Evaluating instrument quality in science education: Rasch-based analyses of a Nature of Science Test

Journal:	*International Journal of Science Education*
Manuscript ID:	TSED-2009-0413.R2
Manuscript Type:	Research Paper
Keywords:	nature of science
Keywords (user):	Scientific Inquiry, Rasch model, instrument evaluation

Taylor & Francis and I.T. Consultant
International Journal of Science Education
Vol. 00, No. 00, 00 Month 2011, 1–25

Evaluating instrument quality in science education: Rasch-based analyses of a Nature of Science Test

Given the central importance of the nature of science (NOS) and scientific inquiry (SI) in national and international science standards and science learning, empirical support for the theoretical delineation of these constructs is of considerable significance. Furthermore, tests of the effects of varying magnitudes of NOS knowledge on domain-specific science understanding and belief require the application of instruments validated in accordance with AERA, APA, and NCME assessment standards. Our study explores three interrelated aspects of a recently developed NOS instrument: (1) validity and reliability; (2) instrument dimensionality; and (3) item scales, properties, and qualities within the context of Classical Test Theory and Item Response Theory (Rasch modeling). A construct analysis revealed that the instrument did not match published operationalisations of NOS concepts. Rasch analysis of the original instrument--as well as a reduced item set--indicated that a two-dimensional Rasch model fit significantly better than a one-dimensional model in both cases. Thus, our study revealed that NOS and SI are supported as two separate dimensions, corroborating theoretical distinctions in the literature. To identify items with unacceptable fit values, item quality analyses were used. A Wright Map revealed that few items sufficiently distinguished high performers in the sample and excessive numbers of items were present at the low end of the performance scale. Overall, our study outlines an approach for how Rasch modeling may be used to evaluate and improve Likert-type instruments in science education.

Key words: Nature of Science, Scientific Inquiry, construct, Rasch model, IRT, Wright Map, validity, reliability, instrument quality

INTRODUCTION

The Nature of Science[1] (NOS) has been a topic of longstanding significance in science education (e.g., Abd-El-Khalick, 2006; Lederman 2007; Lombrozo, Thanukos, & Weisberg, 2008; National Research Council [NRC], 1998; Rubba, & Andersen, 1978; Wilson, 1954). The US National Science Education Standards (NSES; NRC, 1998) and Benchmarks for Science Literacy (American Association of the Advancement in Science [AAAS], 1993), for example, explicitly enumerate several NOS aspects of central importance to science curricula. Comparable standards exist internationally (for an overview, see McComas & Olson, 1998). More broadly, NOS is considered an important element of what has been termed 'scientific literacy'. Bybee (1997, p. 61), for example, emphasised that '[s]cientific literacy extends beyond vocabulary, conceptual schemes, and procedural methods to include other understandings about science…[e.g.] the history of scientific ideas, the nature of science and technology, and the role of science and technology in personal life and society'. Driver, Leach, Millar, and Scott (1996) put forth the most encompassing and diverse arguments for the importance of NOS instruction, and include utilitarian, democratic, cultural, moral, and 'science learning' arguments for including it as a necessary component of scientific literacy and school instruction.

Despite this rather large body of work attempting to justify the inclusion and importance of NOS in science curriculum and instruction, considerably less attention has been directed at the issue of whether NOS understanding in fact has a significant and meaningful impact on domain-specific science learning or scientific literacy. Indeed, Lederman (2007, p. 832) pointed out that 'the arguments are primarily intuitive, with little empirical support.' This

[1] While we are aware that there is a recent debate regarding use of the definite article (e.g. Abd-El-Khalick, 2006), we have decided to use '*The* Nature of Science' being the most grammatically appropriate expression.

perspective is surprising, given that more than 25 NOS instruments have been developed over the past 50 years (for an overview, see Lederman, 2007). Considerable effort has been directed at NOS construct delineation, instrument development and evaluation, and NOS measurement in different populations and samples (Lederman, 2007). Comparatively less work has focused on the putative relationships among NOS understanding and domain-specific learning outcomes (however, see Lombrozo et al., 2008).

Robust conclusions about the effects of varying magnitudes of NOS knowledge on learning outcomes will require the application of rigorously validated and reliable measurement instruments (American Educational Research Association [AERA], American Psychological Association [APA], and National Council on Measurement in Education [NCME], 2004; Authors, 2006; Authors, 2008; Authors, 2010). Although much work has focused on the development of open-response NOS instruments for use with K-12 students (e.g. the VNOS series, Lederman, 2007), there are very few recently developed instruments suitable for use with college students (e.g. Abd-El-Khalick, 2006; Bezzi, 1999; Fleming, 1988; Gilbert, 1991; Ryder, 1999). Rigorously evaluated closed-response instruments would be of great use, as they could be employed to empirically test hypotheses regarding the relationships among domain-specific knowledge and NOS understanding in large samples. Conversely, employing weakly evaluated instruments to test these hypotheses contributes little to resolve whether NOS understanding has a significant relationship to students' science achievement.

Likert-type items and instruments are commonly used in science education research (e.g. Coulson, 1992: the ECEASS; Glynn, Taasoobshirazi, & Brickmann, 2009: the SMQ; Kaya, Yager, & Dogan, 2009: the QASTS; Rutledge & Warden, 2000: the MATE; Southerland, Settlage, Johnston, Scuderi, & Meadows, 2003: the STILT). In the area of NOS research, Likert-type items are also quite common. Of the 23 NOS instruments we could locate in Lederman's review (2007, p. 862), for example, 11 are Likert-type. Despite their common

use, many methodological issues confront users of Likert-type instruments. Quantitative analyses of Likert-type items and instruments, for example, often fail to test whether the item scores meet assumptions of normality (that is, whether subsequent statistics of central tendency are appropriate; e.g. Rutledge & Warden, 2000). If Likert-type data are bimodal or skewed (a common situation, for example, in evolution education instrument data from the MATE, Rutledge & Warden, 2000), then calculating item means is inappropriate (and interpretations may be misleading). Additionally, a question that is inherent to Likert scales is whether participants answered all items in a consistent fashion; that is, whether all category options were comparably scaled by the respondent, independent of the item. The magnitude of agreement differences from, for example, 'disagree' to 'not sure', or from 'agree' to 'strongly agree', may not be of the same quantitative magnitude throughout all items. Only in cases where a consistent scale may be reasonably assumed for each item is the calculation of person sum scores on the whole instrument (or parts of it) meaningful. Usually, Likert-type data are treated as if they do in fact meet these two properties (normal distribution and scale consistency) even if they do not.

Applying Item Response Theory models, such as Rasch models, provide significant advantages for the development and evaluation of Likert-type items and instruments (Bond & Fox, 2001). Rasch analysis converts ordinal data into ratio-scaled data and produces item parameters and person parameters that are of a ratio level of measurement (Bond & Fox, 2001). Rasch-based analyses are also able to test whether item/scale comparability exists for a given sample by testing if all items are answered in the same fashion or not. This allows empirical testing of Likert-type scale assumptions. It also allows for comparisons of students and items on quantitatively equivalent intervals.

Lombrozo et al. (2008) recently developed a Likert-scale instrument for quantifying NOS knowledge in undergraduate students. They employed this new instrument in a series of tests

designed to measure the relationships among NOS knowledge and evolutionary beliefs. Many psychometric aspects of their instrument were unfortunately not provided. In particular, aspects of the validity and reliability of the instrument remain in need of further investigation. Likewise, the instrument was piloted on a sample of psychology students; it is unclear as to whether the instrument would also be appropriate for samples of undergraduate science students.

Our study explores three interrelated aspects of Lombrozo et al.'s NOS tool: (1) validity and reliability; (2) instrument dimensionality; and (3) item scales, properties, and qualities. We examine these attributes within the context of Rasch modeling, an approach for constructing and evaluating measures, which has gained in importance in recent years (Bond & Fox, 2001, Wilson, 2004). We illustrate some of the strengths and limitations of the Lombrozo et al. instrument and provide suggestions for improving it. Given the putative importance of NOS knowledge, our study attempts to advance instrument quality in this research area. More broadly, we hope that our investigation will help to outline a general approach for evaluating many important aspects of Likert-type instruments--and interpreting resulting scores--in science education research using Rasch modeling.

INSTRUMENT AND SAMPLE

Like in the study carried out by Lombrozo et al. (2008), the NOS instrument (Lombrozo, personal communication, October 6, 2008) was administered to a sample of N = 214 college students. The students were undergraduate science majors (the vast majority of which were biology-related) at the end of a second-quarter biology course at a large Midwestern research university in the United States. We took this approach because we assumed that undergraduate science majors would have more or less adequate views of NOS, thereby providing us with information about how well the instrument performed in this capacity. The average age of undergraduates in the sample was 21 years (minimum 19, maximum 31);

59.3% were female; and 23.3% were minorities. The average cumulative grade point average of the sample was 3.2 out of a possible 4.0.

The NOS instrument contained 12 NOS 'themes', each consisting of five Likert-type items (strongly agree, agree, not sure, disagree, strongly disagree). The original instrument contained three items per theme with positive valences and two items per theme with negative valences. In order to prevent student identification of this pattern during testing--but not deviate extensively from the format of the original instrument--reverse valences were applied to three additional items. Thus, the themes *applications*, *provisionality*, and *limits* included three items with negative valences whereas the remaining nine themes contained two items with negative valences. In addition, five items covering Lombrozo et al.'s theme 'limits of scientific inquiry' were included (as in Lombrozo et al. 2008). The NOS instrument was administered to our sample in a manner similar to Lombrozo et al. (2008) except that our computer system was not able to randomise theme order as was done by Lombrozo et al. 2008.

As a matter of fact, designs based on self-scored questionnaires are prone to social desirability bias (i.e. students answering in a manner they think is favored by the teacher); so-called Rosenthal effect (i.e. students improving in their performance according to the teacher's expectation [Rosenthal & Jacobson, 1992]) and Hawthorne effect (i.e. students improving in their performance because they are part of an experimental study [Adair, 1984]). As this study aims to investigate instrument quality, the instrument was administered to the students just after the beginning of the term. As students did not receive any formal teaching on NOS at this point and no experimental study was undertaken, effects related to the lecturers' expectations (cf. Rosenthal or Hawthorne effects) are likely to be minimal. To reduce bias related to social desirability, participants were told that their performance on the questionnaire would not affect their course grade. Additionally, the design of the instrument using items

with both positive and negative valences attempts to control for social desirability effects. Nevertheless, these or other unwanted effects may introduce bias, which necessitates careful analysis and interpretation of the data along with the consideration of these potential artifacts.

METHODS

Validity and reliability

Validity. The standards for educational and psychological testing (AERA et al., 2004, p.1) provide clear benchmarks for instrument development and emphasise the importance of 'evaluating the quality of testing practices'. Validity is concerned, in part, with the theoretical grounding of an investigated construct, such as NOS. As noted by Furr and Bacharach (2008, p. 173), '[c]ontent validity is the degree to which the content of a measure truly reflects the full domain of the construct for which it is being used'. Ensuring that this 'full domain' is reflected in an instrument can be achieved in various ways, including: (1) precisely specifying the content that the instrument developer considers to be part of the construct (and its correspondence to particular instrument items); (2) confirming that the subject matter literature is in alignment with the particular construct definition used by the instrument developers; and (3) establishing expert consensus about the construct using Delphi (or related) methods (AERA et al., 2004; Best & Kahn, 2003; Osborne, Collins, Ratcliffe, Millar, & Duschl, 2003). While these three approaches clearly overlap--expert literature, for example, will likely corroborate expert judgments--this may not always be the case (e.g. Alters, 1997a; Alters, 1997b; Smith, Lederman, Bell, McComas, & Clough, 1997).

