The Nature of Science Instrument-Elementary (NOSI-E): Using Rasch Principles to develop a theoretically grounded scale to measure Elementary Student Understanding of the Nature of Science

Author: Shelagh M. Peoples

Persistent link: http://hdl.handle.net/2345/2925

This work is posted on eScholarship@BC, Boston College University Libraries.

Boston College Electronic Thesis or Dissertation, 2012

Copyright is held by the author, with all rights reserved, unless otherwise noted.
THE NATURE OF SCIENCE INSTRUMENT-ELEMENTARY (NOSI-E): USING RASCH PRINCIPLES TO DEVELOP A THEORETICALLY-GROUNDED SCALE TO MEASURE ELEMENTARY STUDENT UNDERSTANDING OF THE NATURE OF SCIENCE

Dissertation
by
SHELAGH PEOPLES

Submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy

December 2012
Abstract

The Nature of Science Instrument-Elementary (NOSI-E): Using Rasch principles to develop a theoretically-grounded scale to measure elementary student understanding of the nature of science.

Shelagh M. Peoples, Author
Laura M. O’Dwyer, Dissertation Chair

The purpose of this study was to determine which of three competing models will provide, reliable, interpretable, and responsive measures of elementary students’ understanding of the nature of science (NOS). The Nature of Science Instrument-Elementary (NOSI-E), a 28-item Rasch-based instrument, was used to assess students’ NOS understanding. The NOS construct was conceptualized using five construct dimensions (Empirical, Inventive, Theory-laden, Certainty and Socially & Culturally Embedded). The competing models represent three internal models for the NOS construct. One postulate is that the NOS construct is unidimensional where one latent construct explains the relationship between the 28 items of the NOSI-E. Alternatively, the NOS construct is composed of five independent unidimensional constructs (the consecutive approach). Lastly, the NOS construct is multidimensional and composed of five inter-related but separate dimensions.

A validity argument was developed that hypothesized that the internal structure of the NOS construct is best represented by the multidimensional Rasch model. Four sets of
analyses were performed in which the three representations were compared. These analyses addressed five validity aspects (content, substantive, generalizability, structural and external) of construct validity. The vast body of evidence supported the claim that the NOS construct is composed of five separate but inter-related dimensions that is best represented by the multidimensional Rasch model. The results of the multidimensional analyses indicated that the items of the five subscales were of excellent technical quality, exhibited no differential item functioning (based on gender), had an item hierarchy that conformed to theoretical expectations; and together formed subscales of reasonable reliability (> 0.7 on each subscale) that were responsive to change in the construct.

Theory-laden scores from the multidimensional model predicted students’ science achievement with scores from all five NOS dimensions significantly predicting students’ perceptions of the constructivist nature of their classroom learning environment. The NOSI-E instrument is a theoretically grounded scale that can measure elementary students’ NOS understanding and appears suitable for use in science education research.
**Acknowledgements**

It is hard to express my gratitude to Dr. Laura O’Dwyer, my dissertation chair. Any words are completely inadequate. Since I began taking classes as a non-degreed student at the Lynch School of Education’s ERME department in 2006, Dr. O’Dwyer encouraged me to do more. The constancy of her support and her confidence that I was capable of completing the doctoral program provided me with the determination to persevere. My appreciation of Dr. O’Dwyer’s faith in me is heartfelt. What I have learned from her is immeasurable and will stay with me as I enter the next phase of my life.

I would like to thank my readers, Dr. Henry Braun and Dr. Katherine McNeill for their guidance and feedback. The questions and commentary I received from them pushed me to re-evaluate, to think and to consolidate my understanding. Thank you for making me think and pushing me to do better. I would like to also thank Dr. Larry Ludlow for his invaluable feedback on my dissertation proposal.

Finally, I would not have managed to complete my degree and dissertation without the sustained encouragement of my husband, Olly, and of my four daughters, Maggie, Fiona, Kathleen and Kelsey. I love you all and thank you for your enduring support.
Dedication

This dissertation is dedicated to my mother, Dulcie Menzies Dawson-Stormont (deceased) and to my father, George Rodger Stormont (deceased). Together, they inspired my lifelong love of learning and taught me to always “keep the heid”.
# Table of Contents

Acknowledgements ........................................................................................................................... i  
Dedication ........................................................................................................................................ ii  

**Chapter One: Introduction** ........................................................................................................ 1  
Background ...................................................................................................................................... 1  
Statement of the Problem ................................................................................................................. 4  
Purpose of Study .............................................................................................................................. 8  
Research Question ............................................................................................................................ 9  
Description of the Research Problem ............................................................................................ 10  
  US science education reform efforts .......................................................................................... 11  
  US student science achievement in a global knowledge-based economy ............................. 12  
  Structural cracks in the US’s vertically integrated education system .................................. 16  
  The role of the nature of science in US science education reform efforts ......................... 19  
  Measures of NOS.. ................................................................................................................ 23  
  Measurement hegemony. ........................................................................................................... 28  
Significance of Study ..................................................................................................................... 31  

**Chapter Two: Literature Review** ............................................................................................... 36  
The Nature of Science Theoretical Framework ............................................................................. 37  
  Empirical domain...................................................................................................................... 40  
  Theory-laden domain. ................................................................................................................. 41  
  Certainty domain. ...................................................................................................................... 42  
  Inventive domain. .................................................................................................................... 43  
  Socially & culturally embedded. ................................................................................................. 45  
Students’ Understanding of the Nature of Science ........................................................................ 47  
  Characterization of students’ perceptions of the NOS. ......................................................... 48  
  Student understanding of the nature of science through implicit instruction ...................... 57  
  Student understanding of the nature of science through explicit instruction ...................... 71  
  Summary of the Nature of Science Literature Review ......................................................... 86  
The Utility of the Rasch Model in Scale Construction and Science Education Research .......... 93  
The Evolution Readiness: A Modeling Approach Project .......................................................... 96  
  Development and validation of the Nature of Science Instrument-Elementary ................. 99
Chapter Three: Methodology ........................................................................................................................................................................ 122

The Rasch Family of Models ........................................................................................................................................................................ 123

The basic Rasch model ................................................................................................................................................................................ 124

The Partial Credit model for Likert-based responses ............................................................................................................................... 129

The Multi-dimensional Random Coefficients Multinomial Logit model ................................................................................................. 131

Data Sources .......................................................................................................................................................................................... 140

Sample characteristics, sample size and sample parameter estimates ...................................................................................................... 140

Hypothesis Related to Research Question 1 ............................................................................................................................................. 144

Data Analyses Procedures ........................................................................................................................................................................ 144

Premise of Validity Argument ................................................................................................................................................................. 147

Analyses 1: Content and Substantive Validity Evidence ......................................................................................................................... 147

Analyses 2: Generalizability Evidence ...................................................................................................................................................... 151

Analyses 3: Structural Validity Evidence .................................................................................................................................................. 156

Analyses 4: External Validity Evidence .................................................................................................................................................... 163

Validity Argument Conclusion ............................................................................................................................................................... 167

Conclusion ................................................................................................................................................................................................. 167

Chapter Four: Results and Discussion ................................................................................................................................................................. 169

Results from Analyses 1: Content and Substantive Validity Evidence .................................................................................................... 170

Item Technical Quality ................................................................................................................................................................................ 171

Rating Scale Functioning .............................................................................................................................................................................. 173

Expected Item Difficulty Hierarchy ........................................................................................................................................................... 182

Results from Analyses 2: Generalizability Validity Evidence .................................................................................................................. 191

Item Invariance across Models ................................................................................................................................................................. 192

Differential Item Functioning ................................................................................................................................................................. 193

Reliability of Person Estimates ................................................................................................................................................................. 197

Precision of Person Estimates ................................................................................................................................................................. 199

Results from Analyses 3: Structural Validity Evidence .......................................................................................................................... 203

Dimensionality Analyses using Rasch-based Goodness-of-Fit Statistics ................................................................................................. 204

Dimensionality Analyses using CFA Goodness of Fit Statistics ............................................................................................................. 205
List of Tables

Table 2.1: Characterization students’ perceptions of NOS ........................................................... 58
Table 2.2: Summary of NOSI-E’s Validation Activities ................................................................. 102
Table 2.3: NOSI-E Development Cycles .................................................................................... 110
Table 2.4: Item Prompts & Theoretical Domain Placement for the 28-Item NOSI-E ............. 112
Table 3.1: Profile of Samples used in the Study .......................................................................... 141
Table 4.1: Comparison of Outfit Mean Square Error (MNSQ) and Standardized T Statistics .... 174
Table 4.2: Comparison of Infit Mean Square Error (MNSQ) and Standardized T Statistics ....... 175
Table 4.3: Rating Scale Functioning: Monotonicity of Average Measures ............................... 178
Table 4.4: Rating Scale Functioning: Monotonicity of Delta Parameters ................................. 180
Table 4.5: Dimensional Ordering of Mean Population Parameters (in logits) ......................... 183
Table 4.6: Comparison of Item Difficulty Hierarchy (average deltas) ....................................... 190
Table 4.7: Mean Difference in Ability Estimates by Gender (Male – Female) ......................... 194
Table 4.8: Differential Item Functioning across Rasch Models – Gender ................................. 196
Table 4.9: Comparison of Dimension Reliabilities across Rasch Models ................................. 197
Table 4.10: Dimensionality Analyses: Rasch Model Fit Comparison ........................................ 204
Table 4.11: Dimensionality Analyses: CFA Model Fit Comparison ........................................... 207
Table 4.12: Dimensionality Analyses: CFA DIFFTEST for 1-Factor vs. 5-Factor Model .... 210
Table 4.13: CFA Correlations and Covariances between NOS dimensions (5-factor model) ... 211
Table 4.14: Rasch-based Correlations between NOS dimensions ........................................... 218
Table 4.15: Number of Person Strata (H) by Rasch Model ...................................................... 224
Table 4.16: Comparison of Two Students’ Ability Estimates (in logits) ..................................... 231
Table 4.17: Multilevel Models: Predicting Students’ Science Achievement ............................ 252
Table 4.18: Multilevel Models: Predicting Students’ Views of their Classroom Learning Environment ................................................................................................................................ 264
List of Figures

Figure 2.1: Conceptual model for the NOSI-E ................................................................. 39
Figure 2.2: Comparative theoretical representations for NOS data ................................. 121
Figure 3.1: Relationship between person’s ability, item difficulty and probability of success ... 127
Figure 3.2: Multidimensional design matrix ..................................................................... 136
Figure 3.3: Multidimensional scoring matrix ................................................................... 136
Figure 3.4: Validity argument for the NOS construct ....................................................... 146
Figure 4.1: Summary of Content & Substantive Validity Analyses ................................. 170
Figure 4.2: Summary of Generalizability Validity Analyses ........................................... 191
Figure 4.3: Item Invariance across the three Rasch models ............................................. 192
Figure 4.4: Comparison of the Precision of Ability Estimates across the Latent Ability  
Continuum ...................................................................................................................... 202
Figure 4.5: Summary of Structural Validity Analyses ..................................................... 203
Figure 4.6: Five-factor CFA Multidimensional Model for NOS construct ....................... 214
Figure 4.7: Relationship between Certainty and Empirical Person Estimates .................... 220
Figure 4.8: Relationship between Theory-laden and Inventive Person Estimates ............. 220
Figure 4.9: Summary of External Validity Analyses ......................................................... 222
Figure 4.10: Unidimensional NOSI-E Variable Threshold Map ........................................ 233
Figure 4.11: Consecutive Empirical Variable Threshold Map ......................................... 234
Figure 4.12: Consecutive Inventive Variable Threshold Map .......................................... 235
Figure 4.13: Consecutive Theory-laden Variable Threshold Map ..................................... 236
Figure 4.14: Consecutive Certainty Variable Threshold Map .......................................... 237
Figure 4.15: Consecutive Socially & Culturally Embedded Variable Threshold Map ......... 238
Figure 4.16: Multidimensional NOSI-E Variable Threshold Map ..................................... 239
Figure 4.17: Evolution Readiness Project Cohort Research Design ................................... 241
Figure 4.18: Multilevel Model Research Design ......................................................... 245
Chapter One: Introduction

Background

Since World War II the United States (US) has dominated innovation in the sciences by strategically funding basic and applied research within its tertiary level institutions and through its federal laboratories (National Academy of Sciences, National Academy of Engineering, & Institute of Medicine, [NAS et al.], 2007a); US dominance in the sciences provided the main vehicle for economic growth since the world war. US’s continued competitiveness in the global economy is heavily dependent on it maintaining its lead in science and engineering (National Academy of Sciences, National Academy of Engineering and Institute of Medicine, [NAS et al.,] 2007a; NAS et al., 2007b). At the heart of this endeavor is human capital; investment in human resources through education and research and development (R&D) has enabled the US to innovate and progress. In recent years, the National Academy of Sciences (NAS et al., 2007a) has expressed concern that the US may lose its leading position in science and, by corollary, endanger the long-term growth of its economy and its economic competitiveness (NAS et al., 2007a).

The NAS pointedly expresses that the educational challenge is that a “danger exists that Americans may not know enough about science, technology or mathematics to significantly contribute to, or fully benefit from, the knowledge-based society that is already taking shape around us” (NAS et al, 2007a, p. 94). Confounding this situation is that most lay people within the US do not fully understand and appreciate the importance
of encouraging students to study science and engineering and how these skills can stimulate enormous economic and social benefits (NAS et al., 2007a; OECD, 2009a). The American Association for the Advancement of Science (AAAS) expressed similar concern that the “images that many people have of science and how it works are often distorted. The myths and stereotypes that young people have about science are not dispelled when science teaching focuses narrowly on the laws, concepts, and theories of science” (American Association for the Advancement of Science, [AAAS], 1990). As a result, the AAAS recommended that the study of science as a way of knowing (i.e., the nature of science; NOS) needs to be made explicit in the curriculum (AAAS, 1990). In the newly released framework for K-12 science education (National Research Council [NRC], 2011, p. 1-3), the vision for K-12 science education is for high school graduates to be “critical consumers of scientific information related to their everyday lives, and to continue to learn about science throughout their lives” so that they can understand “how science and engineering are instrumental in addressing major challenges that confront society today.” Understanding the nature of science (NOS) through understanding the nature of scientific practices is fundamental to and predicates this vision for K-12 science education.

Science education researchers endorse this vision and believe that enhanced knowledge of the what science is and is not, that is, the nature of science (NOS), will support student learning of science content and applications (Smith, Maclin, Houghton, & Hennessey, 2000; Sandoval, 2005); stimulate interest in science (Carey & Smith, 1993; McComas, Clough, & Almazor, 1998; Ryder & Banner, 2011); make students more
informed consumers of scientific arguments and claims (Sandoval, 2005; Lombrozo, Thanukos, & Weisberg, 2008) and help them better understand the rapid and often complex scientific advancements made by researchers (McComas, Clough, & Almazroa, 1998; Bell, Blair, Crawford, & Lederman, 2003). Essential to the NRC (NRC, 2011) realizing their vision is to provide students with developmentally appropriate learning progressions in each of the core science disciplines; students will acquire skills through taking part in authentic scientific practices. The NRC (2011) reason that by taking part in authentic, in-depth scientific practices and inquiry, students will obtain a better understanding of how scientific knowledge is constructed and develop more informed epistemologies as they progress through the K-12 system. The NRC (2011) postulate that a more sophisticated understanding of science as a way of knowing will enable students to become “critical consumers of scientific information” (NRC, 2011, p. 1-3) by the time they leave high school.

With many elementary students having no exposure to science learning, the NRC’s (2011) framework calls for the teaching of science to all elementary students. The NRC’s concern is that, if students at this young age are not taught science, students will not “personally identify” (NRC, 2011, p. 11-4) with science and this will interfere with the sustained developmental process needed to cultivate students’ interest, motivation and understanding of science as a way of knowing beyond their K-12 education. Given this reality, there are limited number of studies on elementary students’ epistemic knowledge as it relates to scientific practice (Metz, 2011), but of those conducted, each has found that even young students in early elementary grades are able to develop their epistemic
knowledge if the instruction is explicit and reflective (Smith, Maclin, Houghton, & Hennessey, 2000; Metz, 2004; Akerson & Donnelly, 2010; Akerson, Buck, & Quigley, 2011; Quigley, Pongsanon, & Akerson, 2011). The positive results from these studies suggest that changes to the instructional practices of elementary teachers could help young students develop the epistemological knowledge needed to cultivate their interest and “personal identity” toward science which will help them become “critical consumers of scientific information” (NRC, 2011, p. 1-3) by the time they leave high school. This developmental progression starting in the early schooling years is fundamental to the NRC realizing its goal of having children build on their knowledge and understanding of science so that they can develop sophisticated and coherent views of the scientific enterprise.

**Statement of the Problem**

The research to date on elementary students’ epistemic knowledge has been conducted using small qualitative studies. A key area specified by the NRC (2011) for future research is the development of high-quality assessments of student learning and instruction. The NRC suggests that researchers need to develop instruments that can be efficiently used in “large-scale testing contexts” (NRC, 2011, p. 13-6) so that the relationship between engagement in authentic scientific practices and learning of the core ideas can be examined and whether taking part in authentic scientific practices supports the development of “understanding of the nature of science” (NRC, 2011, p. 13-4). The NRC (2011) recommends that student content knowledge should be tracked over time as their learning progresses and that use could be made of assessment methodology that can
provide educators with scores in multiple aspects of their students’ science content knowledge. The goal is to provide teachers with the data needed to inform their instructional practices so that they can help their students construct scientific content and epistemic knowledge. Given the importance that the NRC (2011) and other researchers place on having students understand the nature of scientific knowledge construction, a need existed for a measure that can reliability and optimally assess students’ NOS views. The NRC believes that if successful their reform efforts will sustain students’ future interest in and learning of science; what is needed are instruments that can help determine if their reform efforts are succeeding.

The Nature of Science Instrument-Elementary (NOSI-E) was developed to produce a reliable measurement tool to assess student understanding of the nature of science (NOS) construct (Peoples, O’Dwyer, Shields, & Wang, in review). The NOSI-E was created as part of the “Evolution Readiness: A Modeling Approach” [DRL-0822213] (ER) project at the Concord Consortium and Boston College. The instrument, designed for large-scale assessment use is conceptualized using five construct domains. Data from 741 elementary students were used to pilot the Rasch scale with continuous improvements made over three successive administrations. The psychometric properties of the NOSI-E instrument were consistent with the basic assumptions of the Rasch-model, namely that the items are equal-interval, invariant and well-fitting. Development work on the NOSI-E used a unidimensional Rasch model. The resulting instrument allowed for the student distribution to be divided into three levels of understanding (naïve, adequate and informed). Items from each of the five NOS domains (Empirical,
Theory-Laden, Certainty, Inventive and Socially and Culturally Embedded) are spread along the scale’s continuum and appear to overlap well. Most importantly, the scale seemed appropriately calibrated for elementary school-aged children, the target age group. Details of the NOSI-E’s development are summarized in Chapter 2.

Wilson (2003) emphasizes the need for measurement tools that can reliably model a construct when it “manifests itself in real-world situations” (p. 8). The NOSI-E is deconstructed into five domains with each dimension representing an important subcomponent of the NOS construct. Although the NOSI-E developed was represented by the unidimensional Rasch model (Peoples, O’Dwyer, Shields, & Wang, in review), the unidimensional Rasch model is only one of three possible theoretical models to represent the internal structure of the NOS construct. The internal structure of the NOS construct could, for example, be multidimensional and composed of five inter-related domains. If the NOS construct is considered and modeled as unidimensional when in reality it is multi-dimensional, this is problematic as the measurement tool would not reliably model the NOS construct as it “manifests itself in real-world situations” (Wilson, 2003, p. 8). In addition, mis-representing the NOS construct as unidimensional when it is in-fact multidimensional has important measurement consequences. The parameter estimates derived from the unidimensional model will be biased and it becomes impossible to “rank order test-takers without implicitly or explicitly weighting the dimensions” (Briggs & Wilson, 2003, p. 89). As a result, a multi-dimensional Rasch model may better represent the construct providing unbiased estimates and allowing test-takers to be rank ordered for each dimension separately. Similarly, a third representation of the NOS construct is
plausible. Each dimension may be sufficiently distinct in content that each dimension could be treated as a separate, unidimensional scale; this approach to modeling a construct’s items is commonly called the consecutive approach (Briggs & Wilson, 2003). One of the advantages of the multi-dimensional and consecutive approaches to modeling the NOS construct is that both approaches provide scores on the multiple aspects of the construct. In contrast, the unidimensional model provides a single score for each student. The benefit of having student-level scores on each dimension of the NOS construct is that it will provide teachers and researchers with more diagnostic information on students’ abilities and help them realize the NRC’s goal of having high quality data available to inform their pedagogy (NRC, 2011). It is important therefore to determine which theoretical model best represents the NOS construct in order to reliably and accurately represent students’ views of NOS within the elementary science.

