC7. A science teacher tells her class to 'think and act like a scientist.' What do you think she means?
(a) In every investigation, follow the scientific method by going from hypothesis to procedure to results and then to the conclusion.
(b) Ask questions, invent procedures, and revise your questions and procedures when you discover better ways to do the investigation.
(c) Revise your questions and procedures until you show that your hypothesis is correct.

MC9. Thinking and acting like a scientist is useful for:
(a) students in all classes.
(b) students who want to become scientists.
(c) students in a science class.

MC10. Malia has conducted the same experiment many times, but her data did ***not*** give her the results she expected. What should she do?
(a) Conduct the experiment again.
(b) Report the results she should have gotten.
(c) Revise her research question and procedures and conduct a new experiment.

MC11. Imagine that you are reading about a scientific study. You need to decide how well the study was conducted. Select the ***best*** question to ask about the study.
(a) Did the study's methods include an experiment?
(b) Were the study's methods appropriate for the study?
(c) Was the study conducted by well-known scientists?

MC13. Emma is a young friend. She wants to know if basketballs bounce higher when they are fully inflated with air. What is the ***best*** thing to do to help Emma learn?
(a) Draw a picture that shows how high two different basketballs will bounce when they have different amounts of air in them. Use your picture when you explain the results to Emma.
(b) With Emma, fully inflate one basketball and partially inflate another basketball. Drop the two balls from the same height. Help Emma explain the results.
(c) Show Emma where the information is in a good science book and have her read it. Later, ask her to explain to you what she learned from the book.

MC14. Kai had a hypothesis. He tested the hypothesis by doing an experiment. Based on the results of the experiment, Kai changed his hypothesis. Which of the following best describes Kai's actions?

(a) Kai was practicing good science because scientists revise their hypotheses based on experimental results.
(b) Kai was practicing good science because he conducted an experiment.
(c) Kai was ***not*** practicing good science because the scientific method goes from hypothesis to procedure to results to conclusion.

MC16. Alex has been studying a gumball machine for two hours. He says he has a theory that the next gumball to come out of the machine will be green. Why is Alex's idea ***not*** a scientific theory?

(a) Alex has only used his own observations to explain why the next gumball will be green. He needs to do an experiment to test his idea.
(b) Alex does not describe a large set of observations and explain those observations. He needs to consider scientific facts, laws, and previous studies.
(c) Alex has not shared his idea with others. He needs to present his observations and idea to others before it can be called a scientific theory.

MC17. Elena is conducting an experiment and finds that her procedure ***cannot*** help her answer her research question. After reviewing her experiment, she develops a new set of procedures for answering the question. What should she do?

(a) Repeat the experiment but use the new set of procedures.
(b) Repeat the experiment with the procedures she used the first time.
(c) Record her current results and finish the investigation.

MC19. Kiana has a hypothesis that bananas will ripen faster when placed in a plastic bowl than when placed on the counter. She places some bananas in a plastic bowl and some bananas on the counter and observes the bananas for a week. Kiana's data show that the bananas in the plastic bowl did not ripen faster. What can Kiana conclude?

(a) Her investigation has failed.
(b) Her hypothesis was not supported.
(c) Her hypothesis cannot be answered using the scientific process.

MC20. Lisa's science teacher asked her class to work in groups and to record observations of fish behavior. One group's findings are very different from the others. Select the ***best*** way for Lisa's class to address the differences in the data.

(a) Compare the groups' procedures to see if this explains different results.
(b) Conduct the observations again but make sure each group's findings are the same.
(c) Have the class vote on which group's data should be accepted.

MC22. Scientific knowledge is determined from:
(a) the votes of knowledgeable individuals.
(b) information that is written in books.
(c) explanations that are supported by evidence.