Lombrozo et al. (2008) did not provide a clear justification for their conceptualisation of the construct of NOS. Therefore, in order to determine whether or not Lombrozo et al.'s NOS measure encompassed the full domain of the construct that it intended to measure (i.e. NOS), we conducted a literature review in order to delineate the construct as it is currently accepted

and operationalised by the science education community. A panel of the three science educators placed the items from the Lombrozo et al. instrument into the delineated construct. In cases of inconsistent placement by panelists, consensus was eventually reached through group deliberation. Comparing the construct introduced by Lombrozo et al. (2008) to the results of this review serves to clarify the theoretical grounding of their construct and to reveal discrepancies with current views of NOS. These findings are a necessary prerequisite to interpreting the meaning of the scores that the Lombrozo et al. instrument produces (i.e. AERA et al., 2004, p. 9). In short, these investigations attempt to address the question of whether the Lombrozo et al. instrument sufficiently and appropriately measures NOS knowledge.

Reliability. In addition to validity, our study investigated the issue of instrument reliability (AERA et al., 2004, p. 25). A high quality and valuable instrument is expected to provide consistent measurement outcomes (i.e. it produces reliable inferences). We investigated the reliability of the Lombrozo et al. instrument using both classical test theory (CTT) and item-response theory (IRT; specifically, Rasch modeling). Within CTT, one can distinguish between several reliability coefficients depending on which approach is used for replicating the measurement (e.g. test-retest reliability, reliability of parallel forms, or internal consistency). Given that only one form and one single test were administered, the appropriate CTT approach available for investigating reliability in our study is internal consistency. Internal consistency splits the test into two or more parts, and consistency within those parts is calculated. The most widely used coefficient of internal consistency is Cronbach's α (Haertel, 2006). Using α eliminates a source of error associated with an arbitrary split choice. It is also one of the most commonly used measures of reliability in science education.

Our analyses of the reliability inferences of the Lombrozo et al. (2008) instrument included (1) calculations of Cronbach's α (CTT) and (2) Expected A Posteriori/Plausible Value

tsed-00637-2014.R1 Version 1 - 29 Oct 2014

reliability (IRT). IRT reliability calculations using Rasch modeling (discussed in more detail in the section below) involve every participant being assigned an estimated ability value expressed as a score distribution. Predictive reliability is 1 minus the ratio of the variance of one participant's score distribution relative to the sample variance. Expected EAP/PV reliability (A Posteriori/Plausible Values) represents the mean of such predictive reliabilities in the sample and, thus, is a measure of overall sample reliability (for details, see Zoanetti, Griffin, & Adams, 2006). This reliability value is not the same as but can be interpreted much like Cronbach's α, in which values > 0.7 are supportive of reliability inferences (for a discussion of acceptable α values see Field, 2009). As Lombrozo et al. (2008) did not provide any data regarding reliability inferences derived from their sample, we were unable to compare our reliability findings to theirs.

Rasch modeling

Our study investigates the quality of the NOS instrument developed by Lombrozo et al. (2008) using Rasch modeling, which is a type of item response theory (IRT) (Bode & Wright, 1999; Liu & Boone, 2006). IRT, as well as classical test theory (CTT), comprise two related families 'of statistical models used to analyse test item data' (Yen & Fitzpatrick, 2006, p. 111). Yen and Fitzpatrick (2006) discuss a number of differences between these two families, but one of the more significant differences relates to scaling considerations. Our analyses of the Lombrozo et al. instrument using Rasch involve ordinal scale and ratio scale data, so a brief note about these scales may be helpful (for a detailed discussion, see Field, 2009; those familiar with this topic may choose to move on to the next paragraph). Ordinal measurements include categorical data, but in contrast to other types of categorical data (such as 'yes/no' or 'red/yellow/blue'), ordinal categories have an inherent order (e.g. 'no agreement/medium agreement/strong agreement'). Despite this inherent order, these categories are not scaled at equal intervals. In contrast to ordinal data, data on a ratio scale are continuous (e.g. running

time is measured in minutes and seconds), ordered (one minute running is less than two minutes running), and--unlike ordinal measurements--scaled at equal intervals (the difference between one and two minutes is the same as between 11 and 12 minutes). Additionally, ratios of two values from a ratio scale data set are meaningful whereas those from ordinal scales are not (e.g. five minutes of running is only half as much as ten minutes of running, whereas the category 'slow speed runners' does not necessarily represent a speed half as fast as the category of 'medium speed runners').

Unlike IRT-derived scores, CTT methods use raw scores. One important limitation of such scores is a consequence of their categorical/ordinal nature. Assume, for example, that four students take the same physics test. Student A receives 90 correct answers, student B 100, student C 125, and student D 135. Clearly, the difference between students A and B, as well as between students C and D, is 10 correct answers. Since this sum score of correct answers is ordinal, it is an open question as to whether these score differences among students are equivalent (as, for example, degrees of temperature would be using a thermometer). It is possible, for example, that: student B answered all difficult items correctly; student A only answered the easy items correctly; and students C and D answered items correctly at about the same level of difficulty. In order to determine if the 10 points separating these students were equivalent or not, we require equal interval measures.

Boone and Scantlebury (2006) used the analogy with a meter stick to illustrate the issue of scale equivalency. Again, consider the four participants, but now consider their scores in terms of height (A is 90 cm, B 100 cm, C 125 cm, and D 135 cm tall). Since their height was measured by a meter stick providing a ratio scale, one is now able to say that B is taller that A to the same amount that D is taller than C. Moreover, it is meaningful and reasonable to say that student B is twice as tall as student E, who is only 50 cm tall. This simple analogy elucidates the useful properties that ratio scales provide educational measurements. Analysing

instrument scores on a ratio scale--as provided by Rasch modeling--allows students' performances to be compared directly. Boone and Scantlebury (2006) emphasise that using Rasch models in science education provides methodological rigor and 'confidence in students' computed scores' (p. 256). The following sections describe other advantages of applying the Rasch model, specifically within the context of instrument evaluation. For additional reviews of this growing area of psychometric research, see Bond and Fox (2001), Boone and Scantlebury (2006), Wilson (2004), and Yen and Fitzpatrick (2006).

Instrument Dimensionality. Instruments are typically designed to capture one or more attributes (or 'traits') about the knowledge, performance, or attitudes held by a sample. Quite often, different item sets are used to measure different traits (or aspects of a single trait). These aspects may be referred to as instrument 'dimensions'. Particular assemblages of items in an instrument are typically used to measure these different dimensions. In theory, an individual Rasch analysis may be used to measure each of these dimensions; however, a multidimensional Rasch analysis permits analysis of the dimensionality of a given set of items. In other words, Rasch analysis may be used to determine empirically whether a set of items is in fact measuring two or more different traits. If there is a theoretical justification for assuming multi-dimensionality for a set of items (that is, the instrument was designed to measure several traits), then a multi-dimensional as well as a uni-dimensional Rasch model may be employed to examine instrument structure.

Similar commonly used methods of instrument structure examination include principal component analysis (PCA) and confirmatory factor analysis (CFA). The main difference between CFA and Rasch analysis on the one hand, and PCA on the other hand, is that the first two investigate how well the data fit a hypothesised model, while the latter is a method of data reduction. That is, whereas Rasch analysis provides fit statistics with respect to how well an item fits into a dimension of a hypothesised trait structure and CFA determines the factor

Peer-00637625, version 1 - 29 Oct 2011

loadings of an item with respect to different dimensions, with PCA high-correlating items can be identified with the purpose of removing unnecessary items. When developing a questionnaire to survey a theoretically well-defined trait such as views of NOS, Rasch analysis and CFA would be the more appropriate methods to investigate whether the questionnaire actually represents the theoretically hypothesised structure. Although the results obtained from Rasch analysis and CFA can principally be converted into each other (Edwards & Wirth, 2009), the procedures are based on different assumptions regarding the original data: CFA requires the original data to be normally distributed on an interval scale, whereas polytomous Rasch analysis only requires the data to be ordinal. Thus, although CFA is commonly used to analyse Likert-based questionnaires (assuming Likert-type questions are discrete representations of a continuous interval scale), Rasch analysis would be the more appropriate method in this respect, which is why we chose Rasch analysis for the investigation of the instrument developed by Lombrozo et al. (2008).

When investigating dimensionality using Rasch analysis, a log-likelihood test may be performed to determine statistically whether one or more dimensions in fact characterise a dataset, thereby determining whether multi-dimensionality is supported empirically. Within this test, so-called final deviance of each model is compared. Final deviance is a measure indicating the likelihood of the observed data fitting the assumptions of the estimated model. A smaller likelihood value indicates a better fit. Comparing the efficacy of two models therefore requires comparing their final deviances. Since the deviances, and, thus, their differences are χ^2 distributed, a comparison to a critical value in a χ^2 distribution indicates if this difference is in fact significant. Degrees of freedom are determined by the difference of the number of parameters that are estimated. If such a log-likelihood test reveals a more than unidimensional structure, correlations between those latent dimensions can determine whether students' abilities with respect to these dimensions are parallel, antiparallel or independent.

We performed dimensionality tests with our dataset using Conquest 2.0 (Wu, Adams, & Wilson, 2007). Specifically, we compared a 13-dimensional model (grouping the items in the same themes as Lombrozo et al. 2008) to a unidimensional model (all items together) and a two-dimensional model (NOS items vs. SI items, parsed in accordance with the constructs defined below). These analyses were used to draw conclusions about the dimensionality of the Lombrozo et al. instrument, in particular to test the hypothesis that different NOS themes are supported empirically. When comparing different models, the simplicity of the model needs to be taken into account (cf. Occam's Razor). In the context of Rasch analysis, the simplicity of the model is determined by the number of estimated parameters (n_p). Coefficients based on information theory can be used to compare models. Common coefficients are Akaike's Information Criterion (AIC: where AIC = deviance+2·n_p) and Bayes' Information Criterion (BIC: where BIC = deviance+log$N$·n_p; Wilson, deBoek & Carstensen, 2008). Since BIC weighs simplicity of the model by sample size, BIC is recommended especially for larger sample sizes. Regarding AIC and BIC, lower coefficients are desirable. The disadvantage of such information-based criteria lies in the lack of a significance test. Thus, based on AIC and BIC, models can only be compared relative to one another and researchers have to decide which difference in AIC or BIC is sufficient to indicate the advantage of one model over the other. In addition to deviance statistics, we also employed AIC and BIC to compare models characterised by different dimensionalities.

Generally speaking, in order to determine the best fitting model, researchers must take into account several criteria (for a thorough discussion on model fit, see Bond & Fox, 2004). These criteria include: (1) statistical criteria, e.g., deviance statistics, information criteria, and misfitting items (discussed in more detail in the section "Item quality and item redundancy" below), and (2) non-statistical criteria, e.g., the underlying theoretical basis of the instrument,

and aspects of instrument content. Collectively, researchers must weigh how these criteria inform interpretations of model fit.