In summary, as science education moves toward students learning science through engaging in authentic scientific practices (NRC, 2011), it will be important to assess whether this new approach to teaching science is effective. Understanding the nature of these scientific practices is an important goal of the NRC’s reforms (2011) as the NRC and other practitioners believe that this will help students’ develop their epistemic and science content knowledge (Smith, Maclin, Houghton, & Hennessey, 2000; Sandoval, 2005; NRC, 2011). The NOSI-E, designed to assess elementary student understanding of the nature science, can be used as one measure of whether this reform effort is impactful. However, it is important that student estimates derived from the Rasch model are
unbiased and interpretable, and provide researchers and educators with reliable data to inform their research and practices, respectively.

**Purpose of Study**

The primary purpose of this dissertation was to determine which of three Rasch models (unidimensional, multidimensional or consecutive) best represents the internal structure of NOS construct. The internal structure of a construct is related to the structural validity aspect of construct validity. Empirical evidence from each of these models, therefore, was used to provide support for the structural validity aspect of the NOS construct. If the data fit each of the models, then the measures from each of the models conform to the following Rasch principles: they are linear, additive, interval-level, invariant and hierarchical. All these properties appropriately meet the assumptions of parametric testing. It is not clear however, of the three models, which model will provide the best theoretically-grounded, interpretable, reliable and responsive measure of NOS, which can furnish science education researchers unbiased student parameter estimates. It is essential, as highlighted in the previous section, to represent the internal structure of the NOS construct properly as misspecification of this structure would lead to biased parameter estimates and these estimates may not be sufficiently reliable for the intended purpose of the instrument. Estimates from the appropriate Rasch model can subsequently be used in large-scale investigations advocated by the NRC (2011) to examine how student NOS understanding relates to other important facets of a student’s science learning such as their core content knowledge, science self-efficacy and their science classroom learning environment and to their personal epistemological
development. These relationships are essential for establishing evidence for the external validity aspect of the NOS construct. This aspect of validity is addressed in this dissertation as it provides support that the measures from the best internal model are responsive and can be used to examine their empirical relationships to external measures. Specifically, the purpose of this study is (1) to assess the most appropriate internal structure for the NOS construct; and, by corollary, (2) to assess the instrument’s concurrent external validity by determining whether measures derived from the most appropriate NOS Rasch model are related to students’ science achievement and to their views on the constructivist nature of their science classroom learning environment.

**Research Question**

The research question that was addressed in this dissertation was:

Q1: Is the internal structure of the NOS construct best represented by a unidimensional, multidimensional or consecutive Rasch model? Which of the three Rasch-based models will provide the most reliable, interpretable and responsive Rasch-based measure that will assess elementary students’ understanding of the nature of science (NOS); and, one that can be used in science education research and teaching?

To address this question, data from the NSF-funded Evolution Readiness: A Modeling Approach study will be used. Student Likert responses to NOS items will initially be modeled using a unidimensional Rasch model and then compared to a
consecutive and multidimensional Rasch model. Measures from the best fitting model will be used in multilevel regression models to determine if student understanding of NOS can predict student achievement and students’ views of their science classroom learning environment.

In the following section, the research problem will be described and placed in the perspective of current US reform efforts in science; the concerns over US student achievement in science and its impact on the sustainability of US global competitiveness. In addition, the methodological orientation of this dissertation requires that the research problem be described in the framework of competing test theories.

**Description of the Research Problem**

The goal of current United States (US) reform efforts in science is to change how students are taught in science; these reform efforts are spurred by concerns over US student achievement in comparison to their international peers and by the continuing achievement gaps between sub-groups of students within the US. Similarly, the lack of student interest in studying science has led to structural cracks within the vertically integrated education system with concerns that the US will lose its competitive edge due to insufficient numbers of students graduating from college with science degrees. These contextual factors and the role that students’ understanding of the nature of science can play in US science education reforms will be discussed. Unique to this dissertation is an evaluation of a Rasch-based instrument designed to measure elementary students’ NOS understanding. A discussion of the current instruments available for measuring students’ understanding of NOS and their applicability will first be provided; to conclude, the
rationale for using Rasch-based methodology to develop the NOSI-E is provided in the context of sound measurement practice.

**US science education reform efforts.** The NRC wants to transform how science is taught in the nation’s K-12 schools (NRC, 2011). Three dimensions will premise these reforms: scientific and engineering practices; crosscutting concepts and core ideas. One of the main goals of the new framework for science is to provide students with a more in-depth knowledge of science by exploring fewer topics and by having students engage in scientific practices (e.g., modeling, developing explanation and engaging in argumentation) that are used by scientists to construct knowledge. Similarly, certain core concepts such as “patterns” and “cause and effect” permeate across the scientific disciplines and understanding these concepts will help students appreciate and develop a “cumulative, coherent and usable understanding of science and engineering” (NRC, 2011, p. 4-1). Student learning in the sciences will be focused on four core subject disciplines (Life Sciences; Physical Sciences; Earth and Space Sciences; and, Engineering Technology and the Applications of Science); students will revisit and build content knowledge in each discipline as they progress through each grade level. Standards will be produced in the near future to articulate this framework and to guide teachers in implementing the new path for student science learning (NRC, 2011).

Teaching of the core ideas and crosscutting concepts through engaging all K-12 students in in-depth science activities and practices is predicted to motivate more students to study science at the tertiary educational level and chose careers in science (NRC, 2011). For those not following a science career path, it is expected that they will be more
informed citizens who can better understand the increasingly complex knowledge-based world in which they live (NRC, 2011).

**US student science achievement in a global knowledge-based economy.**

Student achievement in science in the US is of great concern; when compared to their peers in other nations, US students appear to struggle. In the latest Trends in Mathematics and Science Study (Trends in Mathematics and Science Study [TIMSS], 2008), US fourth grade students ranked eighth in overall achievement among the 36 countries that participated. To put this in perspective, the difference between the US students’ grade four mean score (539) and Singapore’s student grade four mean score (587; the top ranked country) was 48 points; the standardized effect size difference is 0.60, a medium/large effect. In terms of TIMSS achievement benchmarks, the US ranked fourth with 47% of their students achieving at the high or advanced performance level; this compares to 66% of students within Singapore. In 2007, compared to the U.S., Singapore had almost 2.5 times the number of students scoring in the top, advanced performance category. Contributing to this difference is the significant improvement in scores by Singapore’s fourth graders between 2003 (565) and 2007 (587); in contrast fourth grade US students’ scores remained static (TIMSS, 2008).

Similar to their fourth grade peers, US eighth graders’ science performance in TIMSS has remained static between 2003 and 2007 (TIMSS, 2008) with no significant improvement in the mean science score. By eighth grade, US TIMSS student performance places them eleventh out of forty nine countries in the overall rankings. The standardized effect size difference between the US students’ mean score (520) and that of
the top performer, Singapore (567), was 0.55, a medium/large effect size; this effect size is comparable to that of their fourth grade peers indicating that U.S. students continue to lag behind. Just over a third (38%) of US eighth grade students achieve at the high or advanced performance level benchmarks (TIMSS, 2008) in 2007; this compares to 61% in Singapore. At the Advanced performance level, three times as many Singaporean students achieve at the highest level than US students (TIMSS, 2008).

Although TIMSS does not assess students at the high school level, a comparable international study, the Programme for International Student Assessment (PISA), provides evidence that US students’ performance stagnates in the sciences when compared to their peers worldwide and to their younger U.S. peers. In the Organization of Economic Cooperation and Development’s 2009 study of science literacy (Organization of Economic Cooperation and Development [OECD], 2009a), US high school students’ average score does not differ from the OECD study’s average and US students score significantly lower than 18 of the 65 countries participating in the study. The standardized effect size difference between US students’ mean score and that of Shanghai-China (the top performing country) is 0.80; the standardized effect size difference (0.4) between US students and Singapore’s fourth ranked students was similar to that highlighted above for their younger peers in the TIMSS. Similar to the finding in the TIMSS studies, the US falls behind other countries in the percentage of students that achieve in the top performance categories of PISA. In 2009, only 9.2% of US students scored within the two top performing categories; this compares to 24.3% for Shanghai China; 19.9% for Singapore and 11.6% for South Korea. If as Bracey (2009, p. 34)
suggests, the “future innovators and leaders are more likely to come from high scorers”,
then the competitiveness of the US economy could be detrimentally impacted. The cost
of this deficit could slow US innovation and economic growth and enable countries such
as China, with a larger overall higher achieving population to become the engine of the
global economy and reap the rewards of this dominant position.

Achievement gaps do not just exist only at the international level. Domestically,
large racial, socio-economic and state achievement gaps are evident in the National
Assessment of Educational Progress (NAEP) study of US student science achievement
(National Center for Education Statistics [NCES], 2011) with similar gaps in reading and
mathematics. NAEP is the principal study used within the US to assess student progress
in reading, mathematics and science at key grades (fourth, eighth and twelfth) in the K-12
educational system. In science, in 2009, large achievement gaps (between 0.8 and 1.0
standard deviations) exist between White and African American and Hispanic students
(NCES, 2011) with White students outperforming African American and Hispanic
students in grades 4, 8 and 12. Similar and related achievement gaps are apparent
between students from high and low socio-economic backgrounds with students from
high socio-economic backgrounds outperforming their peers by almost one standard
deviation at both fourth grade and eighth grade in science (NCES, 2011).

In the newly released 2011 data, the achievement gap in grade 8 has remained
static across socio-economic groups (NCES, 2012); no data were available for grade 4.
There is some evidence that the racial achievement gaps are narrowing (NCES, 2012)
with the increase in average scaled scores for Black and Hispanic students (3 points and 5
points, respectively) outpacing the increase witnessed, on average, by White students (1 point). Although this is encouraging, the standardized effect size difference of the 2011 gaps remains large for both comparisons: the effect size of the Black/White achievement gap was 1.09 in 2011 with the effect size of the Hispanic/White achievement gap equaling 0.82. As a result, a substantial opportunity cost exists for the nation in terms of its competitiveness and economic wellbeing.

In a study of the economic impact of these achievement gaps (based on Reading and Mathematics scores), McKinsey and Company (2009) found that if the US had closed the international achievement gap that existed in 2008, the gain to the economy would have represented between 9-16% of the US’s Gross Domestic Product (GDP). The opportunity cost to the US economy of domestic achievement gaps is similarly onerous with the closing of domestic achievement gaps (based on racial and poverty gaps in Reading and Mathematics) each representing a further 2-5% of GDP (McKinsey & Company, 2009). This is a substantial opportunity cost to the US economy.

Overall, the TIMSS, PISA and NAEP data combined paint a worrisome picture for US science achievement. As previously mentioned, the US’s continued competitiveness in the global economy is heavily dependent on it maintaining its lead in science and engineering (NAS et al., 2007a; NAS et al., 2007b) through its investment in education and R&D. Newly emerging economies pose a particular threat to US supremacy; for example, absolute expenditures in R&D by China increased by 500% between 1991 and 2002 (NAS et al., 2007a); in the same time span, US R&D expenditures increased by only 140%. With this aggressive growth rate in national
spending, China ranked third (in purchasing parity terms) in spending behind the US and Japan in 2005 (OECD, 2007). Research intensity (research expenditures as a percent of Gross Domestic Product (GDP)) is another relative indicator of how countries invest in research and development activities. The growth in research intensity by China and South Korea outstripped that of the US over the last two decades; in 2004, for the first time, South Korea overtook the US in research intensity (Eaton & Kortum, 2007).

In the latest OECD’s Science, Technology and Industry scoreboard (OECD, 2009b), the growth in US government R&D expenditures continues to lag, with 10 other countries outpacing the US. With competition only set to intensify from other countries in the increasingly knowledge-based global economic environment; the US’s position as the leading economy in the world is threatened. In a recent US Department of Commerce’s report (United States Department of Commerce [USDC], 2011) on science, technology, engineering and mathematics (STEM) jobs; the USDC emphasize the critical role STEM workers play in encouraging innovation and in investing and creating new companies and industries within the US that can support the continued growth of the US economy. Highlighted in the USDC report (USCD, 2011) is the concern that the US will not have a sufficient supply of STEM workers to sustain the economy in the future, despite STEM jobs commanding, on average, higher earnings than non-STEM jobs.

**Structural cracks in the US’s vertically integrated education system.**

Contributing to the lack of STEM workers is the rapid decline in the US’s share of the world’s science and engineering doctoral graduates with countries such as China close to surpassing the US in the total number of graduates who obtain Ph.D.’s in science and
engineering (National Science Board, [NSB], 2010). When compared to the OECD countries, the US produces just 28% of all doctoral graduates; this compares to almost 50% produced by the European community (OECD, 2009b). Perhaps more worrisome is that between a quarter and half of the US’s graduates are foreign-born and may intend to return to their home countries and not contribute to the US’s economic well-being (NSB, 2010). At the undergraduate level, the US falls far behind other countries with only 15% of all undergraduate degrees awarded in science and engineering; this compares to 38% in South Korea; France (47%); China (50%) and Singapore (67%), (NAS et al., 2007b). Notably, these are the same countries that have been shown previously to out rank and out achieve US students in international science achievement tests. The US’s share of the world’s undergraduate engineering degrees halved (12% to 6%) between 1991 and 2000 (Freeman, 2007).

Similar patterns are also evident in high school graduation rates. At the K-12 level the US ranks 17th out of 23 OECD countries in terms of high school graduation rates (NSB, 2010). A silver lining is emerging with the number of high school students taking Advanced Placement tests in science subjects increasing fivefold between 1990 and 2008; even though the passing rate has either remained static or declined, many of these students will go on to study science at the tertiary level (NSB, 2010). The challenge is to help these students graduate from college; many potential scientists drop out of the sciences during their undergraduate experience and transfer into non-science subjects (Stephan, 2007). It appears that at every level of the vertically integrated educational
system, US students are falling behind their peers internationally; this has significant, potential repercussions for the US’s ability to sustain its economic growth and security.

The National Academy of Sciences and fifteen of the leading business organizations express concern that the US may lose its leading position in science and, by corollary, may endanger the long-term growth of its economy and economic competitiveness if these achievement gaps and other structural cracks highlighted above persist (NAS et al., 2007a; Tapping America’s Potential [TAP], 2008). The USDC (2011) emphasizes that the projected growth of STEM employment opportunities is going to outpace those of non-STEM jobs between now and 2018; with STEM jobs earning, on average, more than non-STEM jobs. STEM jobs are needed to stimulate the economy and are critical to sustain future innovation and economic growth (TAP, 2008). In order to fill these STEM jobs and meet the future needs of the US economy, business leaders (TAP, 2008) recommend that the US double the number of STEM graduates by 2015 and enact major reforms to help attract students into the sciences, especially those from under-represented groups (e.g., female, African-American, Hispanic). To improve science achievement overall, business leaders (TAP, 2008) further recommend that incentives are needed to attract qualified science teachers at the K-12 level to transform science education.

Another priority of TAP (2008) is to raise public awareness on the importance of STEM to US economic security and the need to improve US student achievement in science. In discussing the new K-12 science frameworks, the NRC (2011) suggests that having students engage in scientific practices will help raise students’ awareness of how
scientific knowledge is produced and by “reflecting on their nature” (p2-2); students will build a more “flexible and coherent…understanding of science” (p2-2). Although the NRC’s vision is that K-12 students’ who have a greater understanding of the nature of scientific practices will chose to study science at college and enter STEM careers; they also hope that it will, for those that do not pursue a STEM career, foster continued interest and awareness of science throughout the lives of students so they will be more informed users of scientific information and of science’s impact on society. Carey and Smith (1993) suggest that the typical student cannot absorb all there is to know in science during their schooling but what is needed is for the science curriculum to stimulate a life time interest in science so that in the future, students can make informed decisions on scientific issues that affect them and society. This philosophy is at the heart of the NRC’s framework with its goal of having students focus on learning a narrower set of core ideas and crosscutting concepts so that they can understand the nature of the scientific enterprise and use that understanding to contribute to controversial issues that influence society.

**The role of the nature of science in US science education reform efforts.** The NRC acknowledges that “although there is no universal agreement about teaching the nature of science, there is a strong consensus about characteristics of the scientific enterprise that should be understood by the educated citizen” (NRC, 2011, p. 3-22). As Carey and Smith (1993) pointed out in their article in 1993 that the teaching of science in the US is performed out with the context of real scientific inquiry with the consequence that this prohibits students from understanding the nature of how science knowledge is
constructed. The NRC’s goal in having students “practice” science is that students will have the time to reflect on the nature of scientific practices and by doing so will better understand how the scientific enterprise functions; how scientific knowledge is constructed and what characterizes science as a way of knowing (NRC, 1996; NRC, 2011). For students to understand how science functions requires students to not only understand scientific content but also to appreciate how scientists go about their work (practice science) and to realize the part epistemic knowledge has in helping students construct knowledge (NRC, 2011).

Epistemic knowledge is defined in the NRC’s framework as the “knowledge of the constructs and values that are intrinsic to science” (NRC 2011, p. 3-22). The constructs highlighted by the NRC (2011, p. 3-22) that are fundamental to understanding the nature of science are based on the same theoretical framework as those established previously by researchers (Lederman, 1992; Lederman, Abd-El-Khalick, Bell, & Schwartz, 2002; Lederman, 2007) studying the nature of science and the one used here in this research. According to these experts (Lederman, 1992; Lederman, Abd-El-Khalick, Bell, & Schwartz, 2002; Lederman, 2007) to understand the nature of science, students need to comprehend that science is (1) based on real-world observations which help explain the natural world; (2) that scientists need to be imaginative and creative to ideate and progress their work; (3) that scientific knowledge is durable but involves uncertainty; (4) that scientific theories are subjective but coherent principles that guide and explain the inference process; and (5) that science takes place within social and cultural contexts.
The theoretical framework used in this dissertation for the nature of science is founded on these five constructs.

Historically, the NOS has been viewed as a multidimensional, complex construct whose conception has evolved over time. Lederman, Abd-El-Khalick, Bell, and Schwartz, (2002), in developing the Views of Nature of Science (VNOS) Questionnaire, acknowledge that “similar to scientific knowledge, conceptions of the NOS are tentative and dynamic” (p. 499). In the early twentieth century student understanding of the NOS focused on their understanding of the scientific method and scientific processes (McComas, Clough, & Almazroa, 1998). In the mid-twentieth century philosophical elements were added to the conceptual framework with scientific knowledge deemed tentative, probabilistic, naturalistic and humanistic (McComas, Clough, & Almazroa, 1998; Abd-El-Khalick & Lederman, 2000). By the late twentieth century, the concept of the NOS was viewed more as a psychological and sociological construct with psycho/social factors such as the role of creativity and culture in science and how scientific claims and knowledge are fashioned considered part of the NOS construct (McComas, Clough, & Almazroa, 1998; AAAS, 1993; National Research Council, [NRC], 1996). There have been no significant changes to how the NOS construct is generally defined since the late twentieth century with most current science researchers viewing the NOS as an affective measure rather than a cognitive measure.

In a review of 105 empirical studies related to NOS, Deng, Chen, Tsai and Chai (2011, p. 963) surmise that the construct of NOS, in relation to scientific inquiry activities “refers to the epistemological underpinnings of these science activities and the
characteristics of the knowledge produced”. To date measurement of students’ understanding of NOS has largely been accomplished using open-ended semi-structured interviews. This form of instrument is ideal for exploring and assessing students’ epistemological beliefs. The NOSI-E uses Likert scales to measure students’ NOS understanding. This closed form of instrumentation is better suited to measuring students’ concretized knowledge of NOS and, under the Rasch framework, measuring the level of NOS understanding that is categorized by naive, adequate and informed conceptions of the NOS construct.