MC23. Your cat, Frisky, is a vegetarian. Based on this, you hypothesized that cats prefer vegetables to meat. However, the data you gathered show that most cats prefer to eat meat instead of vegetables. This means that:
(a) you conducted a poor experiment because your data did not support your hypothesis.
(b) you showed that your hypothesis about cats was not supported.
(c) your hypothesis was poor because most people already know that cats prefer meat.

MC24. Last week, you noticed that your pet fish swam to the top of its fishbowl when the sun came up. You decided to watch your fish every morning for a week to see if the fish always swims to the top of the fishbowl at sunrise. You are:
(a) doing science because you are repeatedly making careful observations.
(b) doing sloppy science because you are only observing at sunrise.
(c) ***not*** doing science because your observations are too informal.

Directions: Please tell us how true you believe these statements are about science. For each question below, circle the number on the scale that matches how true you think the statement is. Circle **only one number** for each question.

		Not at all true					Very true
Lik1	Many different kinds of people can be good scientists	0	1	2	3	4	5
Lik2	Scientific knowledge can be useful away from school	0	1	2	3	4	5
Lik3	All good scientists work in the same way	0	1	2	3	4	5
Lik4	Scientific knowledge is only useful to scientists	0	1	2	3	4	5
Lik5	Sometimes things that scientists thought were right turn out to be wrong	0	1	2	3	4	5
Lik6	Scientists always get the same results	0	1	2	3	4	5
Lik7	There are many different ways to do science	0	1	2	3	4	5
Lik8	Scientific knowledge can change over time	0	1	2	3	4	5
Lik9	Scientists are always right	0	1	2	3	4	5
Lik10	When you follow the scientific way of doing something, you get the right answer	0	1	2	3	4	5

Supplemental Research Material

This supplement provides additional information for "Comparing Models of Nature of Science Dimensionality Based on the Next Generation Science Standards," George M. Harrison, Kanesa Duncan Seraphin, Joanna Philippoff, Lisa M. Vallin, and Paul R. Brandon, *International Journal of Science Education* (http://dx.doi.org/10.1080/09500693.2015.1035357).

Information on the Standard Deviations of the Standardized Subscale Scores

The standard deviation of the standardized subscale scores provides an indication of how different the students' scores are on the separate dimensions. This approach is based on a method presented in Briggs and Wilson (2003), where they examined the discrepancies in standardized student ability estimates on four dimensions. In their study, they calculated a sum-of-squares index, $\mathrm{DI} = \sum_{d=1}^{4}(\theta_d - \bar{\theta})^2$, where θ_d represents a student's subscale score on Dimension d, and $\bar{\theta}$ is that student's mean dimension score across the four dimensions. Because the magnitude of the sum of the squared deviations depends on the number of dimensions, and our model included six dimensions, we calculated the mean-squared deviation and then square-rooted this value to place it back on the original scale. Thus, for each student, there was a standard-deviation (with *df* = number of dimensions) of standardized subscale scores; That is, $SD = \sqrt{\sum_{d=1}^{6}(\theta_d - \bar{\theta})^2/6}$. A large standard deviation would indicate that the student's estimated subscale scores were markedly different from each other, which would suggest that the multidimensional model would provide more information about a student's NOS-understanding profile than would a score with all the subscales collapsed into a single estimate. An examination of the average of these indexes, across all students, provides a summary of the impact of using multiple subscale scores versus a single NOS score.