Likert-scales and item properties. Likert-type instruments are quite common in science education research (e.g. Rutledge & Warden's MATE instrument, 2000). Employing such instruments, individuals are prompted to choose their level of agreement with a particular statement, and these levels of agreement are coded numerically (e.g. 1 = strongly disagree, 2 = disagree, 3 = unsure, 4 = agree, and 5 = strongly agree). While seemingly simple computationally, methodological issues may complicate interpretation of these scales and their resultant scores. One issue that is inherent with Likert scales is the question of whether participants answered all items in a consistent fashion; that is, every answer category option is assumed to have produced a comparably scaled conception in the respondent, independent of the different items. A second issue with Likert scales is that they are ordinal (see above) and as such their response options are typically not of equal intervals. For example, the magnitude of agreement from 'disagree' to 'not sure', and from 'agree' to 'strongly agree', may not differ in the exact same magnitude. Usually, Likert type data are treated as if they meet these two properties even if they do not. Applying the Rasch model transforms Likert type scores, and thus, ordinal data into ratio-scaled data. Based on observed response patterns, Rasch models produce item parameters and person parameters that are of a ratio level of measurement. This allows appropriate comparisons of students and items by comparing quantitatively equivalent intervals. The second issue taken into account by Rasch models is testing for scale effects, as described below.

In general, there are two different methods for obtaining Rasch measures from a dataset using Likert-scale items: employing (1) a Rating Scale Model (RSM) or (2) a Partial Credit Model (PCM). In the RSM, one single, consistent, and unchanging measurement scale is assumed to characterise all items in an instrument (Wu, Adams, & Wilson, 2007). The PCM, on the other

hand, originates from a desire to make use of the advantages offered by Rasch analysis. This model assumes that different measurement scales characterise different items in an instrument. That is, it is not assumed that each pair of adjacent categories (e.g. strongly agree, agree) is equidistant from one another among all items. Thus, PCM takes into account scale differences that may occur among items. The PCM is mostly used when 'richer data than the dichotomous data that are typically generated by traditional assessment practices are available' (Wu et al., 2007, p. 3).

It is possible to test which of these two models (RSM, PCM) better fits a Likert-type data set, thereby determining whether item/scale comparability exists for a given sample. As described in the preceding section, a common approach for comparing the two models with respect to fit is to employ a log-likelihood test. In doing so, it is possible to determine whether it is more likely that the answering scale on an instrument is perceived comparably by participants across all items or not (cf. Wu et al., 2007). We used Conquest 2.0 (Wu et al., 2007) to calculate deviance, upon which log-likelihood tests were calculated.

Independent of the issue of scale is the issue of score distributions. If Rasch models are not used, then particular item response distributions will constrain the analysis options available for calculating instrument summary scores. For example, if Likert-type data are not distributed normally (for example, in evolution education research bimodal or skewed distributions are common), then calculating item means is inappropriate (and interpretations may be misleading). A common approach for investigating the distribution of scores in a data set--and thereby determining if they are suitable for raw score averaging--is performing a Kolmogorov-Smirnov (KS) test (see also Field, 2009). We ran KS tests on all items in the dataset to determine if averages of raw scores could be used.

Item quality and item redundancy. Rasch modeling may also be used to analyse the quality of items within an instrument as a whole, irrespective of the instrument's dimensionality. For

each item, fit indices are calculated (Bond & Fox, 2001), i.e. infit and outfit (both are based on a mean square [MNSQ] statistic) and standardised z scores (ZSTD). These fit measures indicate how well an item fits with the estimated Rasch model (see above). Item fit also indicates the quality of an estimated model. The more items that fit a model, the better the model is assumed to be. Moreover, a traditional index, discrimination, can be used for item evaluation. This index is calculated by the correlation between the persons' score on a particular item and their sum score; it represents an item's power to distinguish among persons. Low discrimination values indicate that a particular item cannot distinguish among students with different mastery levels of expertise with respect to the measured trait. If items are intended to measure only whether students mastered particular skills or learned particular content, analysing the item's discrimination would be inappropriate for evaluating item quality. However, as the instrument at hand (and thus its items) was initially developed to differentiate among students with respect to their NOS views (cf. Lombrozo et al., 2008), our analyses of item discrimination are appropriate.

In addition to identifying poorly fitting items, Rasch modeling allows for the clear identification of redundant items (that is, those that do not help to differentiate participant performance). A so-called 'Wright Map' (or person-item map; Boone & Scantlebury, 2006; Wilson, 2004) may be used to visually display the simultaneous distributions (or 'performances') of items and persons for a particular sample and instrument. Redundant items appear on the same, or nearly the same, point on the Wright map scale. The Wright map also illustrates regions of the scale in which items are absent; that is, it identifies where new instrument items may be needed. Of course, when excluding redundant items from a questionnaire, aspects of construct validity must be taken into account. That is, the revised version of a questionnaire must be re-evaluated with respect to validity.

We used Conquest 2.0 (Wu, et al., 2007) to calculate item fit (infit, outfit, ZSTD, and discrimination values) and produce a Wright Map. These analyses were used to empirically evaluate the quality of items in the Lombrozo et al. (2008) instrument.

RESULTS

Validity

AERA et al. (2004, p. 9) describes validity as 'the most fundamental consideration in developing and evaluating tests'. A central aspect to evaluating the validity of a NOS instrument is determining whether it covers the construct as it is currently delineated and operationalised by the scholarly community. NOS is a large and amorphous construct that has at times proven difficult to circumscribe. Indeed, Abd-El-Khalick (2006, p. 391) noted that '[p]hilosophers, historians, sociologists of science, and science educators are quick to disagree on a definition for NOS'. Fortunately, several recent attempts to delineate--or at least constrain--the construct of NOS from a science education perspective have been met with success. McComas and Olson (1998) reviewed eight science education documents from the U.S., Australia, England/Wales, New Zealand, and Canada in order to find a definition of NOS 'useful in informing science teaching and learning' (p. 41). Importantly, their work revealed that: 'there is clearly consensus regarding the nature of science issues that should inform science education' (p. 48).

Osborne, Collins, Ratcliffe, Millar, and Duschl (2003) used a Delphi-study approach to investigate if consensus could also be reached among experts regarding what aspects of NOS should be taught in schools. Their study of 23 experts from the fields of science, history, philosophy, and sociology of science; science education; public understanding of science; science communication; and teaching revealed nine themes on which there was consensus and a stable agreement on their importance. Lederman, Abd-El-Khalick, Bell, and Schwartz

(2002) likewise argued that consensus about what constituted the construct of NOS could in fact be reached. Moreover, they carefully considered the relevance of NOS aspects for K-12 students' daily lives and their relevance in science education. Specifically, they outlined eight aspects of NOS that met these criteria: 'scientific knowledge is tentative; empirical; theory-laden; partly the product of human inference, imagination, and creativity; and socially and culturally embedded. Three additional important aspects are the distinction between observation and inference, the lack of a universal recipe-like method for doing science, and the function of and relationship between theories and laws' (p. 499). Thus, consensus has emerged regarding many elements of the construct of NOS.

Recently, Lederman (2006) distinguished 'the nature of scientific knowledge' (abbreviated as NOS) from 'the process of scientific inquiry (SI)'. Schwartz, Lederman, and Lederman (2008) went on to develop a framework that specified the elements of the nature of scientific inquiry. Their conceptualisation is closely related to 'Understanding about Scientific Inquiry' within the National Science Education Standards (NRC, 2000, p. 20). Based on standards documents and research reports, Schwartz et al. (2008) identified seven aspects of SI on which there was consensus and on which a case of relevancy for science education could be made. These themes included: 'a) questions guide investigations, b) multiple methods of scientific investigations, c) multiple purposes of scientific investigations, d) justification of scientific knowledge, e) recognition and handling of anomalous data, f) sources, roles of, and distinctions between data and evidence, and g) community of practice' (Schwartz et al., 2008, p. 4). Thus, within the field of science education, the working definitions of NOS and SI have now been clearly delineated.

In their study, Lombrozo et al. (2008, p. 292) did not make reference to the past 30 years of work on NOS construct delineation, but they did refer to and base their new instrument on the Student Understanding of Science and Scientific Inquiry instrument (SUSSI, developed by

Liang et al., 2006). Although the acronym 'SUSSI' is inherently suggestive of measurement of both constructs (NOS and SI), Lombrozo et al. (2008) refer to 12 themes that they consider to characterise NOS as well as one more theme they identify as 'limits of scientific inquiry' (p. 292, which they apparently consider to be distinct from NOS). Other than Lombrozo et al.'s (2008) reference to Liang et al. (2006), who themselves apparently developed their instrument based on the work of Lederman et al. (2002), it is not clear on what theoretical grounding Lombrozo et al. based their 12 (or perhaps 13) themes about NOS, or how they correspond to the current or past literature about the construct of NOS. Given a lack of theoretical grounding (AERA et al., 2004), we required a vantage point from which to conceptualise Lombrozo et al.'s formulation of NOS. Figure 1 situates Lombrozo et al.'s conceptualisation of NOS relative to the current views of the science education community.

[INSERT FIGURE 1 ABOUT HERE]

Item-construct relationships

Given the absence of explicit and precise theoretical grounding, we mapped Lombrozo et al.'s 'NOS' instrument items onto the current operationalisation of NOS and SI by the science education community (described above). Several conclusions may be drawn from this content analysis. First, the distribution of the Lombrozo et al. items across the 14 aspects of NOS and SI are very unbalanced (Table 1). The SI aspect '*Multiple methods of scientific investigations'* is represented by ten items; the NOS aspect '*Empirical'* is represented by 11 items; and the NOS aspect of *'Theory ladenness'* is represented by only one item (Table 1). Second, two items (#48 and 49) could not be clearly assigned to any of the NOS or SI aspects noted in the literature (both of which were assigned to the theme '*Continuity'* within Lombrozo et al.) and the NOS aspect '*Distinction between data and evidence'* is not addressed by any item. Third, five items (# 61-65) placed within Lombrozo et al.'s '*Limits'* theme--which was envisioned as separate from their other NOS items could be assigned to the '*Empirical'* NOS aspect. This

indicates that the '*Limits'* theme is not in fact separate from the main construct of NOS as operationalised by the science education community. Fourth, not all items within a theme hold together or match a single aspect of NOS (or SI). We provide three examples of such conceptual fragmentation using three items from the Lombrozo et al.'s theme of *continuity*. Item 46, ('Scientific investigations usually lead to additional questions for further investigation.') most closely matches the aspect of SI known as '*scientific questions guide investigations'*. In contrast, item 47 ('Science is mainly a collection of facts that can be described in textbooks.') most closely matches the aspect of NOS known as '*tentative'*. Finally, item 48 ('The process of science is iterative; new scientific investigations build on previous scientific knowledge.') does not clearly match any of the NOS or SI aspects (the same is true of item 49). Table 1 (see also the Appendix) provides a detailed categorisation of all of Lombrozo et al.'s items relative to the 14 aspects of the constructs of NOS and SI delineated in the literature. Items 48 and 49 could not be unambiguously assigned to any aspects of NOS or SI and were therefore omitted from the subsequent analyses. Omitting these items does not necessarily mean that they are not part of NOS or SI, but it rather indicates that they are not part of how the science education community currently conceptualises and operationalises NOS and SI.