In its broadest sense, it has been defined as the, “the values and beliefs inherent to scientific knowledge and its development” (Lederman, Abd-El-Khalick, Bell, & Schwartz, 2002, p. 498). The National Research Council (NRC, 1996, p. 21) in their current standards setting document, provide a more succinct definition of the NOS: they state that:

science is a way of knowing that is characterized by empirical criteria, logical argument and skeptical review. Students should develop an understanding of what science is, what science is not, what science can and cannot do, and how science contributes to culture. (p. 21)

Informed by this conceptualization, the NRC incorporated content standards for students to learn and be assessed on the history and nature of science at the elementary and secondary level. Students, for example, by the end of fifth grade, are expected to understand that scientific ideas are tested “using observation, experiments and theoretical
models” (NRC, 1996, p. 171); that scientific ideas are “tentative and subject to change” (NRC, 1996, p. 171); and that scientists often disagree and offer conflicting results from the same research.

At the high school level, students should, for example, be able to understand the more sophisticated concept that science differs from other subjects in that it accumulates knowledge through “empirical standards, logical arguments and skepticism” (NRC, 1996, p. 201). It is unclear at this time if the new standards being developed to operationalize the recently released NRC frameworks (NRC, 2011) will include standards to explicitly teach students about the NOS. However, it is clear that the intent of the NRC is for students to reflect on the nature of the scientific practices they participate in and that this reflection will provide them with a deeper understanding of how science operates; how scientists construct knowledge and how science contributes to their worldviews. The NOS provides the mesh upon which students will practice and construct scientific knowledge. In order to evaluate progress in realizing this goal, it is important to have instruments that can, on a large-scale, reliably and accurately measure student understanding of the NOS. An instrument that is reliable, suitable for large-scale use, will ensure that changes in students’ understanding of the NOS can be accurately measured and enable policy-makers to determine if their reforms are successful.

**Measures of NOS.** To date, research that focused on whether students have an understanding of the NOS has been largely restricted to small-scale and mostly qualitative studies. These studies use predominantly in-depth interviews to assess student understanding of the NOS (Carey, Evans, Honda, Jay, & Unger, 1989; Ryder, Leach, &
Driver, 1999; Smith, Maclin, Houghton, & Hennessey, 2000; Moss, Abrams, & Robb, 2001; Thoermer & Sodian, 2002; Khisfe & Abd-El-Khalick, 2002; Bell, Blair, Crawford, & Lederman, 2003; Sandoval & Morrison, 2003; Kawasaki, Herrenkohl, & Yeary, 2004; Akerson & Abd-El-Khalick, 2005; Akerson & Volrich, 2006; Smith & Wenk, 2006; Khisfe & Lederman, 2006; Khisfe, 2008; Şahin & Köksal, 2010; Akerson & Donnelly, 2010; Akerson, Buck, & Quigley, 2011; Akerson, Buck, Donnelly, Nargund-Joshi, & Weiland, 2011; Wu & Wu, 2011). The majority of these studies are targeted at the middle; high school and college level and not at the elementary level, which is the focus of this research.

The studies highlighted above, by their nature, rely on qualitative instruments to measure student understanding of the NOS. The Nature of Science Survey instrument (NSS: Lederman & O’Malley, 1990) and the Views of Nature of Science (VNOS: Lederman, Abd-El-Khalick, Bell, & Schwartz, 2002) questionnaire are theoretically valid measures of the NOS and often used in science education research. Similarly, Carey, Evans, Honda, Jay and Unger (1989) “Nature of Science Interview” (NOSI), is grounded on the NOS’s theoretical framework and provides researchers with the ability to categorize students’ understanding into three levels (Level 1: naïve; Level 2: moderate; and Level 3: informed) of understanding. The NSS and VNOS use a combination of seven open-ended written questions and follow-up interviews to assess student understanding. The VNOS and NOSI use a semi-structured interview process to examine students’ views on the empirical, theory-laden, tentative and creative nature of scientific knowledge. The VNOS extends to assessing students’ perceptions of the role of theories
and laws in science; and to the role that social and cultural contexts play in science. The VNOS was first designed for pre-service teachers, undergraduate students and high school students (Lederman, Abd-El-Khalick, Bell, & Schwartz, 2002) but was adapted for use with younger, elementary students (Lederman & Lederman, 2005). In very young children (K-2), the VNOS is administered orally with researchers transcribing audiotaped responses. Whilst this methodology is invaluable and suited for classroom level studies; assessing students’ understanding of the NOS using open-ended instruments is time-consuming and resource intensive; the cost of training multiple interviewers and scorers is prohibitive and not practicable for large-scale use (Dogan & Abd-El-Khalick, 2008). The lack of standardization of the interview process negates the use of parametric testing as it calls into question the validity of the tests being used and what actually is being measured.

For the future, the clear advantage of large-scale quantitative studies is the ability to statistically test for changes in student NOS understanding; this is particularly important when the NRC’s new reforms are implemented and measures of their effectiveness are needed. Instruments that measure NOS understanding at the elementary level will help researchers understand early on whether students at this young age are able to develop foundational NOS understanding and to investigate whether this understanding helps improve student content knowledge, interest and “personal identity” with science. Large-scale studies will also present researchers with the opportunity to statistically and unambiguously test whether teaching science through authentic scientific practices, is sufficient to improve elementary student understanding of NOS or, as many
researchers attest (Khisfe & Abd-El-Khalick, 2002; Akerson & Volrich, 2006; Quigley, Pongsanon, & Akerson, 2011) requires explicit and reflective instruction to embed NOS understanding. The NRC has expressed the need for all elementary students to learn science and for students at this young age to understand NOS in order for them to become scientifically literate by the time they leave K-12 education. Similarly, the NRC expresses the need to have quality assessments that can measure student outcomes effectively in large scale studies, yet are accessible to helping teachers in their instructional practices.

Quantitative instruments have been created in the past and used in large-scale studies (Lederman, Wade, & Bell, 1998) to measure student NOS understanding. Most lack rigorous validity evidence (Lederman, Abd-El-Khalick, Bell, & Schwartz, 2002; Khisfe, 2008) and all are ill-suited for elementary students. The Test on Understanding Science (TOUS: Cooley & Kloper, 1961) has been used widely to measure student understanding of the scientific enterprise, their views on scientists and the method(s) used by them (Lederman, Wade, & Bell, 1998). The TOUS, a 60-item multiple choice test, has been criticized in terms of its construct validity related to measuring the nature of science (Lederman, Wade, & Bell, 1998) as many of its items are focused on whether students appreciate science and scientists and not on their understanding of the nature of scientific knowledge construction.

Other instruments available, the Wisconsin Inventory of Science Processes (WISP: Scientific Literacy Research Center, 1967); the Science Process Inventory (SPI: Welch, 1966) and the Nature of Science Test (NOST: Kimball, 1968) are of limited
utility as they only provide one score for students’ understanding of the NOS. As such, it is not possible to compare students’ understanding across the different dimensions of the complex NOS construct and, as a result, the retrievable information is restricted. In addition, the WISP; SPI and NOST all consist of 60-plus multiple-choice items and are targeted for middle and high school students, and college students and/or teachers; these instruments are therefore inappropriate for elementary students in terms of length, cognitive ability and readability.

The View on Science-Technology-Society (VOSTS: Aikenhead & Ryan, 1992) consisting of an item pool of 114 multiple choice items is similarly targeted at high school students. The VOSTS is exceptional in that it used student qualitative information (viewpoints obtained through extensive interviews) to construct responses to each item’s key; as a result, a student often has several options (average of seven positions) from which to choose their viewpoint. The instrument was created in this manner to overcome the criticism of using a small number of response options typically used in convergent instruments; critics believe that a small set of response options imposes researchers’ preconceived notions of the NOS construct on students (Lederman, Abd-El-Khalick, Bell, & Schwartz, 2002) and consequently the instruments may not measure students’ “true” epistemological beliefs. The VOSTS developers did not provide a scoring system limiting its use in inferential statistical analyses. In a large-scale study (2082 students and 378 teachers) of Turkish students’ and teachers’ conceptions of the NOS, Dogan and Abd-El-Khalick (2008) were constrained to using frequency analyses and non-parametric
tests to compare NOS views across sub-populations highlighting the limited utility of the VOSTS.

Most recently, Lombrozo, Thanukos and Weisberg’s (2008) developed a scale to measure college student understanding of the NOS. This instrument, used to study the relationship between students’ acceptance of evolutionary theory and their understanding of the NOS, used a five point Likert-scale to measure students’ NOS understanding. The authors report that students’ understanding of the NOS is important in preventing students from rejecting the theory of evolution; they base this conclusion on the positive correlation between the two variables. However, this instrument and all of the quantitative instruments highlighted above fall prey to unsound measurement principles because they rely on ordinal measures that are not linear, interval-leveled or normally distributed and are, therefore, not suited for use in examining relationships between multiple variables and for use in inferential statistical testing (Boone, Townsend, & Staver, 2011).

**Measurement hegemony.** Parametric statistical tests require interval-level and linear measures and for the data to conform to a normal distribution. Likert scaling, most often used in self-report instruments, assumes that the data structure and sum score of the items are interval level and normally distributed (Van Alphen, Halfens, Hasman, & Imbos, 1994; Boone, Townsend, & Staver, 2011). Ordinal Likert measures do not meet the assumption of interval-level measurement as the difference between item attributes is not necessarily equivalent (or additive) over the full range of the scale and therefore these differences cannot be properly interpreted (Wolfe & Smith, 2007a; Boone, Townsend, &
Staver, 2011). Similarly, Likert scales with four response options are unlikely to be normally distributed. Parametric tests, at a minimum, require interval level, normally distributed measures in order that hypotheses can be tested and correct inferences can be made from the statistical analyses (Smith 2000, Boone, Townsend, & Staver, 2011). Scales created using classical test theory (CTT) methodology do not meet these assumptions.

The Rasch model for Likert-based scale development is distinct from those using CTT methods. The Rasch model is a prescriptive model whose primary purpose is to measure concatenations of items and examinees. It is a strict measurement system in which items and examinees are placed on a common interval scale and enable the linear combinations of items and examinees to be analyzed statistically. It provides for estimation of error for each item and each person estimate. Given these features, the Rasch model provides the expected performance on an item and test; provides estimates of the developmental distances between ordered skills or persons and provides a confirmatory test of the construct validity of a test.

Neumann, Neumann and Nehm (2011) have recently applied the Rasch model to Lombrozo, Thanukos and Weisberg’s (2008) instrument highlighted above. This is the first application of the Rasch model to a NOS instrument in the literature. Neumann, Neumann and Nehm (2011, p. 1396) highlighted that the “raw Likert-type scores did not meet the assumptions of normality” and that the distance between item category responses varied across items indicating that it was erroneous to assume that each item conformed to the same linear, additive measurement scale. The CTT methodology used
by Lombrozo, Thanukos and Weisberg (2008) were not appropriate for use in inferential statistical analyses suggesting their results should be treated with caution. Neumann, Neumann and Nehm (2011) also highlight the difficulty of applying a Rasch model to a pre-existing scale as many items proved redundant (i.e., same difficulty level and similar content) in the Rasch model and the items were, overall too easy for the sample.

In CTT methodology, instruments are developed to maximize the internal consistency (reliability) of the instrument and do not necessarily penalize redundancy as additional items improve the reliability of the scale. Cronbach’s alpha is the most often used measure of internal consistency within the CTT framework. It is a measure of the correlation between all possible split-halves; if items share variance, the sum of the item variances will be reduced relative to the total test score variance and provide an inflated estimate of the reliability of the instrument. In contrast, in the Rasch model, the internal consistency reliability coefficient, the person separation reliability, measures the ratio of the variance in latent person measures to the estimated person measures. Within the Rasch framework, therefore, reliability corresponds to the “reproducibility of relative measure locations” (Linacre, 2010, p. 515) with person reliability increasing with the sample ability variance. Therefore, redundancy of items in the Rasch model limits reliability and the instrument cannot discriminate the sample into enough levels to make it useful to researchers. Similarly, if the sample and items are off target, the instrument will have low person reliability as the standard errors used to calculate the separation coefficient are large and the stability of the person measures for each respondent is
reduced (Wolfe & Smith, 2007b). The benefits of the Rasch model will be discussed in more detail in Chapter 3 of this dissertation.

Neumann, Neumann and Nehm’s (2011) findings highlight the importance of having a scale with appropriate measurement properties that meet the assumptions of parametric tests. The goal of this research therefore, is to determine the most theoretically grounded, interpretable and reliable Rasch-based measure that aptly depicts the true dimensional structure of the NOS construct. This scale, if it fits the Rasch model well, will meet the assumptions of parametric testing and will be suitable for use by science education researchers.

**Significance of Study**

In discussing future directions for research into the NOS, Lederman (2007) highlights that there has been no systematic research into the relationship between the NOS and student learning of other science matter; sparse empirical evidence on the purported benefits of teaching NOS explicitly; and, no studies examining the influence of the NOS on student decision-making as it relates to scientific social issues. To date, this situation has not changed significantly. One reason for this is that there are no instruments available with the appropriate psychometric properties to measure NOS (Neumann, Neumann, & Nehm, 2011) that can stand up to the basic assumptions of large-scale parametric testing. If these types of relationships are to be tested and evaluated in the future, an instrument used to measure the NOS must be properly constructed and up to the task psychometrically. A Rasch-based instrument is up to this
task as instruments designed with Rasch principles are linear and interval-leveled; both fundamental assumptions required for parametric, inference testing.

If students, at a young age, are able to understand NOS, the NRC (2011) suggests that students will be able to “appreciate its (science’s) basic nature” (NRC, 2011, p. 3-23) and students will become more informed in “how any given practice contributes to the scientific enterprise” (NRC, 2011, p. 3-23). Fundamental to the NRC’s vision for science education is to start science education early and for clear learning progressions to be developed for each scientific practice and core idea. The NRC (2011) suggests that students’ understanding of the nature of scientific practices will be inculcated through students “practicing” science and this understanding will be important for students to appreciate the scientific enterprise. In the NRC’s (2011) new framework, the NRC explicitly states that research is needed into “how engagement in specific practices supports the development of both specific (core) ideas in science and understanding of the nature of science” (p. 13-4). Central to this vision is for science education to start at a young age for all students so that student understanding of scientific practices, core ideas and the nature of scientific knowledge is reinforced and becomes more sophisticated as they develop and progress K-12. It seems that the time is opportune for explicating these relationships with the introduction of the new frameworks and impending standards.

In attempting to make large-scale change to the way science is taught within the US, it essential to have reliable measures ready that can measure developmental change and examine the relationships highlighted by Lederman (2007). In particular, the relationship between students’ participation in scientific practices; student understanding
of the NOS; student achievement and student interest in science career paths could prove illuminating in whether the NRC achieves its vision of having “students develop an understanding of the practices of science and engineering” with the aim that they become “critical consumers of scientific information related to their everyday lives, and to continue to learn about science throughout their lives” (NRC, 2011, p. 1-3). With US students’ science achievement falling behind their peers in other countries (TIMSS, 2008; OECD, 2009a) and the opportunity cost of having large internal achievement gaps, it is important to investigate the most effective pedagogy and curricula to try to support and boost student science learning. Researchers hypothesize that a better understanding of the nature of science will enhance student learning of science content (Smith, Maclin, Houghton, & Hennessey, 2000; Sandoval, 2005; NRC, 2011) and help boost student science achievement overall. Qualitative studies (Smith, Maclin, Houghton, & Hennessey, 2000; Metz, 2004; Akerson & Donnelly, 2010; Akerson, Buck, & Quigley, 2011; Quigley, Pongsanon, & Akerson, 2011) have already reported that elementary students are able to develop more sophisticated understanding of NOS through explicit, reflective pedagogy. These findings suggest that teaching NOS early in a student’s education could not only support their NOS understanding but may help students improve their ability to understand science content knowledge. A NOS instrument suited to large-scale use, targeted at the elementary level and with sound psychometric properties is needed in order that parametric analyses can be reliably used to explicate complex multivariate relationships among these variables in young students.
Currently, there are no instruments targeted at assessing elementary students’ understanding of the complex NOS construct. A scale, suitable for large-scale use, is required that is appropriately modeled and constructed for elementary students. In addition to explicating complex multivariate relationships, this scale will provide practitioners with a tool that can measure change in students’ learning progressions; measuring students’ learning progressions is a fundamental goal of the NRC’s (2011) new framework for science. The responsiveness of an instrument refers to “the degree to which an instrument is capable of detecting changes in person measures following an intervention that is assumed to impact the target construct” (Wolfe & Smith, 2007b). By comparing the responsiveness of the instrument across the three competing models, there is the potential to design an instrument that can measure change over time on all five dimensions of the NOS construct (multidimensional model). If the multidimensional model is shown to be the best representation of the NOS construct, researchers and practitioners will receive “multiple aspects of proficiency rather than a single score” (NRC, 2011, p. 13-6) enabling them to tailor their instruction and design interventions needed to support change and improvement in student learning.

If successful, the NOSI-E will have the sensitivity to measure changes in elementary students’ beliefs and help educators and researchers examine the types of relationships highlighted above using appropriate large-scale methodology. If future research indicates that enhancing student views on the nature of science leads to better educational outcomes and promotes student interest in STEM careers; this could help the US maintain its current competitive advantage in the sciences and technology. It is
essential that reliable tools are available to measure the impact of changes to pedagogy and student learning in order that the most appropriate and effective practices take place in the science classroom to the benefit of the teachers and students. The NOSI-E was designed to provide such a tool.

In the following chapter, Chapter 2, a review of the literature that pertains to this study is presented. Literature concerning the theoretical framework for the NOS construct is first discussed. This is followed by a summary of previous research that is related to student understanding of the NOS construct. Rasch methodology and the context of this study is fundamental to this dissertation; a review of the utility of the Rasch model in scale construction and science education research is provided along with a summary of the project that premised and provided the data for this study. Chapter two concludes with an outline of the three theoretical representations for the internal structure of the NOS construct that is at the heart of this dissertation.
Chapter Two: Literature Review

This chapter provides an overview of the literature that is related to the NOS construct and the literature that is related to Rasch methodology. The review begins with the theoretical framework developed for the NOS construct. This will be followed by a detailed description of the underlying or internal structure of the NOS construct that will be modeled in this study. Science education researchers suggest that there are five domains of the NOS construct that elementary students should be able to grasp and understand; these domains are: Empirical; Theory-Laden; Certainty; Inventive; and, Socially and Culturally Embedded. The literature pertaining to and describing each domain will be discussed in turn; this part of the literature review concludes with a section outlining the descriptors researchers normally use to characterize student understanding of NOS. Most studies examining student understanding of NOS have been concerned with whether students are capable of learning about NOS through their classroom science activities implicitly or whether it requires teachers to explicitly teach students NOS in order for students to truly develop their epistemic knowledge. It is important to summarize the literature of both areas of study (implicit and explicit) as they offer the current status of student NOS understanding and provide insight into what factors can affect student understanding and epistemological development.

Rasch methodology was used to develop the NOSI-E instrument that was designed to measure elementary students’ understanding of NOS. Rasch methodology is also fundamental to this study as it is used to determine the internal structure of the NOS construct. As such, the literature review warrants a brief summary of why Rasch
methodology is useful in scale development and in science education research. This summary is followed by a description of the context for the study and of the prior instrument development activities conducted for the NOSI-E. The Evolution Readiness: A Modeling Approach [DRL-0822213] (ER) project provided the specific context for this study and was the impetus behind the development of the Rasch-based NOSI-E. The NOSI-E was developed over the three years of the ER project and the process used to develop the NOSI-E is outlined. The primary purpose of this dissertation is to determine if the internal structure of the NOS construct is best modeled using a unidimensional, consecutive or multidimensional Rasch model. These models represent three different theoretical positions for the internal structure of NOS construct (Wolfe & Smith, 2007b); this chapter concludes with a description of the three construct representations.