Person-Item Maps

Logit	NOS	Content	NOS	Content
+2	X			
	X			
	X			
	X			
	XX			
	XX			
	XXX			
	XXX			
	XXXX			
	XXXX			
	XXXXX			
	XXXXXX	X		
+1	XXXXXXX	X		
	XXXXXXX	X		33
	XXXXXX	X	14	
	XXXXXXX	XX	9	
	XXXXXXX	XX	11 L10	4 18 26
	XXXXXXX	XXX	7 23	21 35
	XXXXXXX	XXXX	10	
	XXXXXXX	XXXX	5	
	XXXXXXX	XXXXXX		8
	XXXXXX	XXXXX		
	XXXXXX	XXXXX	2 17 24 L3	27
	XXXX	XXXXXXX	19 L4	31
0	XXX	XXXXXXXX	1 20	
	XXX	XXXXXXXXXX		30
	XX	XXXXXXXX		
	XX	XXXXXXXXXX	L2 L5	
	X	XXXXXXXX	6 L6	
	X	XXXXXXXX		15
		XXXXXX	3 L1 L7	34
		XXXXX	L8	12 25 28 29
		XXXXX	22	
		XXXX		
		XXXX		32
-1		XX		
		XXX	13	
		X		
		X		
		X		
		X		

Figure S1. Person-item map of Model 2, where NOS is treated as a single dimension. Students are represented on the left, with each X representing 11.4 persons. Items are represented on the right, with the dimension column labels abbreviated; Likert-scale item labels begin with an *L*. Students are positioned from low (bottom) to high (top); i.e. those with higher estimates on the subscales are at the top end of the scale. Items are positioned from easy (bottom) to difficult (top); that is, those estimated to require higher understanding or endorsement are at the top end of the scale.

Logit	NOS theme 1	2	3	5	6	7	Content	NOS theme 1	2	3	5	6	7	Content
+2														
				X										
				X										
				X										
				X										
		X		XX		X								
		X		X		X								
	X	X		X		X	X							
	X	XX	X	XX		X	X							33
+1	X	XX	X	XXX	X	XX	X							
	XX	XXX	XX	XXX	X	XX	XX							4 18 26
	XXX	XXX	XXX	XXX	XX	XXX	XX							21 35
	XXXX	XXXXX	XXX	XXX	XXXX	XXX	XXX							
	XXXXX	XXXX	XXXXX	XXXX	XXXXX	XXXXX	XXXXX							8
	XXXXX	XXXXXX	XXXXXX	XXXXX	XXXXX	XXXXXX	XXXXXX							
	XXXXXXX	XXXXXX	XXXXXXX	XXXX	XXXXXXX	XXXXXX	XXXXXXXX							27 31
	XXXXXXX	XXXXXX	XXXXXXX	XXXXX	XXXXXXXX	XXXXXXX	XXXXXXXXX			14				
	XXXXXXX	XXXXXXX	XXXXXXXXX	XXXXX	XXXXXXXXXX	XXXXXXX	XXXXXXXXX						9	30
0	XXXXXXXX	XXXXX	XXXXXXXXX	XXXXX	XXXXXXXXXX	XXXXXXX	XXXXXXXX	7 11		L10	7			
	XXXXXXX	XXXXXX	XXXXXXXX	XXXXX	XXXXXXXX	XXXXXXX	XXXXXXXXX		5 23		23	5		
	XXXXXXXX	XXXXXXX	XXXXXXXX	XXXXX	XXXXXXXXX	XXXXXX	XXXXXXX	17	2	10	17	24	2	15
	XXXXXXX	XXXX	XXXXXXX	XXXXX	XXXXXX	XXXXXX	XXXXXX		20			20		25 28 34
	XXXXXX	XXXXX	XXXXX	XXXX	XXXX	XXXXX	XXXXXX	L3						12 29
	XXXX	XXXX	XXXXX	XXXX	XXXX	XXXXX	XXXX	L6	19			L6	L4	
	XXX	XXXX	XXX	XXX	XXX	XXX	XXX					1		32
	XXX	XX	XX	XXX	X	XXX	XX							
-1	X	XX	X	XXX	X	XX	X		22	L5	22		L2	
	X	XX	X	XXX		XX	X						6	
		X		XX		X		L7		3			L1	
		X		XX						L8				
		X		X										
		X		X										
		X		X										
				X										
											13			
-2														

Figure S2. Person-item map of Model 4. Students are represented on the left, with each X representing 15.4 persons. Items are represented on the right, with the dimension column labels abbreviated; Likert-scale item labels begin with an *L*. Students are positioned from low (bottom) to high (top); i.e. those with higher estimates on the subscales are at the top end of the scale. Items are positioned from easy (bottom) to difficult (top); that is, those estimated to require higher understanding or endorsement are at the top end of the scale.