[INSERT TABLE 1 ABOUT HERE]

Dimensionality and Items

The Lombrozo et al. (2008) instrument initially contained 12 NOS themes and one 'limits' theme. However, after performing a factor analysis on these themes, Lombrozo et al. subsequently collapsed the 12 NOS themes into three components. Given the lack of explicit theoretical grounding or support for these numerous 'themes' or three 'units', it is not clear whether Lombrozo et al. considered their instrument to represent 13 dimensions (12 NOS themes and 1 limits theme); four dimensions (three units and one limits theme); two

dimensions (one large NOS unit and one Limits theme); or possibly one large interrelated dimension. Additionally, we were not able to map any of these themes or units onto any conceptualisation of NOS or SI found in the literature (see Methods). For these reasons, we analysed Lombrozo et al.'s instrument relative to the current theoretical grounding of NOS and SI noted in the literature (see above).

For all of the items in the Lombrozo et al. instrument, participants were asked to rate their degree of agreement with each statement on a five-point Likert scale. We tested whether these raw scores met assumptions of normality and whether subsequent summary statistics were appropriate. A Kolmogorov-Smirnov test of the 65 items in the Lombrozo et al. NOS instrument indicated that none of the 65 items were distributed normally ($2.708 < Z < 5.926, p < .001$, for all items). Thus, calculating item means, which typically are used to determine an item's difficulty using CTT methods, is inappropriate (see Methods). Therefore, we employed Rasch modeling in subsequent analyses. Note that items 48 and 49 were not included because they did not clearly fit into the constructs of NOS or SI (as discussed above).

As shown above, Lombrozo et al.'s items could in most cases be mapped onto NOS and SI constructs delineated in the literature (see above). Therefore, we fitted a two-dimensional model to the dataset. In order to investigate the internal structure of the instrument, we also calculated a one-dimensional model and a 13 dimensional model. The one-dimensional model examined whether one latent trait characterises the items, whereas the 13-dimensional model examined whether Lombrozo's original structure (12+1 themes) is supported empirically. All models were examined for fit, and the final deviance values for the models were compared (see Methods).

In addition to examining dimensionality, we investigated which of two possible scale models (partial credit and rating scale) best fit the dataset. Table 2 shows that the partial credit models have a lower deviance than the rating scale models, indicating a better fit. Additionally, the 13 dimensional models had convergence problems, suggesting problematic structure (for details on convergence see Wu et al., 2007). The reason for such convergence problems might the relatively small sample size. As the uni-dimensional model is a submodel of the two-dimensional model, which again is a submodel of the 13-dimensional model, the differences among the deviance values for these PC models represents a chi-square distribution with two or 89 degrees of freedom, respectively (i.e. df is the difference between the number of free parameters). Within the χ^2-distribution for df = 2, the critical values are 5.99 (p = 0.05) and 9.21 (p = 0.01). Given the difference of 82, and convergence problems with the 13 dimensional models, we conclude that the two dimensional model fits the data significantly better than the one-dimensional model. Since the difference in deviance is larger when comparing the two-dimensional with all other models (2-dim PC vs. 1-dim RS: $\chi^2[177]=831$, $p<.001$; 2-dim PC vs. 2-dim RS: $\chi^2[272]=716$, $p<.001$), the two-dimensional model appears to be the best of these six possible models (Table 2). Concerning information criteria (see Methods), the two-dimensional models are preferred to the unidimensional models (13-dimensional models were not included in our comparisons because of convergence problems). However, while the AIC indicates that the PC model is superior to the RS model, the BIC indicates the opposite. Subsequent analyses used the two-dimensional partial credit model because: (1) deviance statistics are the only results that provide *statistical* justification for model choice; (2) AIC, which takes into account both deviance and the number of parameters, likewise favors the two-dimensional PC model; and (3) BIC findings are questionable given that this method is often used for much larger samples (we think that 214 students is at the lower limit of what is considered to be a large sample). Within the two-dimensional partial credit model, the correlations between these two latent variables was 0.95.

[INSERT TABLE 2 ABOUT HERE]

RELIABILITY

Within the Standards for Educational and Psychological Testing (AERA et al., 2004) three methodologies may be used to address the issue of reliability in separate but related ways: Classical Test Theory (CTT), Generalisability Theory, and Item Response Theory (IRT). We used CTT and IRT methods, employing 63 items from the Lombrozo et al. instrument on our sample of biology undergraduates. To ensure comparability with reliability measures gained by Rasch modeling, two items were not used (i.e. 48 and 49; see above). Cronbach's α for our dataset was 0.89 (n=213). Investigating NOS and SI subsets separately, revealed values of α_{NOS}=0.81 (n=213) and α_{SI}=0.82 (n=214). Cronbach's α values of 0.7 to 0.8 are considered indicative of sufficient reliability (Field, 2009). However, there are complexities concerning these cut-off values (for an overview, see Field, 2009). For example, Field notes that Cronbach's α depends on the number of items under investigation. Increasing the number of items leads to increasing α values. Moreover, Furr and Bacharach (2008, 122) point out 'that an internal consistency estimate of a test's reliability could be high (e.g. α= 0.75) even if the test is multidimensional or conceptually heterogeneous.' Thus, while our findings may be indicative of sufficient reliability and homogeneity, such results should be treated with caution. However, Rasch analysis on the two-dimensional partial credit model produced EAP/PV reliability values of 0.85 for both subscales, NOS and SI, corroborating CTT findings that suggest sufficient reliability.

ITEM QUALITY

We used Rasch analysis to explore item difficulty, item discrimination, item redundancy, and person-item patterns within the two dimensional partial credit model. Two indices are commonly used to determine item fit levels: (1) infit/outfit and (2) standardised z values

(ZSTD). Item infit and outfit values that are less than the expected value of 1.0 are indicative of over-fitting items (see also Bond & Fox, 2001). In these cases, too little variance occurs relative to the estimated model. Item infit and outfit values larger than the expected value indicate under-fitting items. Here excessive variance occurs relative to the estimated model. Depending on the sample size per item, different acceptability intervals for infit and outfit are typically employed (e.g. Adams, Wu, & Macaskill, 1997, Bond & Fox, 2001). Moderate cutoff levels (i.e. infit/outfit acceptability values) of between 0.8 and 1.2, and ZSTD item fit values between -2 and 2, were applied to our data set. Employing these cutoff values, we could identify 6 (10%) misfitting items based on infit, and 9 (14%) misfitting items based on $ZSTD_{infit}$. Based on outfit values, 19 items (30%) were misfitting, and based on $ZSTD_{outfit}$ 20 (32%) items were misfitting. However, many items exhibited unacceptable values on more than one of those indices (see Appendix A). In addition to these four indices produced in the Rasch analysis, we also evaluated the items based on a traditional discrimination index. In this analysis 20 items (32%) were identified with insufficient discrimination (i.e. < 0.30). In summary, 29 items (46%) were identified to display unacceptable fit values (based on at least one of the above indices). Given the high percentage of misfitting items, it is reasonable to investigate item properties in greater detail to obtain information which items should be removed from the instrument.

Item fit statistics (i.e. infit/outfit and ZSTD) show which items fit the estimated model. Therefore, the number of misfitting items also indicates the quality of a model. To investigate the question of whether a two-dimensional model fits better than a uni-dimensional model, one could also investigate which model displays fewer misfitting items. Based on infit, outfit and ZSTD values we could identify 19 misfitting items regarding the unidimensional PC-model while the two-dimensional revealed 21 misfitting items. This result contradicts the findings regarding model dimensionality as discussed above. However, the number of

misfitting items is only one indicator regarding model fit. Deviance statistics, information criteria, and theoretical considerations (see above) indicate that a two-dimensional model should be preferred over a unidimensional model.

Wright Map patterns

In addition to individual item fit values, a Wright Map revealed person-item mismatch patterns (Bond & Fox, 2001). Figure 2 displays a Wright map for our sample, which visually summarises several aspects of the Rasch analysis. The distribution of persons (on the left) and items of the NOS instrument (on the right) are displayed on the same so-called logit scale. Logits are ratio-level scales. Persons at the same position (or 'height') on the scale as a particular item have a 50% chance of answering the item correctly (notice that one X represents 2.8 persons). Questions of equivalent difficulty lie at the same point on the logit scale (e.g. questions 19, 23, and 25; 10 and 54; and 1 and 41). Individuals ('persons') located above an item, however, have an even greater chance of answering the item correctly (i.e. the item is likely to be easier for such individuals). Those persons located below an item have a lower probability of being able to answer it correctly (i.e. the item is more difficult for them).

[INSERT FIGURES 2a and 2b ABOUT HERE]

Overall, figure 2a demonstrates that the students tend to pick answer options at or above the middle of the Likert scale. This means that students tend to agree with adequate NOS and SI statements. Figure 2b reveals this result in more detail: it shows the distribution of each answer option per item. Most students tend to choose answer options 4 and 5 (agree and strongly agree). This means that the majority of students display an adequate view on NOS and SI; however, the instrument fails to distinguish students at the higher end of the likert scale. Additionally, the Wright map (Figure 2a) reveals that items within the same theme are

not of comparable difficulty. For example, items 1 through 5 ('theory support') are spread over a large range of difficulty levels.

Although we have identified many shortcomings of the Lombrozo et al. NOS instrument, our findings also provide numerous suggestions for improving it. In terms of construct delineation, some NOS and SI aspects are overrepresented while other aspects are underrepresented. The Wright Map revealed person-item mismatch; instrument items are distributed towards the low end of the logit scale whereas persons are distributed at the high end of the scale. The elimination of numerous redundant items, or those items that do not appreciably contribute to person discrimination at the low end of the person-item scale, may decrease instrument length while potentially improving quality. We therefore selected 23 items proportionally distributed across all construct elements (NOS and SI; see Appendix 2). When selecting these items, we also took into account the items' distribution on the Wright map. Based on the dataset at hand, we performed a series of parallel analyses as those conducted on the original instrument in order to determine if this reduced item set could improve the quality of the Lombrozo et al. (2008) NOS instrument.

Reduced item set results

As in our analysis of the original Lombrozo et al. instrument, for the reduced item set, we used Rasch analysis to analyse three interrelated instrument properties: (1) model fit and dimensionality; (2) reliability, and (3) item quality. First, we fitted a one and two-dimensional model to the reduced item dataset and compared the final deviance values for the partial credit and rating scale analyses. Values reported in Table 3 demonstrate that the partial credit model fits better than the rating scale model (2-dim PC vs. 1-dim RS: $\chi^2[65]=223$, $p<.001$; 2-dim PC vs. 2-dim RS: $\chi^2[60]=167$, $p<.001$). Given the difference of 23 units in the deviance between the one and two-dimensional partial credit models, we conclude--as with the full item dataset--that the two-dimensional model fits the reduced dataset significantly better than the one-

dimensional model (χ^2 [2] critical values are 5.99, p = 0.05, and 9.21, p = 0.01, respectively). Information criteria AIC and BIC reveal the same picture as for the original item set: two-dimensions should be preferred over one dimension, and the RS should be preferred over the PC. Similar to the analysis of the original item set, we used the two-dimensional partial credit model in further analyses.

[INSERT TABLE 3 ABOUT HERE]

We used both CTT and IRT methods to calculate reliability measures on the reduced item set. Cronbach's α for all items in the reduced dataset was 0.62. For the SI items $\alpha_{NOS} = 0.35$, and for the NOS items $\alpha_{SI} = 0.54$. Rasch analysis produced EAP/PV reliability values of 0.645 for the NOS items, and 0.658 for the SI items. Because we substantially reduced the number of items, it is not surprising that reliability measures decreased. One solution to the reliability problems that we document is to increase the number of high-quality items. It is important to note, however, that the reduction of the number of items in high quality instruments often does not result in a substantial reduction of α. Thus, comparing α values for both the reduced and original item sets reveals that a thorough revision of the instrument is warranted.