The Nature of Science Theoretical Framework

The theoretical model used for this study is primarily based on Lederman’s conceptualization of NOS (Lederman, 1992; Lederman, Abd-El-Khalick, Bell, & Schwartz, 2002; Lederman, 2007). This theoretical model is supported by the American Association for the Advancement of Science (AAAS; 1993); the National Science Teachers Association (NSTA; 2000); the National Research Council (NRC, 1996); and is integrated throughout the new science framework document developed by the National Research Council of the National Academy of Sciences (NRC, 2011). With this consensus, Lederman, Abd-El-Khalick, Bell and Schwartz (2002) suggest certain aspects of the NOS are acceptable and relevant to the lives of K-12 students. Eight aspects are included their conceptual model; five domains are: “scientific knowledge is tentative;
empirical; theory-laden; partly a product of human inference, imagination, and creativity; and socially and culturally embedded” (p. 499). In addition to these five domains, students should be able to distinguish between observation and inference, understand that there is no prescribed method in science and be informed on the “functions of and relationships between scientific theories and laws” (Lederman, Abd-El-Khalick, Bell and Schwartz, 2002, p. 499). Students’ epistemic knowledge is the “knowledge of the constructs and values that are intrinsic to science” (NRC, 2011, p. 3-22); the eight “constructs” highlighted above predicate students’ epistemic knowledge.

The first five domains of Lederman’s theoretical framework are suitable for elementary students; the target age group of this study. Elementary students are not cognitively ready to understand the relationship between theories and laws in science. In this study, student understanding of the distinction between observation and inference was not measured directly but items are included that measure student views on both concepts separately. Similarly, student understanding that there is no common scientific method will not be tested directly. The NOSI-E’s focus is on whether students understand that scientific knowledge is (a) based upon real-world observations and evidence that are needed to explain the natural world (Empirical: EMP) (b) driven by theoretical perspectives (Theory-Laden: THL) (c) durable but involves uncertainty (Certainty: CER); (d) creative and requires imagination (Inventive: INV); and (e) is socially and culturally embedded (SCE). Portrayed in Figure 2.1 is the multidimensional conceptual model for the NOSI-E and, as can be seen, is guided by the theoretical framework of Lederman, Abd-El-Khalick, Bell and Schwartz (2002) for K-12 students.
Figure 2.1. Conceptual Model for the NOSI-E

Source: Figure template guided by Briggs & Wilson (2003). Conceptualization based on Lederman’s (2007) theoretical framework.

The five dimensions are distinct, each providing information on a unique aspect of the NOS; the curved lines symbolize that the domains are, however, epistemologically interrelated by the overarching construct of NOS. The following sections will discuss each domain in turn and provide the theoretical literature-based reasoning for their inclusion in the NOS conceptual framework. The epistemological ideas developed by students in one domain of the NOS can inform or impact their ideas within the other domains (Smith, Maclin, Houghton, & Hennessey, 2000). For example, scientists perform their work in an interpretive explanatory framework (Carey & Smith, 1993).
which cultivates the notion that scientific knowledge is subjective and often tentative (Lederman, Abd-El-Khalick, Bell, & Schwartz, 2002). Therefore, it is important to remember that although these domains are discussed in turn, they are interrelated and, as a result, to understand the underpinnings of each domain will require reference to the other domains within the NOS construct.

**Empirical domain.** Observation through sensory experiences is at the heart of the empirical nature of scientific knowledge; however, not all scientific knowledge can be built directly using the five senses. Scientists take advantage of their five senses to observe and provide a descriptive interpretation and explanation of natural phenomena (Lederman, 2007). Apedoe and Ford (2010, p. 166) describe an “empirical attitude” which they define as “a habit of mind to actively search for feedback on one’s ideas from the material world.” Scientists’ need an inquiring mind that explores natural phenomena and use what is constructed through observation, experimentation and theory-building to inform and change their ideas. Reciprocity exists between data informing phenomena and phenomena informing scientific ideas about real-world phenomena (Apedoe & Ford, 2010). Observations that use senses are relatively easy to obtain and little argumentation about the knowledge claims derived from these observations are likely to occur among scientists.

In practice, however, observation of phenomena (e.g., the structure of the atom) is not always accessible to scientists’ senses and therefore scientists have to rely on inference, models and theoretical frameworks to explain natural phenomena (Lederman, Abd-El-Khalick, Bell, & Schwartz, 2002). Scientists therefore cannot always trust
observation data alone to explain phenomena but often need heuristic thinking to understand and explain difficult concepts and phenomena (Niaz, 2001). To understand the empirical nature of science, it is important to appreciate that scientific knowledge constructed through observations is necessarily limited by scientists’ sensory capabilities and that scientific knowledge is implicitly theory-laden (Khishfe & Abd-El-Khalick, 2002).

Theory-laden domain. Through experimentation and theory-building, evidence is collected that is used to support scientific claims; this process is inherently subjective with the distinct possibility that scientists provide competing claims for the same evidence (Lederman, 2007). Each scientist has their own mind-set (empirical attitude) which influences how they perform their work and how they interpret and explain their findings; these differing mind-sets provide the source of subjectivity in science and the basis for argumentation prevalent in science knowledge construction. Scientists use empirical evidence to buttress their theories on observable and unobservable entities (Lederman, Abd-El-Khalick, Bell, & Schwartz, 2002) and it is the confluence of evidence that builds and sustains scientific theories over time. Theories can be “characterized as a coherent set of subjective principles or concepts that are used to explain a wide range of phenomena” (Smith, Maclin, Houghton, & Hennessey, 2000, p. 401) and are crucial in guiding research questions, hypotheses setting and the inferential process (Lederman, 2007).

Theoretical perspectives provide the context in which science operates and scientists’ beliefs and theoretical propensities drive the knowledge construction process.
Distinct from many forms of knowledge building, science knowledge construction is tentative and involves uncertainty (McComas, Clough, & Almazroa, 1998). A scientist’s theoretical propensity and research may provide new evidence that disconfirms a prior theory that has been supported and corroborated by previously accumulated evidence. Although science is theory-laden and durable, theories can never be absolutely proved and conclusive knowledge in science is unobtainable (McComas, Clough, & Almazroa, 1998). Explanations and findings in science are always subject to change and this predicates the tentative nature of scientific practices.

**Certainty domain.** Francis Bacon in the 17th century postulated that in order to make a valid conclusion, all facts have to be first collected, assimilated and interpreted. This process called induction enables humans to generalize and form predictions (McComas, Clough, & Almazroa, 1998). Induction is essential to how scientific knowledge is constructed but it is, by its nature, problematic. It is impossible to collect or observe every piece of evidence needed to make a conclusive statement about an entity or phenomenon. New evidence collected may provide support for a different theoretical perspective than previously proposed and prior scientific claims are challenged given the new evidence (Lederman, 2007). To understand the tentative nature of scientific knowledge and theory formation requires an understanding that inferences can only be made about unobservable entities and theories can only provide “inferred explanations” for observable entities (Lederman & O’Malley, 1990, p. 226). Theories endure if the “inferred explanations” or inferences are coherent with prior knowledge; and, if the new evidence collected helps to explain the unexplained or clarifies prior relationships or
findings (NRC, 2011). Gauch (2009) suggests that “evidence” is one of the seven “pillars” of science; in order for conclusions to be reached, science “demands evidence” (p. 674) to support them. Theories that are durable are those that are supported by a body of evidence and have been subject to and survived critique, argumentation and scrutiny (NRC, 2011).

Induction on its own is not sufficient to develop theories and explanations for natural phenomena. The collection of evidence-based facts and/or observations is fundamental to developing theories on real-world phenomena but it is the creativity of individual scientists to collectively transform these facts into a coherent explanatory framework (McComas, Clough, & Almazroa, 1998; Lederman, 2007). This creative element of science provides another reason why science, by its nature, is tentative and often uncertain. Scientists pursue and test their ideas resulting in new evidence and inventive explanations for their findings; these explanations often contradict other scientists’ interpretations of similar data leading to back and forth argumentation and confusion on whose scientific claims are justified (Lederman, Abd-El-Khalick, Bell & Schwarz, 2002). Osborne (2010) believes that critique is at the heart of science knowledge construction and without it, science knowledge would not be reliable. McComas, Clough and Almazroa (1998, p. 59) pointedly expresses that the process of turning evidence into scientific knowledge requires a “creative leap” or “abduction” and that it is this that augments the induction process and enables generalizations to be made.

**Inventive domain.** The field of science, contrary to often held beliefs, requires scientists to be imaginative and creative; science is often purported to be objective,
rational and prescribed (Lederman, Abd-El-Khalick, Bell, & Schwartz, 2002). The “creative leap” described by McComas, Clough and Almazroa (1998) derives from scientists’ personae and it is their individualities that enable scientists to “invent” explanations (Lederman, 2007). Scientists have their own perceptual apparatuses which are governed by preconceptions, prior knowledge and training (Lederman Abd-El-Khalick, Bell, & Schwartz, 2002); their perceptual apparatuses or “mind-sets” manage the creative element they bring to all aspects of their work (Lederman, 2007). Science not only advances through scientists’ “inventing” explanations but also by scientists’ using their creativity and their imaginations to develop research questions, formulate ideas, and design appropriate methodology in order to test their ideas and construct knowledge (Lederman, 2007; NRC, 2011). As Matthews quotes; “scientific explanations ‘go beyond’ the available data and do not simply ‘emerge’ from it but involve creative insights” (Matthews, 2009, p. 655). Theory building as an endeavor inherently requires creativity and imagination but reflection is also integral to the theory building process (Carey & Smith, 1993); it is this reflective stance that often leads to another round of entrepreneurial activity and scientific advancement.

The scientific “mind-set” illuminated by Lederman (2007) does not have to, however, act in isolation; often, collectively, scientists ‘brain storm’ and work in groups to advance their research. The complexity of scientific research is often too burdensome for individual scientists to pursue and requires collaborative efforts and the intellectual capital of other scientists to make advances (McComas, Clough, & Almazroa, 1998). As alluded to in the NRC’s framework (NRC, 2011), scientific studies are “driven by
curiosity” (p. 3-5) and “scientists constitute a community whose members work together
to build a body of evidence and devise and test theories” (p. 2-3). Critiquing and arguing
amongst scientists is commonplace and part of the theory-building process (NRC, 2011).
Scientific knowledge is built through this human exchange of ideas and reflective
practices; this, often revisionary process, is elementarily creative and at times
argumentative and conflictive. This depiction of the nature of scientific practices fits with
Gauch’s (2009) sixth “pillar” of science that purports the universality and social nature of
science and his conviction that sharing scientific information is vital to enabling science
to advance; and, with Driver, Asoko, Mortimer and Scott’s (1994) characterization of
scientific knowledge construction being socially negotiated.

**Socially & culturally embedded.** The context in which science takes place is not
just within the laboratory or within the minds of individual scientists; the ‘culture’ of
science extends beyond these prescriptive settings and encompasses the larger culture of
the world scientific community, of the media, and of the social and political fabric of
countries in which scientists practice (Lederman, 2007). The NRC (1996, p. 201)
elaborate this theme by stressing that “[S]cientists are influenced by societal, cultural and
personal beliefs and ways of viewing the world. Science is not separate from society but
rather science is a part of society.” The extent that social and cultural elements impact
science knowledge construction is a source of debate and perennially leads to divergent
viewpoints in the literature. These viewpoints are briefly discussed here; a more in-depth
discussion of them is beyond the scope of this dissertation research.
One understanding evident in the literature supports a universal view of science; this view reasons that although social and cultural contexts influence scientific practices and knowledge construction, these socio-cultural constructs do not impact the reality that nature is fundamentally structured, orderly and unchanging (Stanley & Brickhouse, 2001). Social and cultural influences are not relevant as they cannot impose on reality as it is the entities themselves that determine the reality and scientist are only observers and interpreters of this reality (Stanley & Brickhouse, 2001). The counter view is that science is multi-cultural and that there is a social aspect to science knowledge construction with the explanatory framework of science heavily influenced by scientists’ presuppositions, values and cultural beliefs (Lemke, 2001).

Driver, Asoko, Leach, Mortimer and Scott (1994) take a pragmatic view when they suggest that scientists from the same culture (whether it be within a classroom, country or continent) have “shared ways of referring to and talking about particular phenomena” (p. 8) yet the way scientists experience natural phenomena are “constrained by the way the world is” (p. 8). This view is shared by Matthews (2009) who suggests that science is an integral part of, and is affected by culture, but that it also has effects on culture. Scientists are rooted in cultures that are at different stages of social, economic, political, technological, and religious development; these differences can have restrictive or positive influence on scientific practices and knowledge construction dependent upon the context in which scientists are working (Matthews, 2009). Cobern and Loving (2001) extend this perspective by suggesting that there is interplay between the different levels of scientific knowledge construction: at one level, which is infinitely related to human
sensory experiences, is a descriptive explanation of what is observed directly and is largely culturally invariant. The second level is related to theory-building processes and Cobern and Loving (2001) suggest that it is at this level of scientific practice that is most influenced by scientists’ presuppositions, values and cultural dispositions. Lederman’s (Lederman, 1992; Lederman, Abd-El-Khalick, Bell, & Schwartz, 2002; Lederman, 2007) theoretical framework for the nature of science and the perspective taken in this thesis encompasses the view that acknowledges that scientists’ practices are socially and culturally predisposed and that this influence extends to every step of the scientific process i.e., from ideation, explanation and theory-building to presenting and communicating results.

The five domains (Empirical, Theory-Laden, Certainty, Inventive and Social and Culturally Embedded) discussed in this section are deemed relevant to the elementary school environment (Smith, Maclin, Houghton, & Hennessey, 2000; Lederman, Abd-El-Khalick, Bell, & Schwartz, 2002; Lederman, 2007), the target age group of this study. In the following section, the research pertinent to student perceptions of NOS and how well students’ understand NOS will be discussed.

Students’ Understanding of the Nature of Science

This section provides an overview of the status of K-12 students’ understanding of the NOS. It will first discuss the distinction between what is considered a “naïve” understanding of the NOS and an “informed” understanding; terms most often used in the literature to characterize student understanding. Following this discussion, an overview of the status of K-12 students’ understanding of the NOS (formerly, NOS) will be presented.
Studies related to student understanding are divided into two main areas, those that: (1) assess students’ implicit NOS views associated with their science learning or activities; and those that (2) measure student NOS understanding related to a specific scientific inquiry activity or intervention designed to teach the NOS explicitly. These two areas of study will be discussed in turn and encompass research undertaken in K-12 classrooms to assess student understanding of NOS. This section will conclude with an overview of findings from these studies.

**Characterization of students’ perceptions of the NOS.** In order to comprehend the literature herein with regard to students’ perceptions of the NOS, it is important to be cognizant of what is meant when researchers report that students are considered to have a naïve understanding of the NOS and, by corollary, when they have an informed understanding. Table 2.1 provides a summary of the relationship between students’ perceptions and their level of conceptual understanding for each of the five NOS domains. The descriptions provided in Table 2.1 are primarily based on Lederman’s classification of what is considered a novice (naïve conception) or an expert (informed conception) response to VNOS items for each of the five domains (Lederman, Abd-El-Khalick, Bell, & Schwartz, 2002). The decision to predominantly use Lederman, Abd-El-Khalick, Bell and Schwartz’s (2002) article for this part of the study was based on (1) Lederman’s theoretical framework for NOS (Lederman, 1992; Lederman, Abd-El-Khalick, Bell, & Schwartz, 2002; Lederman, 2007) predicated the development of the NOSI-E and (2) Lederman, Abd-El-Khalick, Bell and Schwartz’s (2002) study describes the development of an open-ended qualitative instrument, the VNOS, that is used
extensively in science education research to assess the conceptual level (i.e., naïve, adequate or informed) of student NOS understanding. As such, it is a seminal piece of work with their classification system validated across the science education research literature discussed in this chapter.

Lederman, Abd-El-Khalick, Bell and Schwartz (2002) developed the VNOS to assess respondents’ understanding of NOS; the age of the respondent determines which form of the VNOS they complete. VNOS-B has been used with middle and high-school students (Bell, Blair, Crawford, & Lederman, 2003; Khisfe, 2008; Peters & Kitsantas, 2010; Șahin & Köksal, 2010) with VNOS-D or VNOS-E used with younger elementary-aged children (Lederman & Lederman, 2005; Akerson & Volrich, 2006; Quigley, Pongsanon, & Akerson, 2011; Leblebicioğlu, Metin, Yardımcı, & Berkyürek, 2011). VNOS-D and VNOS-E measure the same aspects of NOS as VNOS-B but the context and readability of the form are made suitable for use at the elementary level (Lederman & Lederman, 2005). Lederman, Abd-El-Khalick, Bell and Schwartz (2002) purposively chose “experts” and “novices” in science and had them respond to the seven open-ended VNOS-B items; their responses are described in their study and tabulated illustrative examples of novice and expert views (p. 514) provide an interpretative, qualitative framework for NOS understanding of each of the NOS domains. This offers researchers a mechanism to categorize students on the NOS continuum, with naïve (novice) and informed (expert) understandings anchoring the ends of the continuum. Akerson and Donnelly (2010) adapted Lederman, Abd-El-Khalick, Bell and Schwartz’s (2002) characterization of students’ NOS understanding and provided a coding rubric and
tabulation (p. 108) for the VNOS-D instrument that was developmentally appropriate for use with K-2 elementary students. These two studies (Lederman, Abd-El-Khalick, Bell, & Schwartz, 2002; Akerson & Donnelly, 2010) provide the principal source of information discussed next and are summarized in Table 2.1.

Lederman, Abd-El-Khalick, Bell and Schwartz (2002) probed students on their views of how scientists know what the structure of an atom is to assess their understanding of the empirical nature of NOS. Understanding the role of empirical evidence in science differentiates a naïve conception of the Empirical NOS domain from an informed conception. Novice respondents understood that empirical evidence through observation of natural phenomena is needed to construct scientific knowledge (Lederman, Abd-El-Khalick, Bell, & Schwartz, 2002). However, novice respondents viewed physical evidence and the data collected from experiments as the sole providers of information; this evidence was needed to make and “prove” objective scientific claims.

At the other end of the empirical NOS continuum, the expert group in Lederman, Abd-El-Khalick, Bell and Schwartz’s (2002) student also understood that empirical evidence through observation of natural phenomena is needed to construct scientific knowledge but, in contrast to the novice group, they intimated that empirical evidence is just one determinant of scientific knowledge and this evidence cannot “prove” scientific claims resolutely. The two groups differed substantially on the role of indirect evidence in developing scientific claims; novice respondents, for example intimated that the structure of atoms was directly observable to scientist and this fact excluded inference, subjectivity or human bias and values in developing a model for the atom. The expert
group explicated the role of indirect evidence and inference when scientist, using their interpretations and personal biases, proposed their model for the structure of the atom (Lederman, Abd-El-Khalick, Bell, & Schwartz, 2002).

Using dinosaurs as the context for questions for their young respondents, Akerson and Donnelly (2010) found a similar pattern of responses in their interviews of K-2 students. Similar to the novice respondents in Lederman, Abd-El-Khalick, Bell and Schwartz’s (2002) study, students with naïve conceptions of NOS believed that scientists observed dinosaurs directly or read about them in books. These students similarly did not perceive that conjecture plays a part in scientific knowledge construction as scientists are “sure” of their findings (Akerson & Donnelly, 2010, p. 108). Young students’ informed responses indicated that they understood the role of empirical evidence and inference in science; for example, they understood that scientists use data “to make claims and create ideas” (Akerson & Donnelly, 2010, p. 108) and that scientist used their observations of dinosaur bones to infer what dinosaurs looked like. It is important to realize that the context and the developmental age of the respondents dictate the language, context and sophistication of the responses. Table 2.1 provides a summary of these relative positions on the Empirical domain of NOS; the descriptors, informed by these studies, are geared toward the expected conceptions of fourth grade elementary students’ epistemologies on the continuum.