Table S1

NOS Item Quality Indexes in Model 4, the Best Fitting Model

Item	Assigned NOS theme(s)[a]	Logit	SE	Unweighted mean-square error fit	Unweighted standardized fit	Weighted mean-square error fit	Weighted standardized fit	Item-measure correlation[b]
MC1	6	-0.69	0.06	0.95	-1.30	0.97	-1.70	.27
MC2	2, 7	-0.34	0.04	1.19	3.80	1.10	3.20	.31, .28
MC3	3	-1.18	0.06	0.94	-1.70	0.97	-1.00	.28
MC5	2, 6	-0.16	0.03	1.10	2.70	1.08	3.30	.32, .32
MC6	7	-1.07	0.06	0.91	-2.40	0.96	-1.50	.33
MC7	1, 5	-0.09	0.03	1.16	4.00	1.12	5.20	.17, .32
MC9	7	0.04	0.06	0.94	-1.70	0.95	-3.90	.35
MC10	3	-0.24	0.06	1.00	-0.10	1.00	-0.30	.23
MC11	1	-0.10	0.06	0.98	-0.50	0.99	-1.10	.27
MC13	5	-1.92	0.09	0.82	-4.20	0.91	-1.50	.35
MC14	3	0.13	0.07	1.09	1.80	1.08	4.90	.07
MC16	(4)							
MC17	1, 5	-0.33	0.03	1.00	0.00	0.99	-0.40	.24, .37
MC19	2	-0.63	0.07	0.95	-1.00	0.97	-1.30	.31
MC20	2, 6	-0.43	0.03	0.86	-3.90	0.91	-3.50	.44, .38
MC22	2, 5	-0.92	0.05	0.81	-4.30	0.97	-0.50	.40, .45
MC23	2, 5	-0.14	0.04	1.04	0.80	1.03	1.00	.43, .44
MC24	1, 6	-0.29	0.04	1.11	2.30	1.08	3.20	.19, .19
Lik1	7	-1.24	0.03	1.07	1.70	1.07	1.90	.39
Lik2	7	-1.01	0.03	1.08	2.10	1.04	1.00	.42
Lik3	1	-0.55	0.02	0.97	-0.70	0.98	-0.70	.54
Lik4	7	-0.66	0.02	0.98	-0.60	1.00	0.10	.54
Lik5	3	-0.96	0.03	1.01	0.20	1.01	0.20	.40
Lik6	1, 6	-0.62	0.01	0.98	-0.40	0.99	-0.40	.55, .48
Lik7	1	-1.22	0.04	1.00	0.10	1.01	0.20	.42
Lik8	3	-1.30	0.03	0.93	-1.80	0.95	-1.30	.44
Lik9	(8)							
Lik10	3, (8)	-0.07	0.03	1.10	2.70	1.08	2.20	.41

Note: The output is from ConQuest 3. NOS Theme 1 = Scientific investigations use a variety of methods; NOS Theme 2 = Scientific knowledge is based on empirical evidence; NOS Theme 3 = Scientific knowledge is open to revision in light of new evidence; NOS Theme 4 = Science models, laws, mechanisms, and theories explain natural phenomena; NOS Theme 5 = Science is a way of knowing; NOS Theme 6 = Scientific knowledge assumes an order and consistency in natural systems; NOS Theme 7 = Science is a human endeavor; NOS Theme 8 = Science addresses questions about the natural and material world.

[a] Themes 4 and 8, marked by parentheses, were excluded from the model.

[b] Two correlations are presented, in their respective order, for items measuring two dimensions. The average of the two was used to estimate item quality.