[INSERT TABLE 4 ABOUT HERE]

We used infit/outfit and standardised z values (ZSTD) to determine item fit levels in the reduced item dataset. As shown in Table 3 and the Appendix, for the reduced item analysis of fit patterns, far fewer items displayed poor fit than the original item set. Specifically, 0% of the reduced dataset items had unacceptable fit values using four different item quality measures (infit, ZSTD infit, outfit, and ZSTD outfit). Approximately 48% of items in the reduced dataset displayed poor discrimination values (i.e. < 0.3), in contrast to 32% in the original dataset. Overall, 46% of items showed at least one unacceptable fit value in contrast to 48% of items in the reduced dataset showing unacceptable discrimination values. Thus,

Submitted 06-Oct-2014, Version 1

tsed-00637-2013.R1 Revised Version 29-Oct-2013

significantly more items were characterised by acceptable fit values in the reduced dataset relative to the original dataset (Table 4). In both cases the percentage of items characterised by low discrimination values is high, and within the reduced item set it is higher due to a smaller number of items. Finally, as in our analysis of the original Lombrozo et al. item set, a Wright Map was generated to examine person-item distribution patterns. Figure 3 illustrates better item-sample matching in the reduced dataset relative to the original dataset, as well as significantly fewer redundant items. Thus, the reduced item set demonstrated some improvement over the original dataset in our sample. Nevertheless, it is clear that the instrument retains items poorly matched to our sample.

[INSERT FIGURE 3 ABOUT HERE]

DISCUSSION

Lombrozo, Thanukos, and Weisberg (2008) recently produced an important contribution to science education research by developing a closed-response Likert-type instrument for quantifying Nature of Science (NOS) knowledge in undergraduate students. They employed this new instrument in a series of tests to determine the relationships among NOS knowledge and evolutionary beliefs. Many fundamental psychometric aspects of their instrument--including validity and reliability--were not reported. Furthermore, while the instrument was piloted on a sample of psychology students, it would be very useful to know if this instrument would also be appropriate for use in samples of biology majors studying evolution. Additionally, exploring the quality of the instrument and its constituent items helps to determine the robustness of the conclusions presented in the original study. Finally, analyses of instrument quality using Item Response Theory methods (e.g. Rasch modeling) may serve as a useful case study for other science education researchers interested in applying this important and increasingly used methodology to evaluate and/or improve instruments in existence or to develop new ones.

Our study explored three interrelated aspects of Lombrozo et al.'s new NOS Likert-scale instrument: (1) validity and reliability; (2) instrument dimensionality; and (3) item scales, properties, and qualities. In addition to CTT methods, we used Item Response Theory (specifically Rasch modeling) to explore these issues. Overall, the Lombrozo dataset demonstrated good fit with the assumptions of the Rasch model. Nevertheless, a detailed analysis of items indicated that the instrument is inappropriately matched in difficulty level to the sample of biology majors studied here. CTT (i.e. discrimination) and IRT item fit statistics also indicated that only 54% of Lombrozo et al.'s (2008) items were characterised by acceptable fit values. Numerous items of redundant difficulty were also present at the low end of the scale and there was a lack of items of sufficient difficulty to distinguish high performers. Finally, items within some NOS themes did not display comparable difficulty levels.

NOS and SI constructs

Given the central importance of Nature of Science and Scientific Inquiry in science teaching and learning (Abd-El-Khalick, 2006; Clough, 2006; Lederman 2007; Lombrozo et al., 2008, NRC, 1998, 2000; Scharmann & Harris, 1992; Scharmann et al., 2005) and their central role in national and international science standards (McComas & Olson, 1998), empirical support for the theoretical delineation of these constructs is of considerable significance (AERA, APA, & NCME, 2004). Our literature review documented how the constructs of NOS and SI are currently operationalised by the science education community (Lederman et al., 2002; Lederman, 2006; Lederman, 2007; McComas & Olson, 1998; Osborne et al., 2003; Schwartz et al., 2008). Comparing the construct elements of NOS introduced by Lombrozo et al. (2008) to the results of this review were used to theoretically ground their construct, identify discrepancies and omissions, and produce a reduced item set concordant with current views of NOS. This approach was necessary, as Lombrozo et al. did not provide theoretical

tsed-00637-2014.R1 Revision 29-Oct-2014

justification or support for their construct dimensions (contra the *Standards*, AERA et al., 2004); there was no other reference point from which we could evaluate the construct validity of their instrument.

One of the findings of broad significance to the science education community was that NOS and SI do in fact appear to hold up as two separate dimensions, as suggested in the theoretical literature (Lederman, 2006; Schwartz et al., 2008). Our Rasch analysis of the original Lombrozo et al. (2008) instrument--as well as our reduced item set--indicated that a two-dimensional model fits significantly better than a one-dimensional model (regarding deviance comparisons). Information criteria also support this finding, while the number of misfitting items indicates that a unidimensional model might be preferred (for the original item set). It is important that further work attempting to corroborate the separation of NOS and SI employ additional instruments and samples; it is possible that our instrument, sample, or methods may not extrapolate to other contexts. Nevertheless, our findings provide empirical support for the theoretical arguments of Schwartz et al. (2008).

Methodological issues with Likert-type instruments

Our Rasch-based analyses of the items from the original and the reduced-item Lombrozo et al. (2008) instrument provide important lessons about Likert-scale items for science educators developing or evaluating other instruments of this type. First, the raw Likert-type scores did not meet assumptions of normality, prohibiting the calculation of item means. Second, Rasch analyses indicated that the Likert-type responses confirmed the assumption that different measurement scales characterised the different items in the Lombrozo et al. instrument. That is, each pair of adjacent Likert categories (e.g. strongly agree, agree) was not equidistant across all instrument items. Thus, the Partial Credit Model provided better fit than the Rating Scale Model (see Methods). Our study also provided examples of how to evaluate Likert-type items and instruments for these important assumptions and generate ratio scores that can be

appropriately interpreted. Wright map and fit indices revealed problematic, redundant, and misfitting items, which could be clearly identified, excluded, or modified. The approaches that we illustrated and discussed may be effectively employed in evaluations of other Likert-type instruments in order to strengthen the quality of science education research.

Study Limitations

Content validity is one of the most important aspects of measurement instruments (AERA et al., 2004). Despite clear standards and methods for grounding instrument content and items theoretically, the Lombrozo et al. (2008) instrument--and many other instruments used in science education--have not explicitly revealed how such grounding was made. In our study of instrument evaluation--employing Rasch methods and using Lombrozo et al.'s instrument as an exemplar--this presented a significant challenge: From what standpoint should this instrument be evaluated given its lack of explicit and precise theoretical grounding? Our decision to evaluate Lombrozo et al.'s 'NOS' instrument relative to the current consensus and operationalisation of NOS and SI by the science education community (see Methods) followed approaches recommended in the *Standards* (AERA et al., 2004). However, other approaches could have been used, such as assuming that the 12 or 13 NOS 'themes' identified by Lombrozo et al.--but not supported by any literature or theoretical arguments--in fact existed. Alternatively, we could have evaluated the instrument relative to the three collapsed 'NOS' themes (generated by a problematic factor analysis and again not supported by any literature or theoretical arguments). Thus, after considering many evaluation options, we decided that the most rigorous approach was to align our evaluation most closely to a theoretical grounding as recommended by the *Standards*. But other evaluation options are possible and our study conclusions must be viewed through the lens of our evaluation approach.

Although our study addressed many aspects of Likert-type instrument evaluation, we did not explore the important issue of external construct validation (AERA et al., 2004). In order to empirically ensure that the instrument is measuring the construct under investigation it should be compared with another already validated instrument measuring the same construct. Perhaps the most fruitful way for doing this would be employing an additional open response questionnaire, such as one of Lederman et al.'s (2002) VNOS instruments. Such approaches may be used to externally validate Likert-type instruments and explore the limitations of closed-response instrument formats (Authors, 2008, 2010). Participant interviews should also be performed in order to externally validate the construct dimensions and explore item clarity and bias (AERA et al., 2004). Within the scope of such an external validation the influence of effects introducing bias (e.g., social desirability bias) need to be investigated in greater detail. Moreover, the suitability of the instrument to detect a change in students' views should be investigated as one aspect of validity, which again, requires the consideration of instructional factors that may interfere with measurement (e.g., Rosenthal & Jacobson [1992] and Adair, [1984]). Overall, several additional analyses remain to be performed before the instrument in question can be used to investigate students' views of NOS, not to mention change in students' views of NOS.

Future Directions

Our exploration of the quality of a Likert-type NOS instrument is by no means a comprehensive one (AERA et al., 1999). Indeed, the *Standards* outline a series of additional analyses--such as convergent and discriminant validity--that provide crucial information about instrument properties (but are commonly ignored or misunderstood; see Authors 2010). Nevertheless, our limited analyses have produced a wealth of information for the developers and evaluators of Likert-type instruments in science education in general and Lombrozo et al.'s NOS instrument in particular. In terms of the latter, our study has provided a series of

recommendations for improving this instrument and evaluating the efficacy of such putative improvements using Rasch methods. First, more explicit and precise theoretical grounding of Lombrozo et al.'s (2008) instrument is required--this grounding could also include broadening the working definitions of NOS and SI for undergraduate science majors; second, a more balanced distribution of items across construct elements is needed; third, more items, which elicit a more diverse range of agreements, must be added to the instrument; fourth, numerous redundant items need to be removed as they offer little to the discrimination of respondents; and fifth, convergent and discriminant validity should be investigated. In summary, closed-response instruments suitable for large samples of undergraduate science students would be of great value to the field of science education. But we must ensure that such instruments are valid and reliable tools prior to their use.

Conclusions

(1) A construct validity analysis of Lombrozo et al.'s (2008) 'NOS' instrument revealed that item distributions across NOS and SI aspects were very unbalanced; two items could not be unambiguously assigned to any of the NOS or SI aspects noted in the literature; the NOS aspect 'Distinction between data and evidence' was not addressed; Lombrozo et al.'s 'Limits' theme is not appear to be separate from the main construct of NOS as currently operationalised by the science education community; and not all items within Lombrozo et al.'s themes hold together as a single aspect of NOS (or SI).

(2) NOS and SI hold up as two separate dimensions in Rasch analyses, corroborating theoretical distinctions in the literature (Lederman, 2006; Schwartz et al., 2008). Our analysis of the original Lombrozo et al. (2008) instrument--as well as our reduced item set--indicated that a two-dimensional model fits significantly better than a one-dimensional model. To our knowledge, our study is the first to empirically validate the recent theoretical separation of these two construct dimensions (Schwartz et al., 2008).

(3) Item quality measures revealed that 33% of the items in the original Lombrozo et al. instrument were discordant with model predictions based on both MNSQ and ZSTD values. Eight additional items displayed poor discrimination values (< 0.3). Overall, CTT and IRT item fit values indicated that only 54% of Lombrozo et al.'s (2008) items were characterised by acceptable fit values.

(4) A Wright Map demonstrates that the instrument item properties were poorly matched to the person abilities in the sample. Specifically, items are overwhelmingly distributed towards the low end of the logit scale whereas persons are distributed at the high end of the scale. Items of high difficulty are sparse at the high end of the scale and an abundance of redundant items occur at the low end of the score.