Lederman, Abd-El-Khalick, Bell and Schwartz (2002) used scientists’ disagreement on how dinosaurs became extinct to elucidate expert and novice views on the theory-laden nature of science. The novice respondent intimated that the reason why
scientists came to different conclusions was because they were not present to “witness” the extinction of dinosaurs. Thus, they explained, their theories cannot be directly tested with scientists only able to follow objective, observable procedures to formulate their conclusions. Respondents in Lederman, Abd-El-Khalick, Bell and Schwartz’s study (2002) also differed in the need for theories to be supported by substantial evidence and for theories to provide an explanatory framework which scientists can build upon to further scientific knowledge and to generate their hypotheses and research questions. Novice respondents offered the view that theories can change with new evidence adding on to scientists “old ideas” (p. 515); the purpose of theories is to prevent scientists having to “start all over from the beginning” (p. 515). Novice respondents have little or no conception that theories provide scientists with an explanatory framework which can be used to further develop their ideas and hypotheses (Lederman, Abd-El-Khalick, Bell, & Schwartz, 2002).

Lederman, Abd-El-Khalick, Bell and Schwartz (2002) expert respondents offered a relatively more informed perspective on the theory-laden NOS. This is illustrated by the following quote, “Different scientists may come up with different explanations based on their own education and background or what they feel are inconsistencies in others ideas” (p. 516). These differing explanations account for the prevalence of subjectivity in science but also the durability of science. Scientists build upon the explanations of their peers and these explanations are considered and used to design and carry out their own research. As a result, the subjective nature of science is constrained and a coherent, durable framework is built to guide scientists’ future research.
Akerson and Donnelly (2010) used the same dinosaur context to garner their young students’ views on the theory-laden nature of science. In coding this aspect of NOS, Akerson and Donnelly (2010) reported that a naïve conception for young students would be indicated by students suggesting that once scientists have more data, they would all agree. This conception is remarkably similar to the novice respondents in Lederman, Abd-El-Khalick, Bell and Schwartz’s study, 2002. Young students with informed conceptions in Akerson and Donnelly’s study (2010) understood that scientists differ in background and prior experiences and these could lead to disagreement over how to interpret the same data. Akerson and Donnelly (2010) did not delve into assessing young students’ understanding of how theories are durable and involve a coherent explanatory framework; this was likely due to the developmental age of the K-2 students. In Smith, Maclin, Houghton and Hennessy’s study (2000), sixth grade students were able to express developmentally appropriate informed views on this aspect of the theory-laden nature of science. For example, they understood that scientists’ ideas needed to “fit with other ideas” or “fit with a pattern of evidence” (Smith, Maclin, Houghton, & Hennessy’s, 2000, p385); these findings suggest that these students were beginning to realize the need for ever deeper coherent explanations of natural phenomena and this aspect can be assessed in late-elementary students. This more informed perspective is included in the summary table; Table 2.1.

Lederman, Abd-El-Khalick, Bell and Schwartz (2002) assessed students’ views on the tentative nature of science (Certainty domain) by exploring novice and expert respondents’ views on the nature of theories. The authors found that novice respondents
believed the scientific process is governed by a singular objective, namely to determine if something is right or wrong. This naïve perspective extends to the view that if a scientist repeats an experiment and consistently gets the same results, that the theory is factual, proven and a replica of reality. Expert respondents in Lederman, Abd-El-Khalick, Bell and Schwartz’s study (2002) viewed science in general and theory formation in particular as an inherently subjective process and this is one factor that predicates the tentativeness of science knowledge construction. Subjectivity often results from scientists offering different explanations and perspectives on the same data and these differences in interpretations can make science an uncertain endeavor.

Akerson and Donnelly (2010) used a familiar context, the weather, to assess young elementary students’ views on the tentativeness of science (Certainty domain). Students with naïve conceptions of the Certainty domain indicated that scientists are sure of their forecasts because they have analyzed real and concrete data to make them; there was no acknowledgement that new data may result in meteorologists having to change their forecasts. Students in Akerson and Donnelly’s study (2010) who held informed views of the tentativeness of science understood that knowledge is not certain and meteorologists may get new information that makes them change their forecasts; the informed perspective further reasoned that meteorologists may have to change their predicted forecast by reinterpreting existing data. These two developmental understandings of the tentativeness of science are summarized in Table 2.1.

To have an informed perspective of the creative and imaginative (Inventive domain) nature of science, respondents needed to first believe that scientists use
creativity and their imagination in their work but must also understand that they use both in all aspects of their work whether it is to develop research questions and ideas or to interpret their data (Lederman, Abd-El-Khalick, Bell, & Schwartz, 2002). The novice respondents in the study agreed that scientist might use their creativity to develop ideas but once they begin their investigations, it is not needed. This relatively uninformed understanding is related to their view that scientists use a single scientific method in their work; as a result, there is no need for scientists to use their creativity or imaginations because the scientific method employed will determine if their ideas are correct. In contrast, the expert group understood this informed conception of the Inventive domain of NOS believing that “creativity permeates the scientific process” (Lederman, Abd-El-Khalick, Bell, & Schwartz, 2002, p. 508).

Young elementary students possess a similar taxonomy of views on the Inventive nature of science (Akerson & Donnelly, 2010). Students with naïve conceptions in this study indicated that scientists do not need creativity or their imaginations to do their work because they can look at and see what they are doing. In contrast, students with informed conceptions realized that scientist may have to use their creativity and imaginations because scientists will not always be sure of what their results mean and need to use their imaginations to “figure it out” (Akerson & Donnelly, 2010, p. 112) and develop explanations for their findings. Table 2.1 portrays the anchors for students’ views for the Inventive domain that late-elementary students could harbor.

Lederman, Abd-El-Khalick, Bell and Schwartz (2002), in assessing respondents’ views on how science is socially and culturally embedded, found that naïve respondents
had no conception of how social and cultural contexts influences the construction of scientific knowledge. The expert group acknowledged the importance of social and cultural contexts in constructing knowledge in science and identified two areas where their impact is felt. These respondents emphasized that science has a culture in of itself. Scientists make scientific claims that are peer reviewed and these claims are often argued, defended and in need of justification. Over time, hopefully, a consensus is reached and the subjective nature of science is constrained. As a result, scientists’ different backgrounds, experiences and training affect their perspectives they bring to this process. The scientific endeavor is also influenced by the societies and cultures in which scientist work. Politics and religion can affect how and what scientists can work on with funding for their work impacted by the strength of the economies in which they reside. Expert respondents understood the role of society and culture play in the scientific enterprise (Lederman, Abd-El-Khalick, Bell, & Schwartz, 2002); in contrast, novice respondents were not cognizant of its importance in understanding NOS.

Akerson and Donnelly (2010) did not assess this aspect of NOS directly but insight into students’ views of this domain of NOS can be elicited in their discussion of students’ views on the subjective nature of science. Students with informed conceptions on the subjective nature of science did not formally suggest that scientists’ differing backgrounds and experiences led to disagreements and argumentation but they implied this when, for example, one stated, “they all think their way of seeing stuff is right, so they argue” (Akerson & Donnelly, 2010, p. 117). Those that held onto a naïve conception of the subjective nature of science indicated that scientist should provide identical
interpretations for the same data; this view suggests congruence with the novice respondents’ views of Lederman, Abd-El-Khalick, Bell and Schwartz’s study (2002). Table 2.1 provides a summary of the coding used by Lederman, Abd-El-Khalick, Bell and Schwartz (2002) and Akerson and Donnelly (2010) to characterize student views of the five aspects of NOS used in this study. The naïve conception and the informed conception anchor the spectrum of students’ NOS views found in the literature, and are presented to guide readers in the review of the literature that follows.

**Student understanding of the nature of science through implicit instruction.**

Historically, the assumption was that students learn about NOS implicitly by engaging in science-based inquiry activities within the classroom (Abd-El-Khalick, Bell, & Lederman, 1998). As mentioned in the previous chapter, the AAAS (1990) recommends that the study of science as a way of knowing needs to be made explicit in the curriculum. In this section, a summary of the studies pertaining to students’ implicit understanding of the NOS is provided. This is followed by a section devoted to studies that examine the impact of teaching NOS explicitly to students as a part of their science learning. These two areas of study provide context to this study and communicate the different perspectives taken to help researchers assess student understanding of NOS.

Ryan and Aikenhead (1992), in a national survey of Canadian high school students’ epistemology views determined that the vast majority of students held naïve views on several aspects of the nature of science. Less than a quarter of high school students in the study realized that scientific theories were created and built from ideas of
Table 2.1

*Characterization of Students’ Perceptions of NOS*

<table>
<thead>
<tr>
<th>NOS Domain</th>
<th>Naïve Conception</th>
<th>Informed Conception</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Empirical</strong></td>
<td>Science is viewed as observing and describing nature and conducted principally through the use of hands-on methods [experiments]. There is little conception that scientists are collecting evidence, attempting to explain their observations or making inferences about entities that are not directly observable. In order to be sure of something, it has to be observable and seen to be believed.</td>
<td>Scientific knowledge is based on evidence derived from observations and from the inferential process. Observations are used to explain the real-world around us and students understand that evidence is needed to guide this process. Similarly, they understand that scientists have to make inferences and conjectures on the natural phenomenon as many are not observable to scientists’ senses.</td>
</tr>
<tr>
<td><strong>Theory-Laden</strong></td>
<td>Students with naïve conceptions believe theories are “educated guesses”. There is minimal conception that science involves building a complex explanatory framework. A naïve conception expects scientists will follow procedure and find the “right answer” if they persevere. Any disagreements between scientists are due to some scientists having made mistakes in carrying out their experiments and procedures.</td>
<td>Students with an informed conception understand that the goal of science is to provide ever deeper coherent explanations of natural phenomena by building on prior evidence, knowledge and observations and on new evidence. Scientists may contest theories based on new evidence. An informed perspective understands that evidence obtained may equally affirm or contradict current theories resulting in an inherently subjective process.</td>
</tr>
<tr>
<td><strong>Certainty</strong></td>
<td>Science is perceived as factual and objective. Science is viewed as unchanging with little or nothing new to discover. It is concrete in that we can get the same answer to an experiment, no matter who conducts it. A naïve conception would suggest that, provided enough evidence, the knowledge constructed is ensured, irrefutable and an exact replica of reality.</td>
<td>Students with informed conception grasp that science is tentative and that scientist may have to change their ideas or theories when provided with new evidence, new explanations or improved scientific techniques to investigate problems. Similarly, scientists may reinterpret existing findings in light of new discoveries; new theories evolving or scientific claims. As a result, science is durable but subject to change.</td>
</tr>
<tr>
<td>NOS Domain</td>
<td>Naïve Conception</td>
<td>Informed Conception</td>
</tr>
<tr>
<td>---------------------</td>
<td>------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------</td>
<td>----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------</td>
</tr>
<tr>
<td>Inventive</td>
<td>Science is objective and absolute so creativity and imagination are not required to conduct science. The scientific method is empirical; universal; routine; and does not vary resulting in knowledge that is factual and unchanging.</td>
<td>Creativity and imagination is needed at every step of the scientific process; from developing ideas and theories (inferred explanations), to designing methods to test ideas and theories, to interpreting the results, and to communicating the results. No single method can account for diversity of findings and success of science; it is the creativity of scientists’ minds that provide the foundation for theory building and scientific discoveries.</td>
</tr>
<tr>
<td>Socially and Culturally Embedded</td>
<td>Scientists exchange factual information with each other and work together to perform concrete activities and/or experiments. Prior knowledge or experiences (social or cultural) are irrelevant and have little influence as science and the scientific process is dogmatic and in pursuit of the universal and certain truth.</td>
<td>Scientists prior knowledge and social and cultural experiences influence how, why and what scientists work on. Science is an inherently human endeavor with scientists influenced by their social and cultural backgrounds. The social and cultural milieu in which scientists perform impacts the way they conduct science, interpret evidence and accept other scientists’ ideas and theories.</td>
</tr>
</tbody>
</table>

Source: Author developed
scientists with the vast majority believing that theories were simply waiting to be “discovered” or accidently happened upon. The majority of these students were unaware that theory formation was subjective, inventive and involved scientists building a consensus. Solomon, Scott and Duveen (1996) found similar results in a study of Year 10 (grade 9) students in the United Kingdom where very few students (~15%) understood that scientists have to use their imaginations to develop theories and even fewer students held the informed conception that theories support scientists in making predictions in their scientific investigations. Most students’ views characterized theory formation as “blind idealism” in Ryan and Aikenhead’s study (1992, p. 575), where scientists are considered completely objective and their personal values or traits take no part in the construction of scientific knowledge. In line with this finding, most students in this study also dismissed the role of contextual values (culture, religious, ethical) in how scientists create scientific knowledge with many relying on the naïve view that scientists need to be factual and objective. This, in turn, will eventually lead to scientists making the same discoveries as each other. In probing students’ views on the tentative nature of science, students appear to have the most informed views with many students believing that scientific knowledge does change over time. However, the student reasoning provided for their beliefs was naïve with many either holding a falsification’s position that scientists work to disprove theories making the process inherently tentative or of the view that science is uncertain because new knowledge simply adds on to what was discovered in the past. Few students reasoned informatively that scientific facts themselves can change in light of new interpretations or conceptualizations (Ryan & Aikenhead 1992).
In a large-scale, representative national study of Turkish students’ epistemological views, Dogan and Abd-El-Khalick (2008) determined that, in general, the majority of students possessed naïve views of the NOS. However, using 14 items from Aikenhead and Ryan’s VOSTS instrument (1992), Dogan and Abd-El-Khalick (2008) found that students’ epistemology did vary across the different domains of the NOS. Similar to the study above, students held more informed views on the tentative nature of science with a majority of students (68.2%) rejecting the view that if done correctly, the results of an experiment are “unchangeable facts” (p. 1109) and endorsed the view that “old knowledge is reinterpreted in light of new discoveries” (p. 1109). In contrast, they held, on average, moderately informed views that to perform science, scientists should use any method (“including the method of imagination and creativity”, p. 1110); with two-thirds of the students holding the naïve view that theories are simple and do not involve a complex explanatory framework. Similar naïve epistemologies were evident in Kang, Scharmann and Noh’s study (2005) of 1702 Korean students’ NOS views; in this study, students’ views were essentially realist (i.e., scientific knowledge is irrefutable and an exact replica of reality). Although students views did vary across domains (students had more informed views on what a scientific theory was, than on the creativity needed to develop these theories); these largely naive conceptions did not differ across the developmental levels measured (elementary, middle and high school levels).

The results from a large-scale study of elementary students’ views on the NOS corroborate these previous studies in that they found students’ levels of NOS knowledge varied across the three aspects of NOS examined (Huang, Tsai, & Chang, 2005).
Elementary students held, on average, more sophisticated views on the invented and changing nature of science and less informed views on the role of social negotiation and on the cultural context of science. Akerson and Abd-El-Khalick (2005) correspondingly examined fourth grade students’ perspectives towards three aspects of NOS; the distinction between observation and inference [Empirical domain]; the creative and imaginative nature of science [Inventive] and the tentative nature of science [Certainty]. This qualitative case study supports the findings of the larger quantitative studies highlighted above. Fourth grade students, as a whole, held naïve views of the NOS but appeared to have relatively more informed views on the tentative nature of science than those they held on the creative and imaginative nature of science or on the role of observation and inference in performing science. The informed view of the tentative nature of science was expressed by one fourth grader in the quote, “Well, you find evidence all the time, so science changes” (Akerson & Abd-El-Khalick, 2005, p. 7). Similar to Ryan and Aikenhead’s study (1992), Akerson and Abd-El-Khalick (2005) opined that students’ more informed views were falsely premised on the belief that new findings simply “added on” to previous discoveries and not that new scientific claims could change facts and theories completely. The studies described to date provide a “snapshot” of students’ view on NOS. There are several studies that have been conducted to study students’ views that are associated with specific scientific activities or interventions; in addition, these and other “implicit” studies involved measuring student understanding at two time points. These studies are discussed next.
In a case study of five high school students, Moss, Abrams and Robb (2001) studied students’ NOS views within the context of a project-based classroom environment; these students were purposively chosen to span the range of scientific ability within the grade. Uniquely, the students were allowed to go off campus four times during the year to work on projects with scientists in the field. Students initially held relatively more sophisticated views on the tentativeness, creativity and social embeddedness of science with students’ tentativeness views becoming even more informed over time. Students intimated that science is always changing (a relatively sophisticated view) but, by the end of the study, they were able to articulate the more informed view that this revisionary process could involve re-interpreting “old knowledge” (p. 785). Similar to the status studies above, students were more open to changing their views of this domain. Most participants confirmed that they thought of science as a social activity and that it is affected by societal needs; they were less united in their interviews on the role of scientists’ backgrounds and differing perspectives plays in how scientists perform their work. Similarly, from the outset, the authors reported that the students understood the need for scientists to be creative and imaginative to perform their work; however, it was not clear from the reading if their view was limited to the need to use these elements for ideation or extended beyond to using them in all aspects of the scientific process (e.g., making inferences, analyzing and presenting results) (Moss, Abrams, & Robb, 2001).

On the whole, students held relatively naïve epistemologies on the empirical and theory-laden nature of science with students limiting their views of science as being
based on the scientific method and a “prescribed set of procedures” (Moss, Abrams, & Robb, 2001, p. 779) which are used to conclusively prove scientific facts. Four of the five had no conception that science involves conjecture and prediction of phenomena. Similarly, they had little awareness of the notion that science is theory-laden involving the development of an explanatory framework which can lead to new questions being developed. These views essentially remained constant over time but one student’s epistemology appeared to take on a slightly more informed perspective as he reported that some scientific questions “may not be fully answerable” (p. 784), indicating a shift in his thinking concerning this element of the NOS.

Sandoval and Morrison (2003) assessed students’ theory-laden and certainty domain views associated with a four-week evolution and natural selection unit. The unit encompassed a combination of labs; computer-based simulated activities specially designed for the unit; and extensive in-class discussions intended to stimulate theory-building and explanations for the phenomena observed. Similar to Moss, Abrams and Robb (2001), a cross section (based on abilities level) of students were interviewed prior to their unit and then again upon completion. Although the high school students understood, prior to the unit, that the purpose of science is to explain things and that experiments are used to test out scientists’ ideas; the perception that pervaded pre- and post-unit interviews was that ideation and the testing of these ideas were both related to a known absolute truth. Similarly, they perceived theories to be hypotheses that have been proven due to extended testing of an idea with few, even post intervention, understanding that theories “encompass explanations in development that can be advanced over time in
the light of new data” and can kindle further ideation or differing perspectives (Sandoval & Morrison, 2003, p. 380). Their relatively naïve epistemology was not affected by the inquiry-based activities despite the attempt of the teacher and researchers to provide an authentic inquiry-based science learning experience.

Bell, Blair, Crawford and Lederman (2003) selected 10 high school students participating in a 8 week science and engineering mentored apprenticeship program to study high school students’ views on NOS; this apprenticeship enabled students to work in a laboratory full time and perform genuine research. Students’ views on NOS (all five domains relevant to this study) were measured using the VNOS-B prior to the apprenticeship and upon completion. Additional information was gathered by the researchers through observation of their activities and through informal discussions. The researchers’ findings are consistent with those of Sandoval and Morrison (2003) in that students expressed the naïve view that science is resolutely empirically based with no understanding that theories could change with scientists reinterpreting existing evidence or data with a new, creative perspective. Creativity was compartmentalized to the beginning of the scientific process (ideation and choice of method) with no students perceiving that it also played a role in the interpretation of data and how the results connected with the current explanatory framework(s). In assessing the SCE domain of science, no students expressed an informed view that the social and cultural contexts of science can influence how scientists perform their work. These views essentially remained intransigent over the course of the apprenticeship with only one of the ten students endorsing a more informed view of how theories can change and compete with
each other and the role of creativity in analyzing and reporting results. This student professed to be very reflective of her discussions with her mentor, which the authors suggested contributed to her changed viewpoints.

A nine-week hands-on, inquiry-based unit on the chemical properties of substances was the context for Conley, Pintrich, Vekiri and Harrison’s (2004) study of elementary students’ epistemological beliefs. These authors did not use the conceptual framework for NOS as outlined in this study; however, the pre and post-test (self-report survey) results did suggest that students’ understanding of NOS improved slightly by the end of the unit across the four domains measured. Similar to the studies of high school NOS views, elementary students’ views, on average, appeared to improve their understanding that scientific knowledge is tentative and that not “all questions have one right answer” (p. 202). The remaining domains studied by Conley, Pintrich, Vekiri and Harrison (2004) possessed items that cut across the conceptual domains of this study so the epistemological improvement observed could not be broken down succinctly; regardless, the inquiry-based activity did appear to help students form more informed views of NOS.