(5) Our attempts to improve the instrument were marginally successful. Using the Wright map, for example, we were able to identify redundant items and reduce the item set. Analysis of the reduced item set produced improved IRT fit indices relative to the original item set. Item discrimination values remained problematic, however, suggesting item revisions may be necessary.

(6) Our study provided examples of how instrument developers and evaluators may use Rasch analyses to test two central attributes of Likert-type items: normality and scale equivalency. Additionally, our study demonstrated how Rasch analyses may be used to transform Likert-derived scores into ratio scores, thereby permitting appropriate comparisons across different Likert items. While complex, such analyses are necessary in order to ensure that the data and inferences derived from Likert-type instruments are accurate and meaningful.

Acknowledgements

We thank _____, _____, and _____ for item placement assistance; ____ and _____for helpful reviews of the manuscript; and _____ for funding portions of this study.

REFERENCES

Abd-El-Khalick, F. (2006). Over and over again: College students' views of nature of science. In L. B. Flick & N. G. Lederman (Eds.) Scientific Inquiry and Nature of Science (pp. 389-425). Dordrecht, The Netherlands: Springer.

Adair, G. (1984). The Hawthorne effect: A reconsideration of the methodological artifact. *Journal of Applied Psychology*, 69*(2)*, 334-345.

Adams, R.J., Wu, M.L., & Macaskill, G. (1997). Scaling methodology and procedures for the mathematics and science scales. In M.O. Martin & D. L. Kelly (Eds.) TIMSS technical report, Volume II: Implementation and analysis (pp. 111-145). Chestnut Hill, MA: Boston College.

Alters, B. J. (1997a). Whose Nature of Science? *Journal of Research in Science Teaching*, 34, 39-55.

Alters, B. J. (1997b). Nature of Science: A diversity of uniformity of ideas? *Journal of Research in Science Teaching*, 34, 1105-1108.

American Association for the Advancement of Science [AAAS] Project 2061 (1993). *Benchmarks. For Science Literacy*. New York: Oxford University Press.

American Educational Research Association [AERA], American Psychological Association [APA], & National Council on Measurement in Education [NCME] (2004). *Standards for educational and psychological testing*. Washington, DC, American Educational Research Association.

Authors (2006). *Bioscience*.

Authors (2008). *Journal of Research in Science Teaching*.

Authors (2010). *Journal of Research in Science Teaching*.

Best, J. W., & Kahn, J. V. (2003). *Research in Education*. Boston, MA, Allyn and Bacon.

Bezzi, A. (1999). What is this Thing Called Geoscience? Epistemological Dimensions Elicited with the Repertory Grid and their Implications for Scientific Literacy. *Science Education*, 83, 675–700.

Bode, R. K., & Wright, B. D. (1999). Rasch measurement in higher education. In Smart, J.C., & Tierney, W.G. (Eds.) Handbook of Theory and Research, Volume XIV. New York, NY: Agathon Press.

Bond, T.G. & Fox, C.M. (2001). *Applying the Rasch Model: Fundamental Measurement in the Human Sciences*. Mahwah NJ: Lawrence Erlbaum Associates.

Boone, W.J., & Scantlebury, K. (2006). The role of Rasch analysis when conducting science education research utilizing multiple-choice tests. *Science Education*, 90, 253-269.

Bybee, R. W. (1997). Toward an Understanding of Scientific Literacy. In Gräber, W., & Bolte, C. (Eds.) Scientific Literacy (pp. 37-68). Kiel, Germany, Institut für Pädagogik der Naturwissenschaften.

Clough, M. P. (2006). Learners' responses to the demands of conceptual change: Considerations for effective nature of science instruction. *Science Education*, 15, 463-494.

Coulson, R. (1992). Development of an instrument for measuring attitudes of early childhood educators towards science. *Research in Science Education*, 22, 101-105.

Driver, R., Leach, J., Millar, R., & Scott, P (1996). *Young People's Images of Science*. Buckingham, Open University Press.

Edwards, M. C. & Wirth, R. J. (2009). Measurement and the study of change. *Research in Human Development*, 6(2-3), 74-96.

Field, A. (2009). *Discovering statistics using SPSS*. Thousand Oaks, CA, Sage.

Fleming, R. (1988). Undergraduate Science Students' Views on the Relationship between Science, Technology and Society. *International Journal of Science Education*, 10, 449–463.

Furr, R. M., & Bacharach, V. R. (2008). *Psychometrics*. Thousand Oaks, CA, Sage.

Gilbert, S. W. (1991). Model Building and a Definition of Science. *Journal of Research in Science Teaching*, 28, 73–80.

Glynn, S.M., Taasoobshirazi, G., and Brickman, P. (2009). Science Motivation Questionnaire: Construct Validation With Nonscience Majors. Journal of Research in Science Teaching. 46, 2, 127–146.

Haertel, E. H. (2006). Reliability. In Brennan, R. L. (Ed.) Educational Measurement (65-110). Westport, CT, American Council on Education [ACE] and Praeger.

Kaya, N., Yager, R., & Dogan, A. (2009). Changes in Attitudes Towards Science–Technology–Society of Pre-service Science Teachers. *Research in Science Education*, 39, 257-279.

Lederman, N. G., Abd-El Khalick, F., Bell, R. L., & Schwartz, R. (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*, 30, 497-521.

Lederman, N. G. (2006). Syntax of Nature of Science within Inquiry and Science Instruction. In L. B. Flick, & N. G. Lederman (Eds.) Scientific Inquiry and Nature of Science (301-317). Dordrecht, The Netherlands, Springer.

Lederman, N. G. (2007). Nature of Science: Past, Present, and Future. In Abell, S. K., & Lederman, N. G. (Eds.) Handbook of research on science education (831-879). Mahwah, NJ, Lawrence Erlbaum.

Liang L. L., Chen S., Chen X., Kaya O. N., Adams A. D., Macklin M., & Ebenezer J. (2006). *Student understanding of science and scientific inquiry: revision and further validation of an assessment instrument*. Paper presented at the Annual Conference of the National Association for Research in Science Teaching (NARST), San Francisco, CA.

Liu, X., & Boone, W.J. (2006). *Applications of Rasch Measurement in Science Education*. Maple Grove, MN, JAM Press.

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.

McComas, W. F., & Olson, J. K. (1998). The Nature of Science in International Science Education Standards Documents. In McComas, W. F. (Ed.) The Nature of Science in Science Education: Rationales and Strategies (pp. 41-52). Dordrecht, The Netherlands, Kluwer Academic Publishers.

National Research Council [NRC] (1998). *National science education standards*. Washington, DC: National Academic Press.

National Research Council [NRC] (2000). *Inquiry and the National science education standards*. Washington, DC: National Academic 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, 692-720.

Rosenthal, R. & Jacobson, L. (1992). *Pygmalion in the classroom*. New York: Irvington.

Rubba, P. A., & Andersen, H. (1978). Development of an instrument to assess secondary school students' understanding of the nature of scientific knowledge. *Science Education*, 62(4), 449–458.

Rutledge, M. L., & Warden, M. A. (2000). Science and high school biology teachers: Critical relationships. *The American Biology Teacher*, 62, 23-31.

Ryder, J., Leach, J., & Driver, R. (1999). Undergraduate Science Students' Images of Science. *Journal of Research in Science Teaching*, 36, 201–219.

Scharmann, L. C., & Harris, W. M. (1992). Teaching Evolution: Understanding and Applying the Nature of Science. *Journal of Research in Science Teaching*, 29, 375-388.

Scharmann, L.C., Smith, M.U., James, M., & Jensen, M. (2005). Explicit Reflective Nature of Science Instruction: Evolution, Intelligent Design, and Umbrellaology. *Journal of Science Teacher Education*, 16, 27-41.

Schwartz, R., Lederman, N. G., & Lederman, J. S. (2008). *An Instrument To Assess Views Of Scientific Inquiry: The VOSI Questionnaire*. Paper presented at the annual meeting of the National Association for Research in Science Teaching, Baltimore, MD.

Smith, M. U., Lederman, N. G., Bell, R. L., McComas, W. F., & Clough, M. P. (1997). How great is the disagreement about the Nature of Science: A response to Alters. *Journal of Research in Science Teaching*, 34, 1101-1103.

Southerland, S. A., Settlage, J., Johnston, A., Scuderi, A., & Meadows, L. (2003). Development and application of a web-based NOS instrument: Making students aware of their NOS conceptions. Paper presented at the annual meeting of National Association for Research in Science Teaching International Conference, Philadelphia, PA.

Wilson, L. (1954). A study of opinions related to the nature of science and its purpose in society. *Science Education*, 38(2), 159–164.

Wilson, M. R. (2004). *Constructing Measures*. Mahwah, NJ, Lawrence Erlbaum.

Wilson, M., deBoek, P. & Carstensen, C. H. (2008). Explanatory item response models: a brief introduction. In J. Hartig, E. Klieme & D. Leutner (Eds.). Assessment of competencies in educational contexts (p. 91-120). Hogrefe: Cambridge, MA.

Wu, M. L., Adams, R. J., & Wilson, M. R. (2007). *ACER ConQuest version 2.0: generalised item response modelling software*. Camberwell, Vic., ACER Press.

Yen, W. M., & Fitzpatrick, A. R. (2006). Item Response Theory. In Brennan, R. L. (Ed.) Educational Measurement (65-110). Westport, CT, American Council on Education [ACE] and Praeger.

Zoanetti, N., Griffin, P., & Adams, R. (2006). *Applications of item response theory to identify and account for suspect rater data*. Retrieved online: http://www.aare.edu.au/06pap/zoa06310.pdf, Nov. 22nd, 2009, 18:34 EST.

Table 1. Construct analysis, delineation, and assignment. Aspects of the nature of science and scientific inquiry are abbreviations based on the cited studies.

Construct Element	Items from Lombrozo et al. (2008)	Total items
NOS (cf. Lederman et al., 2002)		
Empirical Nature	1, 3, 12, 13, 14, 53, 61, 62, 63, 64, 65	11
Observation vs. Inference	4, 5	2
Theory vs. Law	16, 17, 18, 19, 20	5
Creativity	11, 36, 37, 38, 39, 40	6
Theory-Ladenness/ Subjectivity	2	1
Socio-Cultural Embeddedness	21, 22, 23, 24, 25, 33, 34	7
Tentative Nature	6, 7, 8, 9, 10, 15, 47, 50, 58	9
SI (cf. Schwartz et al., 2008)		
Scientific questions guide investigations	46	1
Multiple methods of sc. investigations	41, 42, 43, 44, 45, 51, 53, 54, 55, 59	10
Multiple purposes of scientific investigations	31, 32, 35	3
Justification of scientific knowledge	28, 56, 60	3
Recognition and handling of anomalous data	57	1
Distinctions between data and evidence	---	0
Community of practice	26, 27, 29, 30	4
Extraneous items	48, 49	2

Table 2. Item dimensionality test results using a Rasch analysis of all 63 items. *convergence problems.

	Rating Scale		Partial Credit	
	deviance (# free parameters)	AIC BIC	deviance (# free parameters)	AIC BIC
1-dimensional model	30554 (67)	30688 30914	29805 (242)	30289 31104
2-dimensional model	30439 (72)	30583 30825	29723 (244)	30211 31032
13-dimensional model	29738* (193)	30124 30774	29249* (333)	29915 31036

Table 3. Reduced item (n = 23) dimensionality test results using a Rasch analysis.

	Rating Scale		Partial Credit	
	deviance (# free parameters)	AIC BIC	deviance (# free parameters)	AIC BIC
1-dimensional model	12186 (27)	12240 12331	11986 (90)	12166 12469
2-dimensional model	12139 (32)	12203 12311	11963 (92)	12147 12457

Table 4. Comparison of item quality statistics for the original and reduced item sets.

Item analysis variable	**Original (63 Items; 2-d, PC)**	**Reduced (23 items; 2-d, PC)**
α all items	0.892 (n_I=63, n_S=213)	0.623 (n_I=23, n_S=214)
α NOS items	0.811 (n_I=41, n_S=213)	0.350 (n_I=13, n_S=214)
α SI items	0.824 (n_I=22, n_S=214)	0.543 (n_I=10, n_S=214)
EAP/PV NOS items	0.852 (n_I=41, n_S=214)	0.645 (n_I=13, n_S=214)
EAP/PV SI items	0.854 (n_I=22, n_S=214)	0.658 (n_I=10, n_S=214)
% of misfitting items (infit)	10	0
% of misfitting items ($ZSTD_{infit}$)	14	0
% of misfitting items (outfit)	30	0
% of misfitting items ($ZSTD_{outfit}$)	32	0
% of misfitting items (discrimination)	32	48

redundant items (Wright map)	many	less
item-sample match	poor	better

Figure 1. Nature of Science and Scientific Inquiry construct delineation in Lombrozo et al. (2008) on the left, and current conceptualisations from the literature on the right. See text for sources and methods.

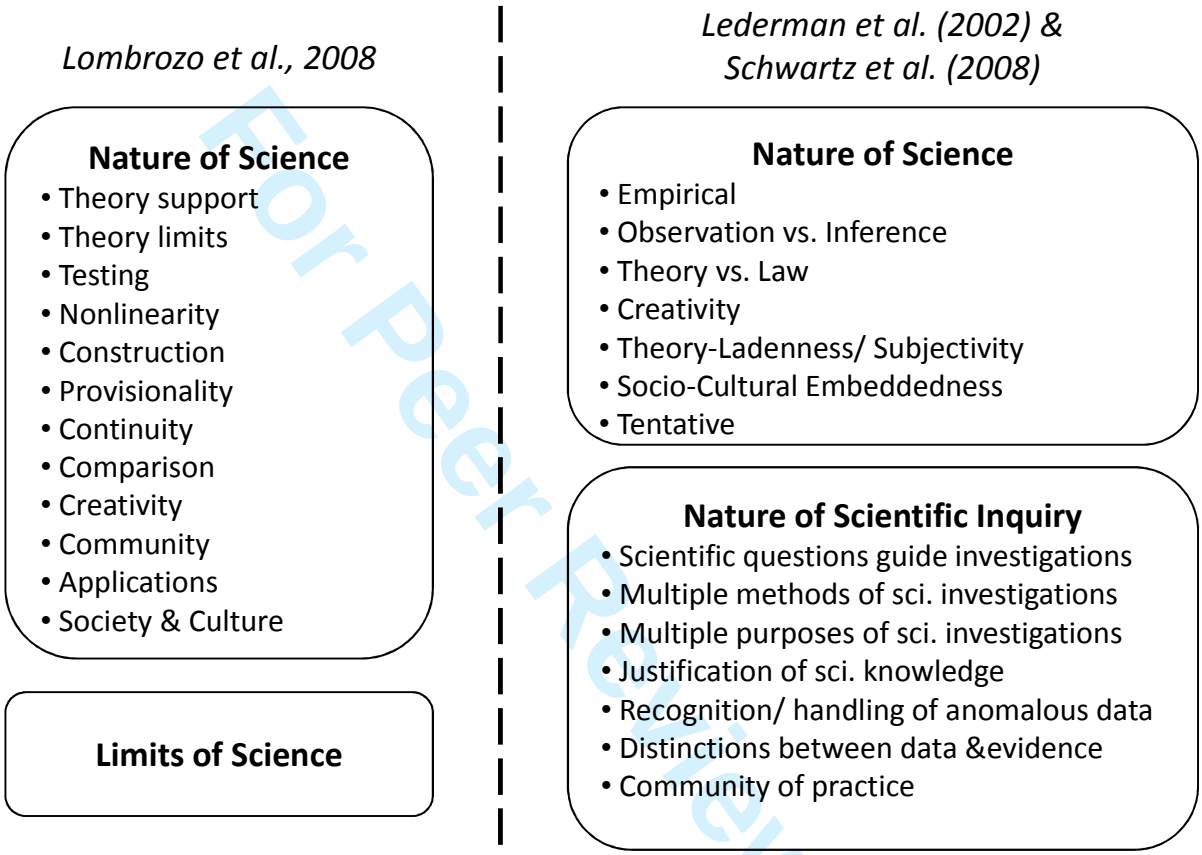


Figure 2a. Wright Map of two-dimensional analysis of the original Lombrozo et al. (2008) instrument (n = 63 items). Individuals or persons in the analysis are displayed on the left and instrument items are displayed on the right. Each X = 2.8 individuals in the sample. See Appendix for item numbers and additional data.

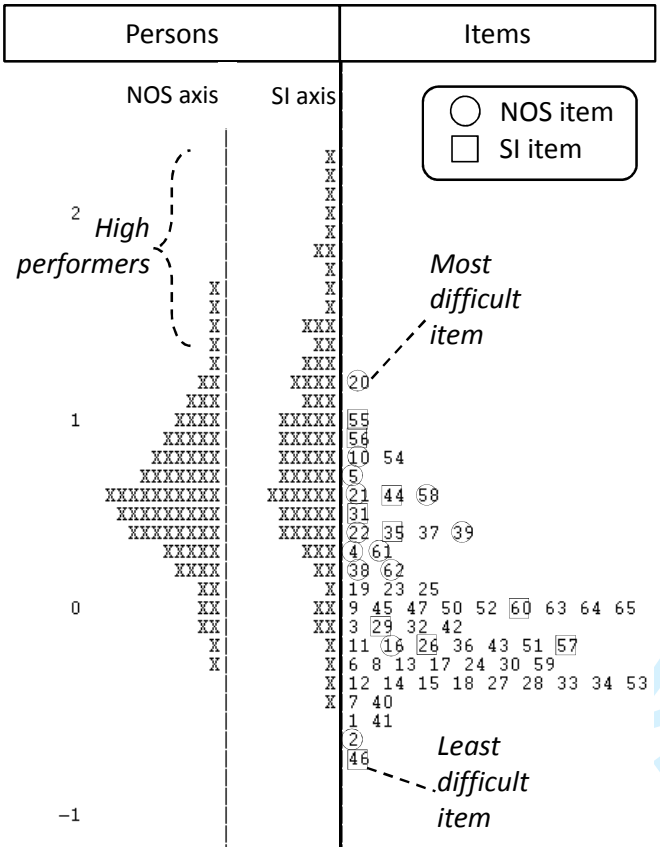


Figure 2b. Wright Map of item thresholds for the original Lombrozo et al. (2008) item set (n = 63 items). The number behind each item number indicates the option of the Likert scale. Each X indicates 2.6 students.

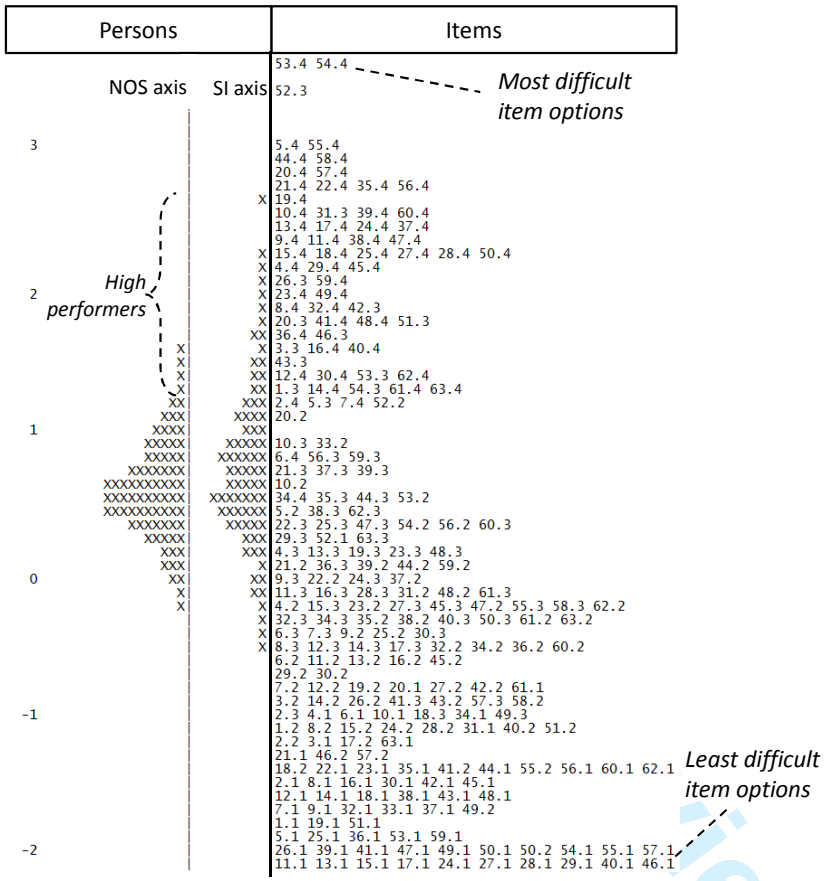
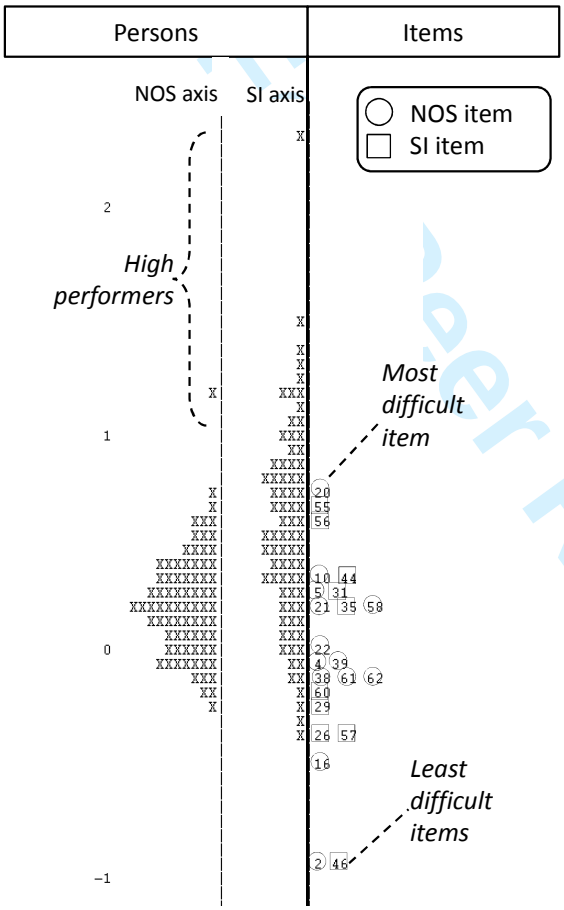


Figure 3. Wright Map of two-dimensional analysis of the reduced item set (n = 23 items). Individuals or persons in the analysis are displayed on the left and instrument items are displayed on the right. Each X = 2.7 individuals in the sample. See Appendix for item numbers and additional data.