In a recent study, Wu and Wu (2011) examined fifth grade epistemologies related to a 5 week inquiry-based activity on force and motion. The activities were designed to improve students’ explanation skills and provide the setting for investigating whether students’ understanding of the nature of scientific knowledge changes over the course of the unit. Self-report instruments and semi-structured interviews were used to assess students’ views. In many instances, students’ views became more naïve over the course
of their activities; more students took the view that doing experiments is the only goal of science. Although fewer students professed that scientific knowledge is only constructed through experimental results post-activities, they switched their opinions to an equally naïve view that scientific knowledge is concrete and derived from textbooks.

However, this limited view of scientific knowledge did not prevent students from improving their epistemologies with regard to the nature of scientific questions [Empirical domain]. The majority of students on the pre-tests and in the pre-interviews expressed no understanding of the types of scientific questions scientists may ideate with most of the remaining indicating that they are posed to describe what is observed. In the post-assessment, the students held more informed views in that the majority of the students felt that scientific questions were used to see “why something happens” or “how something works.” This improvement in epistemology did not translate, however, into improved understanding that scientists use different methods to perform their work and when asked about their work, more students expressed that a single method is needed to perform experiments post-activities than pre-activities (a decline in student understanding). Similarly, students did not improve on their understanding that experimental results could be problematic, relying on the view that if the experiment was done “right,” the results are irrefutable [Certainty]. Only one student out of the sixty eight expressed a more informed view of, “although my group’s result is different from others, both of them could be accurate” (Wu & Wu, 2011, p. 335).

Most students suggested that experiments needed to be repeated if results did not agree and this would lead to the “correct” solution rather than trying to find alternative
explanations for their conflicting results (Wu & Wu, 2011). As the students were unable to articulate the interplay between ideas, evidence and explanations [Theory-laden], these findings are perhaps understandable and support the conclusion by the authors that, in general, students’ epistemologies were very basic. The authors indicated several instances where the teacher missed the opportunity to have students reflect on what they had learned and for the teacher himself to engage and scaffold discussions about the nature of scientific knowledge; these missed opportunities were purported to explain the limited improvement in students’ epistemologies (Wu & Wu, 2011).

In Smith, Maclin, Houghton and Hennessey’s (2000) extensive and seminal research, these authors found that student reflection combined with constructivist teaching pedagogy were instrumental in improving students’ NOS epistemologies. Two classrooms were compared in this study; one classroom used a constructivist teaching approach while the other followed a traditional approach. The goal of the constructivist teaching was for students to take on the responsibility of developing scientific questions and ideas to explore and test; reflect and make meaning of their ideas and findings and take part in in-depth dialogue to explain and, if necessary, revise their ideas. The teacher in the study had taught science to her sixth grade class since first grade using this teaching approach. This teaching approach, designed to engage students in authentic scientific practice, is shares similar characteristics to the one espoused by the NRC (2011) in the new science frameworks. In contrast, the comparison classroom’s teacher (who also had taught her students for the last five years) used a more traditional approach where activities were prescribed (often leading to only one possible solution) and were
targeted at students’ factual content knowledge. Although this was essentially a qualitative study, the authors quantified (using Carey’s Nature of Science Interview protocol and scoring rubric) and compared the results the two classrooms.

Student epistemologies in the comparison classroom were similar those found in Wu and Wu’s (2011) study of fifth graders; for example, students expressed that science and experiments were predominantly concerned with “gathering information” and “finding answers” to procedural (e.g., how to do something); journalistic (e.g., who, what and where and when) and/or explanation questions (e.g., how something works). Students in the constructivist classroom predominantly expressed that while one of the goals of science (and experiments) was to gather information, it was also to understand, test and develop ideas. Their more sophisticated understanding of the empirical and theory-laden domains of science extended to the types of questions scientist explore with the majority (61%) indicating that scientists answer explanation questions; theoretical questions (e.g., about entities that are unobservable; 78%) and metacognitive (e.g., why am I doing this experiment?; 72%) questions.

The consequence of constructivist students’ appreciation that scientists often pose complex questions was their more informed understanding of the processes needed to answer these types of questions. In contrast to the majority of students in the comparison classroom, the majority of students in the constructivist classroom understood that hypotheses need to be ideated and tested; observations and inferences need constructed [Empirical; Theory-Laden] and that there is a relationship between the evidence collected and the inferences that can be made [Theory-laden]. Similarly, students’ conceptions in
the constructivist classroom went beyond those of students in the comparison classroom (whose predominant conception was that scientists need to find out if their ideas are right or wrong) to a more sophisticated epistemology of realizing that their ideas needed to “fit with other ideas” or “fit with a pattern of evidence” (Smith, Maclin, Houghton, & Hennessey’s, 2000, p. 385); these findings suggest that students within the constructivist classroom were beginning to realize the need for ever deeper coherent explanations of natural phenomena. No students in either class could grasp that theories may limit the ideation process and interpretation of results (Smith, Maclin, Houghton, & Hennessey’s, 2000).

When queried on the role of social interaction [Social and Culturally Embedded], students in both classrooms equally understood that scientists exchange information and ideas but students in the constructivist classroom expressed significantly more often that scientists try to understand and discuss explanatory ideas and influence each other’s interpretations of evidence. With more informed perspectives on the theory-laden nature of science, students in the constructivist classroom also had a better sense of the subjectivity and tentativeness of science (Smith, Maclin, Houghton, & Hennessey’s, 2000). A majority of students in the comparison classroom indicated that scientists will “abandon” an idea if an observation or experiment provided conflicting evidence with only a third acknowledging that scientists may have to think or put more effort in before changing their ideas. In contrast, the majority within the constructivist classroom felt that complex evidence is needed before scientists would change their ideas with the process driven by the need for better explanations. As one constructivist student intimated when
discussing with the interviewer the need to change an idea, “Because if your ideas are that strong you can’t just change them. You have to go at it a different way” (Smith, Maclin, Houghton, & Hennessey, 2000, p. 386) and this different way the student suggested may “make it obvious to you that you have to change your ideas” (p. 386). In stark contrast to students in the comparison classroom, some students in the constructivist classroom (as opposed to none in the comparison classroom) conceived that the subjective nature of science is constrained by prior ideas, indicating a high level of NOS epistemology.

**Conclusion.**

In reviewing the “implicit” studies in this section, the authors, for the most part agree with Akerson & Abd-El-Khalick’s (2005) and Ford’s (2008) views that students do not inculcate informed NOS views by simply taking part in inquiry-based scientific activities, especially if they involve “declarative knowledge” (how to directions) and even if they have a teacher with informed views. Ford (2008) opines that NOS understanding should be viewed as an “active ability” (p. 172) rather than “passive declarative knowledge” (p. 172) and suggests that even if NOS is taught explicitly within the classroom, this may not necessarily be sufficient to “translate their grasp into an explicit, declarative form” (Ford, 2008, p. 172) that can be assessed.

**Student understanding of the nature of science through explicit instruction.**

This next section examines studies that have tried to teach NOS explicitly during students’ science activities to determine if, by doing so, students can improve their epistemologies. Explicit instruction of NOS involves the teacher taking an active role in
incorporating NOS learning activities into their lesson plans. These activities were either stand-alone (non-integrated) or embedded (integrated) into their units of study. The studies examined here include young students (as early as kindergarten) to high school students; take place within different settings (science camps; school classroom and after school programs); and try to identify the factors that affect students’ internalization of NOS understanding.

Khishfe and Abd-El-Khalick (2002) compared the outcomes of a group of sixth grade students who received instruction of NOS explicitly to a group of students where they were expected to learn NOS implicitly through the same guided inquiry-unit used in the study. At the end of each of the six activities in the unit, the explicit group took part in scaffolded discourse and reflection that focused on the NOS domains that were associated with each activity. Students’ understanding of NOS was assessed using an open-ended questionnaire and semi-structured interviews; pre-intervention, both groups had roughly similar NOS views with 85% of both groups having a naïve understanding of NOS (between 3 and 9% informed views). For each of the four NOS domains (Certainty; Observation vs. Inference; Empirical; and Inventive) assessed, students’ NOS views in the implicit group remained constant with only a slight improvement (7% informed view pre-intervention to 18% post-intervention) in the “Observation vs. Inference” aspect.

In contrast, post-intervention, in the explicit group, the percent increase in the number of students with informed views significantly improved by 46% for the tentative nature of science; by 42% for the empirical nature of science; 31% for both the “Observation vs. Inference” domain and inventive nature of science. Of the four NOS
domains, students taught NOS explicitly had a hard time understanding that creativity and imagination [Inventive] are needed in science and even this level of understanding was context dependent with students having a better understanding of its role when discussing a dinosaur activity than when discussing an atomic structure activity (Khishfe & Abd-El-Khalick, 2002).

These findings were replicated in a study of Turkish science camp students’ epistemologies; these students showed substantial improvement in their views on the empirical and tentative nature of science but were reticent in their views on the creative and imaginative nature of science, and of the theory-laden nature of science (Leblebicioğlu, Metin, Yardımcı, & Berkyürek, 2011). Of note in the Leblebicioğlu, Metin, Yardımcı and Berkyürek’s study (2011) was that among students who were willing to accept that science involves creativity, there was a lack of awareness of what phases of science inquiry required scientists to use their imaginations and creativity. These two non-experimental, descriptive studies suggest however, that treating NOS understanding as a cognitive endeavor may help improve students’ NOS understanding.

Chuy et al. (2010) conducted a “natural experiment” where they compared two classrooms (in separate schools) who conducted science similarly (both pedagogies allowed for significant student decision-making in what, why and how science activities were performed). One “treatment” differentiated their approaches; in the experimental classroom, the focus of pedagogy was ‘Knowledge Building’ with a strong emphasis on theory-building and understanding the theory-building process. Students in the experimental classroom also had access to a computer program “Knowledge Forum” that
scaffolded the theory-building approach for students and stimulated student-led reflective discussions (on-line and in class) on their scientific activities. Students in both fourth grade classrooms had been exposed to their respective teaching approaches for two years and both undertook an identical 4 month unit on light. Chuy et al. (2010), using Carey’s Nature of Science Interview (Carey, Evans, Honda, Jay, & Unger, 1989), examined students’ overall NOS understanding and Empirical, Theory-laden (two aspects) and Inventive domain understanding before and after the science unit. With the exception of the Inventive domain, students in the experimental classroom had more informed views of the empirical and theory-laden nature of science (pre- and post-). In the Inventive domain both groups did not differ significantly before or after the activities but both groups improved their understanding of the Inventive domain (both groups could articulate that creativity is needed to develop new ideas). In both classrooms, the gain scores were similar for the aggregate NOS measure but the experimental classroom’s students began their science inquiry activity with, on average, a more informed baseline measure. The authors concluded that the sustained use of the Knowledge Building (the treatment to help students understand the theory-building process) supported by the technology scaffold led to more informed views within the experimental classroom with students understanding the role of ideas, evidence, and need for coherent explanations to build a theory that can serve to predict phenomena.

In a mixed-methods study examining the impact of using scaffolded NOS metacognitive prompts embedded within students’ science activities, Peters and Kitsantas (2010) similarly determined that these metacognitive prompts (designed to explicitly
support student NOS understanding) improved student NOS views. The metacognitive prompts were in the form of checklists and questions used to draw students’ attention to different aspects of NOS. Peters and Kitsantas (2010) measured student NOS views (VNOS-B; Lederman, Abd-El-Khalick, Bell, and Schwartz, 2002) under two conditions. Two classes (experimental) were exposed to the embedded prompts throughout their eighth grade electricity and magnetism unit; two comparison classes were provided instruction on the same unit with no explicit instruction on NOS. The prompts were metacognitive in nature and intended to have students reflect on aspects of NOS directly related to the module of instruction within the unit of study. Using VNOS-B, the authors found significant differences in student NOS understanding post unit (differences between the comparison and experimental groups were not significantly different at the start of the unit) with the experimental classes having more informed views. To provide context for these results, the authors conducted interviews; think out loud and focus groups; the qualitative results suggested that the comparison class students were motivated to learn content and felt that science was conducted in labs and was the series of steps needed to obtain the “right answer” (p. 393) to a question. The experimental group, in contrast, viewed science as trying to explain phenomena [Theory-Laden] and involved scientists providing evidence [Empirical] and justification [Socially and Culturally Embedded] for their ideas and theories.

Reflection and metacognition were important components of a case study designed to investigate theory building epistemology in third and fourth grade students (Kawasaki, Herrenkohl, & Yeary, 2004). Students designed activities to test and report on
whether different objects would sink or float and then became involved in discussions related to students’ ideas and ways of thinking. Researchers used discourse analyses and observation of oral reports as the primary vehicles to obtain the nature and quality of students’ views on theory building; discussions were either small group-based or whole class discussions guided by the teacher. Initially, student discourse suggested that students held naïve views on theory formation and could not distinguish between a theory or explanation and the actual observation of the phenomenon (what they observed was the explanation). As the activities evolved and through the questioning of each other and by the teacher, students’ epistemology shifted toward a more relational point of view with students beginning to realize that the behavior of the entities (different objects) they observed was interrelated and required them to search for an explanation divorced from their observations. Naively, at this stage, they expressed that their one emerging theory (that an object sinks or floats is dependent on its weight) would suffice even though evidence seemed to conflict with their explanation. It was not until they began to explore the association between theories and evidence that students began to understand that some of their evidence was not coherent with their theory and another theory might be warranted. Through these guided experiences and discussions, students at this young age were able to develop more sophisticated epistemologies on the theory-building process.

Carey, Evans, Honda, Jay and Unger (1989) developed a three-week long NOS unit linked to student-led and teacher-led science activities that were designed to help the seventh grade students understand the nature of yeast. The first week was spent on scaffolded activities in which the teacher led reflective discussions on students’ scientific
inquiry and thinking processes; students were then instructed to develop experiments that would test their ideas as to why bread rises. Guided and reflective discussions were used to help students understand the nature and purpose of science and scientific ideas/questions [Empirical; Theory-Laden] and the nature of scientific results [Certainty]. In all aspects, students significantly improved their epistemologies with, on average, epistemologies improving from Level 1 understanding to between Level 1.5 and Level 2.0 NOS understanding based on Carey, Evans, Honda, Jay and Unger’s (1989) three-level scale. Students made their greatest gains understanding that scientific questions and ideas form the basis for exploration [Empirical] and moved to a more sophisticated stance in understanding that scientists’ ideas may have to change [Certainty] due to the results of experiments. Although students made gains in their understanding that a goal of science is to provide ever deeper explanations of phenomena [Theory-Laden]; their gains were relatively small compared to the other aspects of NOS measured. Through constructivist practices, Carey, Evans, Honda, Jay and Unger (1989) suggest that “by engaging students in reflecting upon the relationship between ideas and the activities of science” (p. 527); it is “possible to move them beyond their initial (naïve) understanding” (p. 527) of NOS.

The importance of scaffolding by the teacher to improving student NOS understanding is apparent in reviewing Tao’s research (Tao, 2003) on using science stories to teach dyads of seventh grade students NOS. Students read stories (for example, on Alexander Fleming’s discovery of penicillin) and discussed in their pairs the content of the stories. The stories were chosen to explicitly cover three aspects (Empirical, Theory-Laden and Inventive) of NOS; these aspects were integrated within the stories
and founded the pre and post-test content (multiple-choice test). The teacher did not scaffold or guide the dyad discussions and only led a whole class discussion on providing students feedback on their performance post-test. Tao concluded from the test and follow-up interviews that without critical guidance from the teacher, students held on to their realist views of NOS by choosing selective elements of the stories to argue and confirm their naïve viewpoints. Therefore, despite students taking part in a critical aspects of the scientific process (collaboration and argumentation); they were not able to improve upon their naïve views of NOS. Tao suggested that to improve student epistemology, “the teacher should also actively scaffold students’ understandings” (Tao, 2003, p. 169) and this scaffolding should provide for time to reflect on their understandings.

Tao’s (2003) conclusion is supported by Khisfe and Lederman (2006) whose study confirms the need for explicit reflective discussions of NOS to enhance student NOS understanding. Khisfe and Lederman’s study (2006) compared two different teaching approaches to improve NOS understanding in ninth grade students. In one approach, NOS was integrated into the science content (global warming unit) and made explicit by relating NOS to the content of the lessons. The alternative approach was to teach NOS intermittently throughout the global warming unit; in this non-integrated approach, the teaching of NOS was similarly explicit and related to content but taught through activities designed to teach NOS separate from the content. Both approaches focused heavily on reflective discussions to help student understanding. When comparing the results from the two approaches, both were found to improve student epistemology in
the four aspects (Empirical; Theory-Laden; Certainty; and Inventive) of NOS measured. Interestingly, the integrated approach appeared to preferentially help students move from an intermediate NOS pre-unit understanding to an informed understanding whereas the non-integrated approach was more successful in shifting students from a naïve NOS understanding to an intermediate understanding in all four aspects. The non-experimental and limited number of participants in this study prevents the authors from recommending one teaching approach over the other; however, as the previous studies have shown, using explicit, reflective cognition appears to support improvement in student NOS understanding.

In a more recent study, Khisfe (2008) extended the work of the 2006 study (Khisfe & Lederman, 2006) by investigating the processes by which seventh grade students transitioned from a naïve NOS perspective to a more informed view when exposed to an intervention that explicitly integrated NOS teachings into their comprehensive inquiry-based activities associated with their biology unit. As in Khisfe’s previous study (Khisfe & Lederman, 2006), the teacher used reflective discourse following each of the three extended activities in order for students to build an understanding of the related NOS. In this study, Khisfe (Khisfe, 2008) measured student NOS understanding pre, mid and post-intervention and as in the previous study, found that student understanding improved over the course of the intervention. Khisfe (2008) postulates that the transition from naïve NOS understanding to an informed NOS perspective is one that shifts student epistemology along a developmental continuum. It involves a transition phase that can include fluctuating student views that can be unstable
and context dependent and one that can include multiple co-occurring and, at times, conflicting levels of understanding. For example, one student provided an informed view on the distinction between observation and inference as it related to how scientists developed the structure of atoms but reverted to a naïve view when explaining how scientists were able to know the color of dinosaurs.

Khisfe’s (2008) study sustains the finding that explicit reflective instruction improves student epistemologies but also provides evidence that students’ NOS understanding would be further enhanced if it was taught in multiple contexts and reinforced repeatedly so students can inculcate their NOS understanding. Liu and Lederman’s (2002) study of Taiwanese gifted students suggests the importance of having students exposed to explicit and reflective NOS instruction where concepts are continuously reinforced. The authors partially attribute the lack of improvement in students’ NOS views to the short duration (1 week science camp) of the camp and insufficient time given to reflective discussions on NOS. The sustained reinforcement of NOS principles through metacognitive activities surrounding their science activities seems critical for students’ to develop informed NOS views.

The influence of context on students’ NOS views is apparent in Tsai’s (2006) large-scale study which compared high school students’ epistemological views on the tentativeness and creativity nature of science as they relate to two subjects, biology and physics. He found that Taiwanese students’ views expressed differential NOS views depending on the referent of biology or physics. Students viewed biology as a more tentative enterprise that involves changes of concepts and uncertain explanations about
nature; whereas their views on physics suggested they viewed the subject as a significantly more stable and factual enterprise. Interestingly, Tsai (2006) found no significant difference in their views on the role of creativity within biology and physics with students equally agreeing that both require scientists to be creative in their scientific pursuits. These data do, however, suggest that Khisfe’s (2008) assertion that NOS needs to be taught across multiple contexts has validity and may have helped students in Tsai’s (Tsai, 2006) study understand that claims and theories within physics can be as tentative as those ones applied to biology.

The positive impact of teaching NOS explicitly is also evident in studies used to assess student understanding of NOS in elementary students. Akerson and Volrich’s case study (2006) measured first grade students’ conceptions of NOS (using VNOS-D; Lederman, Abd-El-Khalick, Bell, & Schwartz, 2002) prior to and after receiving explicit NOS instruction during their “living things” unit. The student teacher (who possessed fully informed views of NOS) incorporated NOS instruction by using introductory storytelling material on NOS, followed by reflective discussions on targeted aspects of NOS connected to their science content and by concluding the lesson with a summary question on NOS. This intensive approach to incorporating NOS in to the teacher’s instruction resulted in these young students improving their NOS understanding of the tentative, creative and empirical nature of science. Even at this young age, students, post-instruction, were able to grasp that scientists “create ideas”; “use evidence”; “scientists change their minds” and “get new evidence” rather than hold on to their naïve views that
science is “done”; “imagination is not real” and “seeing is believing” (Akerson & Volrich, 2006, p. 384).