Appendix A: Detailed Item properties on the original dataset

Item #	Theme (Lombrozo)	Aspect (NOS/SI)[a]	Construct	difficulty (error)	discriminability	infit (ZSTD)	outfit (ZSTD)
1	Theory support	emp	NOS	-0.547 (0.057)	0.56	0.86 (-1.2)	0.83 (-1.8)
2	Theory support	sub	NOS	-0.638 (0.057)	0.45	0.91 (-0.5)	0.90 (-1.1)
3	Theory support	emp	NOS	-0.128 (0.057)	0.47	0.90 (-0.7)	0.89 (-1.2)
4	Theory support	obs	NOS	0.264 (0.049)	0.32	1.02 (0.3)	1.12 (1.2)
5	Theory support	obs	NOS	0.700 (0.050)	0.06	1.21 (2.6)	1.24 (2.3)
6	Theory limits	ten	NOS	-0.271 (0.051)	0.62	0.81 (-1.4)	0.77 (-2.5)
7	Theory limits	ten	NOS	-0.434 (0.053)	0.25	1.04 (0.4)	1.24 (2.3)
8	Theory limits	ten	NOS	-0.299 (0.055)	0.33	0.98 (-0.1)	1.06 (0.6)
9	Theory limits	ten	NOS	0.051 (0.051)	0.47	0.92 (-0.8)	0.94 (-0.6)
10	Theory limits	ten	NOS	0.764 (0.046)	0.21	1.17 (2.2)	1.20 (2.0)
11	Testing	cre	NOS	-0.145 (0.053)	0.30	1.02 (0.2)	1.03 (0.3)
12	Testing	emp	NOS	-0.404 (0.054)	0.40	0.95 (-0.4)	0.93 (-0.7)
13	Testing	emp	NOS	-0.244 (0.052)	0.29	1.02 (0.3)	1.02 (0.3)
14	Testing	emp	NOS	-0.409 (0.054)	0.57	0.84 (-1.3)	0.76 (-2.6)
15	Testing	tent	NOS	-0.339 (0.055)	0.46	0.91 (-0.8)	0.90 (-1.0)
16	Nonlinearity	the	NOS	-0.149 (0.051)	0.49	0.90 (-0.8)	0.90 (-1.0)
17	Nonlinearity	the	NOS	-0.324 (0.057)	0.30	0.99 (0.0)	1.02 (0.3)
18	Nonlinearity	the	NOS	-0.400 (0.059)	0.55	0.87 (-0.7)	0.85 (-1.6)
19	Nonlinearity	the	NOS	0.072 (0.053)	0.16	1.09 (0.8)	1.15 (1.5)
20	Nonlinearity	the	NOS	1.182 (0.051)	-0.18	1.37 (3.3)	1.44 (4.0)
21	Construction	soc	NOS	0.569 (0.048)	0.13	1.19 (2.2)	1.24 (2.4)
22	Construction	soc	NOS	0.399 (0.049)	0.07	1.23 (2.5)	1.34 (3.2)
23	Construction	soc	NOS	0.133 (0.049)	0.34	1.02 (0.2)	1.11 (1.1)
24	Construction	soc	NOS	-0.267 (0.055)	0.30	1.00 (0.0)	0.99 (-0.1)
25	Construction	soc	NOS	0.083 (0.050)	0.50	0.90 (-1.0)	0.93 (-0.7)
26	Provisionality	com	SI	-0.189 (0.076)	0.44	0.93 (-0.6)	0.94 (-0.6)
27	Provisionality	com	SI	-0.356 (0.068)	0.39	1.07 (0.7)	1.12 (1.2)
28	Provisionality	jus	SI	-0.345 (0.069)	0.41	1.02 (0.2)	1.07 (0.7)
29	Provisionality	com	SI	-0.051 (0.065)	0.51	0.98 (-0.2)	1.04 (0.4)
30	Provisionality	com	SI	-0.262 (0.066)	0.52	0.90 (-0.8)	1.10 (1.1)
31	Continuity	pur	SI	0.487 (0.070)	0.36	1.04 (0.4)	1.08 (0.8)
32	Continuity	pur	SI	-0.127 (0.065)	0.61	0.84 (-1.4)	0.89 (-1.1)
33	Continuity	soc	NOS	-0.358 (0.058)	0.49	0.89 (-1.4)	0.88 (-1.3)
34	Continuity	soc	NOS	-0.332 (0.051)	0.62	0.80 (-1.5)	0.68 (-3.7)
35	Continuity	pur	SI	0.406 (0.063)	0.40	1.14 (1.4)	1.31 (2.9)
36	Comparison	cre	NOS	-0.179 (0.050)	0.58	0.84 (-1.6)	0.84 (-1.7)
37	Comparison	cre	NOS	0.360 (0.049)	0.14	1.16 (1.9)	1.22 (2.2)
38	Comparison	cre	NOS	0.230 (0.050)	0.56	0.87 (-1.5)	0.89 (-1.2)
39	Comparison	cre	NOS	0.342 (0.049)	0.18	1.13 (1.6)	1.15 (1.5)
40	Comparison	cre	NOS	-0.500 (0.054)	0.46	0.91 (-0.8)	0.93 (-0.7)
41	Creativity	met	SI	-0.562 (0.074)	0.53	0.88 (-0.9)	0.88 (-1.3)
42	Creativity	met	SI	-0.077 (0.074)	0.59	0.81 (-1.7)	0.78 (-2.5)
43	Creativity	met	SI	-0.232 (0.073)	0.49	0.91 (-0.7)	0.90 (-1.0)
44	Creativity	met	SI	0.560 (0.060)	0.39	1.19 (2.1)	1.36 (3.4)

45	Creativity	met	SI	0.003 (0.066)	0.56	0.90 (-0.8)	0.96 (-0.4)
46	Community	que	SI	-0.744 (0.077)	0.57	0.82 (-1.9)	0.80 (-2.2)
47	Community	ten	NOS	0.038 (0.050)	0.41	0.96 (-0.4)	1.01 (0.2)
48	Community	N/A	N/A	---	---	---	
49	Community	N/A	N/A	---	---	---	
50	Community	ten	NOS	0.052 (0.048)	0.54	0.88 (-1.4)	0.91 (-1.0)
51	Applications	met	SI	-0.171 (0.076)	0.54	0.84 (-1.3)	0.81 (-2.0)
52	Applications	emp	NOS	-0.023 (0.057)	0.42	0.92 (-0.7)	0.92 (-0.8)
53	Applications	met	SI	-0.406 (0.075)	0.57	0.82 (-1.2)	0.74 (-3.0)
54	Applications	met	SI	0.739 (0.063)	0.26	1.31 (3.4)	1.39 (3.6)
55	Applications	met	SI	0.991 (0.063)	0.25	1.30 (3.3)	1.33 (3.1)
56	Society and Culture	jus	SI	0.859 (0.064)	0.32	1.22 (2.5)	1.28 (2.7)
57	Society and Culture	ano	SI	-0.174 (0.074)	0.17	1.17 (1.3)	1.23 (2.2)
58	Society and Culture	ten	NOS	0.607 (0.047)	0.25	1.11 (1.5)	1.14 (1.4)
59	Society and Culture	met	SI	-0.311 (0.079)	0.41	0.92 (-0.4)	0.88 (-1.2)
60	Society and Culture	jus	SI	-0.040 (0.320)	0.26	1.13 (1.0)	1.30 (2.8)
61	Limits of Science	emp	NOS	0.253 (0.048)	0.24	1.10 (1.3)	1.12 (1.3)
62	Limits of Science	emp	NOS	0.241 (0.051)	0.21	1.08 (0.8)	1.15 (1.5)
63	Limits of Science	emp	NOS	-0.009 (0.049)	0.22	1.13 (1.1)	1.29 (2.8)
64	Limits of Science	emp	NOS	0.033 (0.047)	0.27	1.11 (1.3)	1.17 (1.7)
65	Limits of Science	emp	NOS	0.027 (0.329)	0.24	1.12 (1.4)	1.17 (1.7)

Appendix B: Detailed item properties of the reduced dataset

Item #	Theme (Lombrozo)	Aspect (NOS/SI)[a]	Construct	difficulty (error)	discriminability	infit (ZSTD)	outfit (ZSTD)
2	Theory support	sub	NOS	-0.931 (0.063)	0.28	0.98 (-0.1)	0.98 (-0.2)
4	Theory support	obs	NOS	-0.074 (0.052)	0.36	0.97 (-0.3)	0.97 (-0.3)
5	Theory support	obs	NOS	0.295 (0.053)	0.23	1.03 (0.4)	1.02 (0.3)
10	Theory limits	ten	NOS	0.365 (0.048)	0.39	0.99 (-0.1)	0.99 (-0.1)
16	Nonlinearity	the	NOS	-0.464 (0.054)	0.42	0.95 (-0.4)	0.94 (-0.6)
20	Nonlinearity	the	NOS	0.727 (0.054)	0.00	1.1 (1.0)	1.13 (1.4)
21	Construction	soc	NOS	0.195 (0.051)	0.24	1.05 (0.6)	1.06 (0.6)
22	Construction	soc	NOS	0.044 (0.051)	0.16	1.07 (0.8)	1.10 (1.0)
26	Provisionality	com	SI	-0.366 (0.074)	0.28	0.99 (0.0)	0.99 (-0.1)
29	Provisionality	com	SI	-0.213 (0.062)	0.47	0.95 (-0.5)	0.95 (-0.4)
31	Continuity	pur	SI	0.277 (0.067)	0.28	1.03 (0.3)	1.02 (0.2)
35	Continuity	pur	SI	0.210 (0.060)	0.55	0.93 (-0.8)	0.95 (-0.5)
38	Comparison	cre	NOS	-0.124 (0.053)	0.56	0.91 (-1.1)	0.91 (-1.0)
39	Comparison	cre	NOS	-0.031 (0.052)	0.27	1.03 (0.4)	1.03 (0.3)
44	Creativity	met	SI	0.346 (0.057)	0.45	1.04 (0.5)	1.09 (0.9)
46	Community	que	SI	-0.955 (0.075)	0.42	0.91 (-0.9)	0.90 (-1.1)
55	Applications	met	SI	0.681 (0.060)	0.36	1.06 (0.8)	1.06 (0.7)
56	Society and Culture	jus	SI	0.568 (0.061)	0.43	1.03 (0.3)	1.03 (0.4)
57	Society and Culture	ano	SI	-0.350 (0.071)	0.20	1.07 (0.6)	1.09 (1.0)
58	Society and Culture	ten	NOS	0.223 (0.050)	0.39	0.98 (-0.3)	0.98 (-0.1)
60	Society and Culture	jus	SI	-0.197 (0.196)	0.22	1.07 (0.6)	1.13 (1.3)
61	Limits of Science	emp	NOS	-0.121 (0.051)	0.28	1.03 (0.4)	1.03 (0.3)
62	Limits of Science	emp	NOS	-0.105 (0.183)	0.30	1.00 (0.0)	1.02 (0.3)

[a]emp: emprical nature; obs: observation vs. inference; the: theory vs. law; cre: creativity; sub: theory-ladenness/subjectivity; soc: socio-cultural embeddedness; ten: tentative nature; que: sc. questions guide investigations; met: multiple methods of sc. investigations; pur: multiple purposes of sc. investigations; jus: justification of sc. knowledge; ano: recognition and handling of anomalous data; dat: distinction between data and evidence; com: community of practice.