Akerson, Buck and Quigley (2011) and Akerson, Buck, Donnelly, Nargund-Joshi and Weiland (2011) designed similar studies to teach other groups of first grade students NOS during their 30-day plant units with comparable results; in both studies, first graders were able to articulate a more informed view of the creative and empirical nature of science and the distinction between observation and inference by the end of the unit. In the first study (Akerson, Buck, & Quigley, 2011) which placed an emphasis on the social and culturally embedded aspect of science; the students demonstrated through their pictures and writings in their science journals that they were also able to grasp that scientists may differ in their interpretations of evidence making science an inherently human process. In contrast, in the second study (Akerson, Buck, Donnelly, Nargund-Joshi and Weiland, 2011), students views remained naïve on the social and cultural embeddedness of science; this inconsistency may have resulted from the different aspects of NOS that were emphasized across the two studies.

Akerson was involved in two other NOS studies (Akerson & Donnelly, 2010; Quigley, Pongsanon, & Akerson, 2011) of elementary students; these studies took place in an after-school Saturday program and involved K-2 students. The researchers (Akerson & Donnelly, 2010; Quigley, Pongsanon, & Akerson, 2011) designed their programs similarly but had important differences in their implementation. Each study lasted 6 weeks and involved explicit reflective instruction on NOS; integrated and non-integrated NOS activities and guided and student-led inquiry-based learning. Both studies used
students’ written work (worksheet prompts or journal prompts) to informally assess students’ NOS views and employed VNOS-D (Lederman, Abd-El-Khalick, Bell, & Schwartz, 2002) to interview and assess students’ views pre and post program.

In contrast to Quigley, Pongsanon and Akerson, 2011, Akerson and Donnelly (2010) were not able to individually interview students pre-program and performed this function in groups of 3-4 students. A significant difference in implementation between the two studies was the choice of activities used to construct NOS knowledge. Akerson and Donnelly (2010) used a selection of activities from life sciences; earth sciences and chemistry to emphasize aspects of NOS; Quigley, Pongsanon and Akerson (2011) used a more focused approach with all activities related to invention and inventing things to draw out students’ views on NOS. This latter study appeared to provide more opportunities to evaluate students’ views on the empirical, tentative, creative, subjective, and cultural nature of science assessed in both studies.

The difference in the nature of the activities used in the two studies played a role in the results obtained. Students in the Quigley, Pongsanon and Akerson (2011) study were constantly involved in inventing and reinventing their creations; this focus led to a substantial number (8 out of 19 pre to post program) of the students changing their views on the tentative nature of science [Certainty] from inadequate to adequate. A smaller number (3) in this study changed their perceptions of the subjectivity of science toward more informed views; this concept across all studies reviewed is hard for elementary students to grasp so any improvement in such a short intervention is considered meaningful. Akerson and Donnelly (2010) in their study equally found most
improvement in students’ views on the tentative aspect of science but provided a caveat for their results; the improvement was context dependent with student views improving when opined on the dinosaur activity (a topic they had been taught previously) but not when discussing weather (a topic they had not yet covered in class). Students in the Akerson and Donnelly (2010) study also were not able to articulate the relationship between subjectivity of science and the theory-laden nature of science; most students in this study discussed the subjective nature of science in the context of scientists having different opinions. Students in the Quigley, Pongsanon and Akerson (2011) study voiced more informed, albeit emergent views, on this aspect of NOS with one student’s comment of “Scientists do not always come up with the same idea” (p. 145) representing these emerging viewpoints. Students in this study also had evolving views on concept that scientists’ explanations can be influenced by scientists holding to different theoretical perspectives. The subjective nature of science was heavily emphasized in this study (Quigley, Pongsanon, & Akerson, 2011) and the nature of activities (having students invent their own devices) lends itself to students internalizing this aspect of NOS. In both studies, students held or developed fairly informed views on the creative and empirical nature of science but found it difficult to understand that scientists’ backgrounds [Socially and Culturally Embedded] could impact their work. Both studies argue that properly scaffolded, reflective discussions and giving students the agency to conduct their own scientific investigations across multiple contexts is key to improving students’ epistemologies. Both these studies were conducted out-with the setting of the classroom environment and short in duration; as a result they have limited
generalizability. Akerson, Buck, Donnelly, Nargund-Joshi and Weiland (2011) undertook a longer-term, classroom-based study of third grade students’ epistemologies that spanned one school year and their findings are discussed next.

Akerson, Buck, Donnelly, Nargund-Joshi and Weiland (2011) created and embedded NOS activities within science content of a third grade curriculum; the authors similarly created inquiry-based hands-on activities in which to embed the NOS activities. Purposeful questioning and reflective discussions were used to stimulate student NOS understanding during and at the end of each science investigation. Students read stories and discussed the NOS aspects incorporated within the stories to further reinforce their understanding. Ten students (the school population was transitory) provided the results for the study. Although the results are not openly quantified, the authors postulated that students in this study made similar, if not greater, gains in improving their NOS understanding to those in Akerson’s other short-term intervention studies (Akerson & Donnelly, 2010; Quigley, Pongsanon, & Akerson, 2011). However, they also cautioned that developmental differences and the duration of this study could explain these differences.

Solidification and gains in understanding were evident in this study (Akerson, Buck, Donnelly, Nargund-Joshi, & Weiland, 2011) across all aspects of NOS with the greatest gains within the creative and tentativeness domains of NOS. By the end of the year, students were also able to understand and articulate the important difference between observations and inferences and the interplay between them. Similarly, gains were apparent in students’ views on the subjectiveness and cultural embeddedness of
science with some students (5 of the 10), by the end of the year, having informed views when one representatively elaborated that scientists have “different ways of thinking and different prior knowledge” (p. 548). The overarching conclusion from Akerson’s three studies (Akerson & Donnelly, 2010; Akerson, Buck, & Quigley, 2011; Akerson, Buck, Donnelly, Nargund-Joshi, & Weiland, 2011) was that elementary students are not too young to understand and grasp concepts that underpin the nature of science and that explicit, reflective NOS instruction combined with inquiry-based, hands-on activities are important pedagogical practices to improve young students’ understandings.

**Conclusion.**

Teaching NOS explicitly to students appears effective in improving students’ epistemological views. With one exception, students in all of the “explicit” studies improved their general NOS understanding. Across the studies, it appears especially important that teachers stimulate students’ metacognition by continuously having them reflect on what aspects of NOS are related to their inquiry-based activities. To help students internalize NOS understanding, researchers suggest that students of all ages need to be provided with sustained, reinforced and appropriate pedagogy related to NOS and students need to be exposed to these practices across multiple contexts. The studies relating to elementary students suggest that it is appropriate to target this age group for NOS instruction and that they are not too young to grasp the nature of scientific practices.

**Summary of the Nature of Science Literature Review.** A synthesis of the “implicit” studies reveals common elements. Students, whether at the high school or elementary level, have relatively limited understanding of NOS when the assumption is
made that their epistemology can be improved through their regular science activities; even when these activities involve inquiry-based learning. Of the five domains of NOS used in this study, students appear to have the greatest understanding of the tentative nature of science and this aspect of NOS is most amenable to improvement. However, even this finding may be suspect given that two studies (Ryan & Aikenhead, 1992; Huang, Tsai, & Chang, 2005) determined that their reasoning behind their relatively ‘more informed’ views was erroneous. Another commonality across the studies was that even though some students could get beyond the conception that science is used to test ideas, there was nominal conception of the more advanced understanding that theory building involves sustained effort with explanations having to be coherent and durable.

The overwhelming conclusion from the “implicit” studies is that students are unable to internalize NOS concepts and hold on to their naïve conceptions when students are expected to passively learn NOS through their science activities. In one study (Wu & Wu, 2011), students’ views became relatively less informed on certain aspects of NOS by the end of the study; the authors explained their disappointing findings on the teacher not taking advantage of the opportunities available to guide students’ reflection and discussion of their NOS views. In the one “explicit” study that found no improvement in students’ NOS views (Tao, 2003), the researcher stressed the lack of improvement in students’ NOS understanding was likely due to the teacher not scaffolding and guiding students’ dyad discussions. Both studies point to the importance of student metacognition to their understanding of NOS.
The results from Smith, Maclin, Houghton and Hennessey (2000) stand out in that they suggest that a constructivist approach to teaching can implicitly enhance students’ understanding of NOS. The classroom environment essential to this process is collaborative, interactive and uses discourse that pushes students’ mean-making and students’ reflection on activities and learning. Corroborating that metacognition is important to improving NOS epistemology is the one student in Bell, Blair, Crawford and Lederman’s (2003) apprenticeship study that took it upon herself to reflect on her activities with her mentor; this reflection appears to have contributed to her being the only student to improve her personal epistemology of NOS. Supporting the role the classroom learning environment can play in students’ epistemologies is evidence from Tsai’s (1998a) study examining the relationship between eighth grade students’ epistemologies and their learning orientation. Tsai (1998a) reported that students who preferred a classroom environment that emphasized real-life problems, genuine scientific discourse and where students decide upon their activities were associated with more informed views on NOS than those students who expressed they preferred a teacher-led learning environment.

In a recent study (McNeill & Pimentel, 2009) examining the critical role of argumentation [Certainty; Social and Culturally Embedded] within the high school science classroom; student-on-student argumentation was most evident in the classroom where the teacher used open-ended questions; scaffolded and encouraged students to take a reflective stance in their discussions with their peers. Ford (2008, p.161) alludes to the need for students to have a “grasp of practice” whereby they take on the roles of
“Constructors” and “Critiquers” of scientific claims [Theory-Laden; Socially and Culturally Embedded] in order for students to become scientifically literate and internalize the nature of scientific knowledge. The learning environment appears critical to improving students’ NOS views with evidence that constructivist learning environments may best model the metacognitive practices that can help improve students’ epistemologies of NOS; these practices are similar to those used in explicitly teaching students NOS.

The difference in practices used by teachers between the “implicit” and “explicit” studies is palpable. The practices of the vast majority of “explicit” studies’ teachers cited are constructivist in nature and appear to help students develop a more informed understanding of NOS. In the “explicit” studies, the science activities were for the most part hands-on, often student-led and the teacher expressively and frequently used reflective discussions and metacognitive prompts to elicit students’ views on NOS and their discernment of its relationship with science knowledge construction. In contrast, the pedagogy used in the “implicit” studies relied more heavily on teacher-led science activities where “how to” directions were provided to students and the activities were expected to lead to a given result.

The results from the two sets of studies suggest that students who reside in classrooms they perceive as being more reformed in nature are likely to have a more informed understanding of NOS; assessing the empirical relationship between scores on the NOSI-E and other measures is at the heart of providing external validity evidence for the NOS construct. The analysis of this evidence provides an opportunity to assess
whether students who perceive their classrooms as more constructivist in nature do in fact have a greater understanding of NOS. Specifically, the analyses examined the concurrent relationship between students’ NOS understanding and their views on the degree to which constructivist pedagogy is used in their science classrooms using multilevel regression analyses. Given that NOS was not “explicitly” taught in the ER project or was a part of the students’ regular curriculum, a positive and strong relationship between NOS understanding and students’ views on the degree to which they view their science classroom learning environment reformed would support the argument that a constructivist learning environment contributes to improved student NOS understanding found in the “explicit” studies.

What is clear from the “explicit” studies is that placing an emphasis on NOS instruction is impactful and can influence the uptake of NOS concepts. When certain aspects of NOS instruction were stressed by the teachers (Akerson, Buck, & Quigley, 2011; Akerson, Buck, Donnelly, Nargund-Joshi, & Weiland, 2011), these concepts were internalized by the students and students differentially developed more informed views of these aspects. Across both sets of studies, researchers stress the need for students to be exposed to NOS instruction repeatedly throughout their K-12 instruction and for students to receive this sustained instruction in multiple contexts.

Ryder and Leach (2008) suggests that for students to developed informed views of NOS will require teachers to exemplify “epistemic issues within a range of science content areas” (Ryder & Leach, 2008, p. 292). This will entail students understanding the different “contextual nuances” (p. 292) between science content and how practices within
each content area can influence student understanding of different aspects of NOS. Tsai (2006) found students’ NOS views differed according to whether students were referring to biology or physics and Khisfe and Abd-El-Khalick (2002), and Khisfe (2008) found that improvement in certain aspects of students’ epistemologies were dependent on the context in which NOS activities were embedded. These studies suggest that teaching NOS across multiple contexts is likely essential for students to internalize NOS views and for them to fully understand the relationship between their views and authentic scientific inquiry within different scientific disciplines. Sandoval (2005) proposes that a comprehensive research agenda is needed to study practical epistemologies across the disciplines given that student NOS understanding appears context dependent.

Akerson, Buck, Donnelly, Nargund-Joshi and Weiland (2011) similarly recommend that future research should help develop a better understanding of how different teaching practices influence student NOS understanding and whether an effective learning progression can be identified if students are exposed to NOS instruction (integrated and non-integrated) across disciplines and throughout their education K-12. To implement this type of research will require researchers to conduct studies beyond the scope of most studies examined in this literature review. Most of the studies to date involve small-scale case studies where the results are not generalizable to the K-12 population. No studies are experimental in nature where the causal mechanisms between student NOS understanding; student science content knowledge and the classroom environment could be elucidated. This type of research will require large-scale assessments to support these types of analyses in order to establish causality. The
objective of this dissertation is to develop a Rasch-based instrument that can reliably and best represent elementary students’ views of NOS for use in large-scale studies needed to examine these types of relationships. Before the context and prior NOSI-E development and validation activities are discuss, it is important to realize why the Rasch model is particularly useful in scale development activities and for science education research.
The Utility of the Rasch Model in Scale Construction and Science Education

Research

Measurement instruments play a vital role in science education research; they are used, for example, to measure conceptual understanding; affective variables such as self-efficacy; learning progressions and students’ views on their science learning environment (Thomas, Anderson, & Nashon, 2008; Liu, 2010; Oon & Subramaniam, 2011; Boone, Townsend, & Staver, 2011; Neumann, Neumann, & Nehm, 2011; Peoples, O’Dwyer, Wang, Brown, & Rosca, in press). If science instruments meet the assumptions of the Rasch model, the measures derived are linear, interval-level and invariant. It is these measurement qualities that make them especially beneficial in scale construction and science education research as the measures can reliably be used in hypotheses and parametric testing.

The Rasch model (Rasch, 1960) uses an exponential transformation to place ordinal Likert responses on an interval logit scale. Placement of ordinal, Likert-scale NOS data onto an interval-level, linear scale ensures that student attitudes or understandings are measured appropriately and the data will meet the assumptions of parametric testing, namely the measures are linear and of equal-interval (Boone, Townsend, & Staver, 2011). By taking the natural log of the odds ratio, stable replicable information about the relative strengths of persons and items is derived with equal differences in logit ability estimates translating into equal differences in the probability of getting an item right no matter where on the scale an item is located. This is not true for changes in scores based on CTT methodology where the relationship between ability and
item difficulties is non-linear, especially at the lower and upper ends of a test’s distribution (Boone, Townsend, & Staver, 2011).

The variable map provides a visual of the ordering and spacing of the items (Boone & Scantlebury, 2006) and due to the interval-level scale, the gaps between items, between persons, and between items and persons have “substantive meaning in terms of the underlying construct” (Callingham & Bond, 2006, p. 2). Gaps between items and persons on the continuum, for example, could mean that facets of the underlying construct are not being measured and the instrument is not covering the breadth of meaning needed to represent the construct. In this study, the variable maps will be especially useful in assessing whether the consecutive and multidimensional approaches provide subscales that cover the person measures well (i.e., are reliable); provide enough information on each domain for researchers; and, are responsive to measuring interval-level change in student understanding.

The sample independence features of the Rasch model overcome other fundamental drawbacks of the CTT analyses highlighted by Wright (1967); Hambleton and Jones, (1993); Smith, (2000); and Neumann, Neumann and Nehm (2011). In CTT, the difficulty of a test or survey is sample dependent and makes measuring change unnecessarily problematic; Wright (1967) states that the “scales on which ability is measured are uncomfortably slippery” (p. 86) as they are dependent upon a “specific set of items” and the “particular ability distribution of the children who happened to appear in the standardized sample” (p. 86). In a Rasch model, the estimation of item parameters is free of the sample’s observed mean and variance (Wright & Stone, 1979). Item
invariance implies that the relative endorsements and location of the items do not change or are independent of the sample responding; by corollary, the relative item endorsements should behave as expected across different samples.

When items are invariant, the Rasch model is particularly useful for developing scales. Scales from prior administrations can be used to anchor new scales developed on to the same metric. In this manner, scale developers can determine if they have improved the psychometric properties of their scale (Boone & Scantlebury, 2006; Boone, Townsend, & Staver, 2011) The Rasch model is also discerning in differentiating between high and low scorers (Gable, Ludlow, & Wolf, 1990) on a measurement scale as it places persons and items on a common scale (Hambleton & Jones, 1993). This makes it ideal for measuring change on a construct (Smith, 2000; Boone & Scantlebury, 2006). The NRC (2011) expresses the need to understand how engagement in scientific practices impacts student learning of the core ideas and nature of science; if items on the NOS instrument prove invariant and have good distributional characteristics, researchers will be able to discern low from high NOS scorers and further be able to measure developmental change over time. In this study, it will be important to establish that items are invariant in each of the three Rasch models used to represent the internal structure of the NOS construct. If this assumption is violated for any one of the three models, it would likely exclude the model from further consideration.

Although Rasch-modeling has beneficial quantitative properties (linear, equal interval measurement), a Rasch-based scale also provides useful qualitative information to a researcher. For example, the NOSI-E could prove particularly useful for science
researchers in assessing whether classroom interventions designed to teach students NOS explicitly in their science curricula are effective. At the beginning of a science unit that will employ pedagogically embedded NOS instruction, students can be placed on the NOS continuum. Their location on the continuum denotes a level of understanding of NOS that can be described by the types and level of items that correspond to their person score. Students can be re-tested upon completion of their science unit and any change in NOS understanding can be qualitatively referenced by the types and level of items corresponding to their new score and quantified appropriately using descriptive statistics and parametric tests. Therefore, the NOSI-E could help assess whether the instructional intervention is fostering their epistemological development and enable researchers to examine the sensitivity of the intervention designed to move students from relatively naïve conceptions to relatively more informed conceptions of NOS.

The three models under study will be assessed to ensure that each one is sufficiently and reliably able to detect changes in individual and group-level person measures on the NOS construct. If the models are responsive, then the qualitative assessment highlighted above is feasible providing researchers with the information needed to describe the development of students’ understanding. These beneficial features were used to develop the NOSI-E. The context for this study and the development activities used to create the NOSI-E scale are summarized in the next section.

**The Evolution Readiness: A Modeling Approach Project**

The data for this study were collected over a period of three years (spring 2008 to spring 2011) as part of an NSF-funded project titled, “Evolution Readiness: A Modeling
Approach” [Horwitz & O’Dwyer, DRL-0822213], a collaboration between content and technology specialists at the Concord Consortium, a non-profit educational and research development organization, and research and measurement specialists at Boston College. The goal of the ER study was to help fourth grade students appreciate elemental evolution concepts that would allow them to better understand the inherently complex topic of evolution in their later learning. Once students comprehend the conceptual understructure needed to understand evolution, the acquisition of the basic concepts underlying the theory of evolution will be easier to grasp. Concord Consortium of Massachusetts developed the ER curriculum intervention for fourth grade students and researchers at Boston College developed the instruments and conducted the research to examine the effects of the intervention. This curriculum intervention included hands-on activities, combined with computer hyper-models to scaffold and teach basic concepts fundamental to understanding evolution.

The Concept Inventory for Evolution Readiness CIER was developed to measure student understanding of the intervention’s eleven learning goals (e.g., basic needs of organisms; inter-specific differences and heritability of traits) designed to ready students to understand evolution. Details of the eleven goals are found in Horwitz and O’Dwyer (2012). Original items were developed and pilot tested for the CIER to measure student understanding of the eleven learning goals. Forty items (61 prompts) were administered in two sessions to fourth grade students with each session designed to be completed in one class period. The CIER was composed of 32 multiple choice items, 5 short-answer
and 24 open-response prompts. The data were analyzed using a partial credit model and were reliable (person separation reliability of 0.88).

In addition to the CIER inventory, two other self-report instruments were developed for the study: the Nature of Science Instrument-Elementary (NOSI-E), the topic of this dissertation; and, the Elementary School Science Classroom Environment Survey (ESSCES; Peoples, O’Dwyer, Wang, Brown, & Rosca, in press). The ESSCES was designed to assess students’ perceptions of constructivist teaching within their science classroom learning environment. The instrument, designed to complement the Reformed Teaching Observation Protocol (RTOP), is conceptualized using the RTOP’s three construct domains: Lesson Design and Implementation (LDI); Content (CT); and Classroom Culture (CC) (Piburn & Sawada, 2000a; Piburn & Sawada, 2000b). Although the items measure three aspects of the science classroom learning environment, the ESSCES items formed a unidimensional scale. A partial credit Rasch model (Masters, 1982; Wright & Masters, 1982) was used to develop the scale; the scale had a person separation reliability of 0.91. Further details of the development of the ESSCES are provided in Peoples, O’Dwyer, Wang, Brown and Rosca (in press). Scores from these three instruments (CIER, NOSI-E, and ESSCES) were used to examine the relationship between students’ content knowledge; students’ understanding of NOS and their views on their classroom environment; investigating these relationships was the premise for establishing external validity for the NOSI-E measure.

The ER study used a cohort-based research design to examine the impact of the intervention. Pre-intervention data (CIER; ESSCES; NOSI-E) was obtained in Year 1
from students within the classrooms of the ten participating elementary teachers at the end of their regular 6-week life science units (this cohort of students provide the baseline data for the three measures of the study). In Year 2 and Year 3 of the project, new cohorts of students within these same ten classrooms were exposed to the curriculum intervention and were assessed on the three measures highlighted above. The CIER and ESSCES administered Year 2 and Year 3 were identical to the one administered to students in Year 1 (the baseline year). The development of the NOSI-E was closely associated with the phases of the ER cohort-based designed study. To improve the reliability and construct validity of the NOSI-E, the composition of the items of the NOSI-E changed in Year 2 and in Year 3; however, the administrations were placed on the same metric using 16 anchor items developed and administered in the pilot and Year 1 of the ER project. A description of the NOSI-E instrument development process and validation activities is provided in the next section.

**Development and validation of the Nature of Science Instrument-Elementary.** The NOSI-E was developed using Rasch principles (Rasch, 1960) to measure elementary students’ understanding of the nature of scientific practices. Specifically, a partial credit Rasch model (Masters, 1982; Wright & Masters, 1982) was used to develop the NOSI-E (Peoples, O’Dwyer, Shields, & Wang, in review). The Rasch model is a strict measurement system in which items and examinees are located on a common interval scale (ruler) and enable the linear combinations of items and examinees to be analyzed statistically. As mentioned, creating scales based on Rasch modeling has a significant advantage; researchers can continuously improve their scale over time whilst
retaining the same metric (Boone & Scantlebury, 2006). This is especially useful when a new instrument is being innovated and involves item development that is targeted to measure a complex construct such as the NOS. In this study, students’ understanding of the NOS was measured as part of a three year project. The successive administrations of the NOSI-E enabled the instrument to be improved over the three years of the ER study. Results obtained from the earlier administrations were used as a guide for new item ideation and development. The composition of the items of the NOSI-E changed in Year 2 and in Year 3; however, as mentioned, the Year 2 and Year 3 items were placed on the same metric using 16 anchor items developed and administered in the pilot and Year 1 of the ER study.

In this type of extended project, the scale development process is fluid, amenable and allows researchers to use evidence (i.e., item and person locations) from previous administrations to guide the development of new items that can improve the reliability and construct validity aspects of the scale. As Boone and Scantlebury (2006, p. 262) state, “the ability to constantly revise a measurement instrument improves its reliability and validity.” The item anchoring process facilitates instrument development activities and enables students’ NOS understanding to be compared across the years of the study (Boone & Scantlebury, 2006; Boone, Townsend, & Staver, 2011). Messick (1980, p.1019) similarly purports that “construct validation is a continuous, never-ending process developing an ever-expanding mosaic of research evidence” with the distinct possibility that new evidence could require a change in the construct, the theory supporting its construction or its measurement. It is the preponderance of evidence and its
sustainability that helps researchers justify their claims and inferences when using data from their assessments.

Messick’s (Messick, 1980; Messick, 1995) unified concept of construct validity guided the development activities for the NOS scale and provides the framework for this dissertation’s analyses. Messick (1995, p741) defines validity as “an evaluative judgment of the degree to which empirical evidence and theoretical rationales support the adequacy and appropriateness of interpretations and actions on the basis of test scores or other modes of assessment.” Evidence from six aspects of test validity combine to provide test developers with the justification to claim that the meaning or interpretability of the test scores is trustworthy and appropriate for the test’s intended use. Messick’s (1995) six aspects of test validity are: content; substantive; structural; generalizability; external and consequential. Wolfe and Smith (2007b) used Messick’s validity conceptualization to detail instrument development activities and evidence that are needed to support the use of scores from instruments based on the Rasch measurement framework. A summary of the construct validation activities used to assess the most appropriate measurement model for the NOSI-E is provided in Table 2.2.

Previous to this dissertation, three validity aspects (content, substantive and generalizability) of construct validity guided the activities used to develop and choose the items for the unidimensional Rasch scale. These same validity aspects also helped determine the appropriate internal model for the NOS construct as evidence was needed that the items behave similarly and are of the same technical quality and reliability regardless of the model used to represent the internal structure of the NOS construct.
Justification for the need for this evidence is presented in Chapter 3 with the results of the empirical analyses used to provide the evidence in Chapter 4.

Table 2.2

**Summary of NOSI-E’s Validity Evidence**

<table>
<thead>
<tr>
<th>Validation Evidence</th>
<th>Validity Aspect</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Content</strong></td>
<td><strong>Substantive</strong></td>
</tr>
<tr>
<td>Instrument Purpose</td>
<td>Rating Scale Functioning</td>
</tr>
<tr>
<td>Item Development</td>
<td>Item Difficulty Hierarchy</td>
</tr>
<tr>
<td>Evidence</td>
<td>Expert Reviews</td>
</tr>
<tr>
<td>Field Test</td>
<td></td>
</tr>
<tr>
<td>Item Technical Quality</td>
<td></td>
</tr>
</tbody>
</table>

**Validation Evidence**

<table>
<thead>
<tr>
<th>Validity Aspect</th>
</tr>
</thead>
<tbody>
<tr>
<td>Structural</td>
</tr>
<tr>
<td>Dimensionality Analyses:</td>
</tr>
<tr>
<td>Comparison of Rasch Models;</td>
</tr>
<tr>
<td>Confirmatory Factor Analyses</td>
</tr>
</tbody>
</table>

Rasch Subscale Correlations

Discrepant Case Analyses

2. Research that provides evidence for the content validity aspects highlighted in italics was performed prior to this dissertation (Peoples, O’Dwyer, Wang, & Shields, in review) and discussed below.
3. Evidence to support the consequential validity aspect of the NOSI-E (italicized) was beyond the scope of this dissertation.
A summary of the preliminary NOSI-E development work and associated validation activities are discussed next in the context of Messick’s (1995) validity framework for test development and Wolfe and Smith’s (2007b) adaptation of this framework for developing a Rasch-based scale. Full results of these validation activities are available in Peoples, O’Dwyer, Shields, & Wang (in review).

**Content Aspect.**

The content aspect of construct validity examines the “content relevance, representativeness and technical quality” (Messick, 1995, p.745) of the items used as indicators of the NOS construct. In a self-report survey, items are developed and used to measure the breadth of the construct; these items are themselves a sample of all possible items that can be used to measure student understanding of the NOS construct (Wolfe & Smith, 2007a). Gable, Ludlow and Wolf (1990) in their study of enhancing the validity of affective measures using Rasch latent trait models, stressed that “when the targeted variables are conceptually identified and then operationally defined through the item writing process, instrument developers need to place greater emphasis on spanning the underlying psychological continuum for a targeted variable” (p. 877). Approximately, 175 items were authored for the NOSI-E measure. In addition, 18 items from prior nature of science surveys were included in the pilot (Moore & Foy, 1997; Conley, Pintrich, Vekiri, & Harrison, 2004; Deniz, 2007; Hampton, 2007). The items spanned the five domains of the NOS construct; at the start of the process, each domain had 4-5 times the number of items needed for the NOSI-E. A hierarchical perspective was taken in creating
the items for each domain with the items created representing a continuum that spans naïve conceptions on the domain construct to informed conceptions.

The initial pool of items were reviewed for content validity (definition, representativeness, relevance) and specificity (simplicity, clarity, directness, reading level) using a panel of educational researchers and practitioners. At this stage of the instrument’s development, items were flagged to be kept, revised or removed based on the above criteria used in the review process. An iterative process was used to reduce the number of items to 57 for a field test of the items. Fifty-seven items were field tested with four groups of elementary school students; seven items were removed as the students were confused by the content of the items. The remaining 50 items were administered in a pilot study.

The pilot study was foundational in nature and was used to guide initial instrument development efforts. The pilot data were not associated with the ER study; as such, the pilot data are not included in this dissertation. Evidence (item-measure correlation and item fit data) of the technical quality of the items was also used to make decisions on whether to keep, revise or reject items for the NOS scale. The item-measure Pearson product moment correlation coefficient (Pearson correlation thereafter) examines the consistency of the scores for any one item to the respondents’ average scores for the remaining items. The mean square error fit statistics were also examined to ensure that the difference between observed values related to an item (i.e., item difficulty estimates) were minimally different from the expected values predicted by the Rasch model; a large discrepancy would indicate that an item(s) was not related to the NOS construct.
Misfitting items were removed from further consideration. The 28 items that form the NOSI-E were well-fitting (Peoples, O'Dwyer, Shields, & Wang, in review) and all have positive item-measure Pearson correlations of above 0.4 (data not shown).

**Substantive Aspect.**

The substantive aspect of construct validity assesses whether the responses to the items are consistent with the theoretical framework used to develop the items; in this study, Lederman’s theoretical framework (Lederman, 1992; Lederman, Abd-El-Khalick, Bell, & Schwartz, 2002; Lederman, 2007) predicated the development of items for the NOS scale. Two pieces of evidence (rating scale functioning and item difficulty hierarchy) was used to demonstrate that responses to NOS items were coherent with theoretical expectations. Using guidelines recommended by Wolfe and Smith (2007a) a small, even number of response options was used to operationalize the NOSI-E. A Likert scale with four response options was used to rate the degree of understanding toward each statement; coding (or recoding) for all items dictated that a response of “0” would be indicative of the lowest level of NOS understanding with a “3” denoting the highest level or more informed understanding of NOS. The expected performance of a well-functioning item is that the calibrated steps (item difficulty deltas) and the average of the respondents’ measures are monotonic; that is, increasing levels of the latent trait are associated with more affirmative categories. If the functioning of the rating scale is appropriate, this provides evidence that the respondents are using the scale according to the intent of the item developers (Wolfe & Smith, 2007b).
The expected ordering of the items on the construct variable map between and within domains similarly provides evidence for the substantive aspect of construct validity. Van Alphen, Halfens, Hasman and Imbos (1994, p. 197) posit that the Rasch model’s response-centered approach to modeling survey data “enable the researcher to test in a more explicit way hypotheses about the obtained data using the measurement instrument under development”. The item-person variable maps were examined to determine if each item’s difficulty placed on the continuum in a priori hypothesized hierarchical locations based on Lederman’s theoretical framework (Lederman, 1992; Lederman, Abd-El-Khalick, Bell, & Schwartz, 2002; Lederman, 2007). By demonstrating that the empirical item difficulty hierarchy is consistent with Lederman’s theoretical framework, a validity argument is made to support the substantive aspect of NOSI-E’s construct validity.

Using information from the rating scale analyses and the expected theoretical ordering of items, items were removed, revised or kept for future administrations. Disordinal items were removed from the scale; disordinal items indicate that students did not likely have a common understanding of the concept embedded in the item (Green & Frantom, 2002). Items from the pilot and Year 1 that required recoding due to their negatively worded statements were problematic with all of them being, on average, hardest to understand. These items shifted position when positively phrased in Year 2’s administration and the decision was made to remove them from further consideration. Items that, in hindsight, were redundant in content (located at the same item difficulty in the hierarchy) were removed; of the redundant items, the items with the best item fit and
rating scale performance were retained. The evidence from the rating scale performance and expected item ordering combined to make decisions on item removal and retention. The 28 items remaining that form the NOSI-E (Appendix 1) exhibit monotonic rating scales (average measures increased systematically) and expected item hierarchical structure (Peoples, O’Dwyer, Shields, & Wang, 2012).

**Generalizability.**

A measure is considered generalizable when the score meaning and properties function similarly across multiple contexts or time points (Messick, 1995; Wolfe & Smith, 2007a). In the Rasch framework, item invariance across multiple contexts or administrations is an important assumption of the model (Bond & Fox, 2007; Wolfe & Smith, 2007b); this ensures that the score meaning is equivalent across “test” forms or “test” administrations. Although NOS content was not taught explicitly during the ER project, one of the analyses conducted was to compare student NOSI-E scores across the three cohorts participating in the project to determine if students’ understanding of the nature of science changed. Due to efforts to improve the psychometric properties of the NOSI-E, the items changed across the three administrations. Item anchoring was used to equate the “test” forms across the three years of the instrument’s administration. Item anchoring fixes the location of items common to each of the “test” forms to specific values along the latent trait. Sixteen items were common to all three NOSI-E “test” administrations of the ER project and were used to place the responses from the three cohorts on to the same metric (Peoples, O’Dwyer, Shields, & Wang, in review). These sixteen items were invariant across the three years thereby supporting the generalizability.
of the scale and the use of the “test” in comparing student achievement across the three cohorts.

The reliability of an instrument also provides evidence to support the generalizability aspect of construct validity (Bond & Fox, 2007; Wolfe & Smith, 2007b). Reliability analyses provide the degree to which measures are consistent or stable across instrumentation. In the Rasch model, the internal consistency reliability coefficient, the person separation reliability, measures the ratio of the variance in latent person measures (true) to the estimated person measures; similar to Cronbach alpha, it ranges from zero to one. The person-item maps from the Pilot and Year 1 indicated that there were regions of the person-item continuum that lacked items; one of those regions was around the mode of the person distribution with another toward the very high (difficult) end of the person distribution (Peoples, O’Dwyer, Shields, & Wang, in review). These gaps were a major source of error variance that could impact score interpretability, inference making and consistency of the measures across time; the person separation reliability was 0.78 for the pilot data (Peoples, O’Dwyer, Shields, & Wang, in review). As a result, the pilot or Year 1’s instrument did not have the capacity to reliably differentiate students with high ability from those with lower ability. Two opportunities existed during the ER project to improve the psychometric properties and reliability of the scale.

No items were removed or added between the pilot and Year 1; this decision was predicated on the belief that reproducibility (i.e., item invariance across test administrations) was important to establish. In Year 2, twenty seven items were retained, four items were revised and eighteen new items were authored. In Year 2 revisions to the
instrument were targeted toward increasing the reliability and variance of the scale as a whole; as a result, eight of the eighteen newly authored items were retained to fill the gaps in the construct map. Instrument development efforts continued in Year 3 of the ER project. In Year 3, eighteen common items from the Pilot/Year 1 administrations were administered; eight from Year 2’s administration and five new items were authored and tested out. The goal of these revisions was to improve the variance of the individual subscales; four of the five newly authored items were retained. The final NOS scale includes twenty eight items (16 from Pilot/Year 1; 8 from Year 2 and 4 from Year 3’s development cycle); the person separation reliability improved from 0.78 in the pilot to 0.84 (Peoples, O’Dwyer, Shields, & Wang, in review). The consistency and interpretability of score meaning for the 28-item instrument will enable researchers to categorize students into three distinct ability groups (naïve; adequate and informed). A road map of each development cycle in the instrument development process is summarized in Table 2.3.

Additional analyses were undertaken to ensure the generalizability of score meaning; Differential Item Functioning (DIF) analyses were used to establish item invariance across sub-groups. Items were invariant (within standard error) across gender and state membership groupings providing evidence that different groups were interpreting the items similarly and the quality of the measure was maintained regardless of the context.
Table 2.3

*NOSI-E Item Development Cycles*

<table>
<thead>
<tr>
<th>Time of Administration</th>
<th>Development Cycle (Number)</th>
<th>Number of Items Administered</th>
<th>Number of Items Revised from Previous Cycle</th>
<th>Number of Items Retained from Previous Cycle</th>
<th>Number of New Items Authored</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fall/ Winter 2008</td>
<td>Expert Panel</td>
<td>105</td>
<td>0</td>
<td>57 for Field Test</td>
<td></td>
</tr>
<tr>
<td>Winter/Spring 2008/2009</td>
<td>Pilot (N = 741)</td>
<td>50</td>
<td>0</td>
<td>50</td>
<td>0</td>
</tr>
<tr>
<td>Spring 2009</td>
<td>Year 1 (N = 122)</td>
<td>50</td>
<td>0</td>
<td>50</td>
<td>0</td>
</tr>
<tr>
<td>Spring 2010</td>
<td>Year 2 (N = 173)</td>
<td>49</td>
<td>4</td>
<td>27</td>
<td>18</td>
</tr>
<tr>
<td>Spring 2011</td>
<td>Year 3 (N = 161)</td>
<td>31</td>
<td>0</td>
<td>18 from Year 1</td>
<td>5</td>
</tr>
</tbody>
</table>

The 28 items remaining at the end of the NOSI-E development process cut across the five domains of NOS available for this study. Theoretically, there are six items each representing the Empirical (EMP), Theory-Laden (THL), and Certainty (CER) domains with five items each representing the Inventive (INV) and Socially and Culturally Embedded (SCE) domains. The item prompts for the 28-items remaining after the
development process are presented in Appendix 1 with anchor items denoted by an “A” adjacent to the item difficulty statistic. The developmental cycle (year of administration) is also identified. These items form the basis for this dissertation; the item prompts and theoretical domain designation are presented in Table 2.4.

This dissertation examined and compared three theoretical representations for the internal structure of these twenty-eight items and these representations are discussed in the next section with the Rasch measurement models used to assess these representations provided in the methodology section. The primary purpose of this study is to determine which theoretical representation provides for the most reliable, interpretable and responsive measure of the internal structure of the NOS construct. Determining which model best represents the internal structure of the NOS construct provides the structural validity evidence for the construct as measured by the NOSI-E. As a result, this ensures that the parameter estimates derived from the model are not biased and can be used for their intended purpose.

**Theoretical Representations for the Internal Structure of the NOS Construct**

Ludlow, Enterline and Cochran-Smith (2008) posit, “Rasch models are used as confirmatory tests of the extent to which scales have been successfully developed according to explicit a priori measurement criteria” (p. 196). Three approaches, and by corollary three Rasch models, will be assessed to provide a “confirmatory test” of the internal structure of the NOSI-E data. Data from each of the three Rasch models was assessed and used to provide evidence for Messick’s structural validity aspect of the NOS
### Table 2.4

*Item Prompts and Theoretical Domain Placement for the 28-item NOSI-E*

<table>
<thead>
<tr>
<th>Theoretical Domain</th>
<th>Item</th>
<th>Item Prompt</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Empirical</strong></td>
<td>EMP8D</td>
<td>A good way to know if something is true is to do an experiment.</td>
</tr>
<tr>
<td></td>
<td>EMP8I</td>
<td>Science describes what happens in nature.</td>
</tr>
<tr>
<td></td>
<td>EMP9A</td>
<td>Scientists explain how something works.</td>
</tr>
<tr>
<td></td>
<td>EMP9I</td>
<td>Experiments are used to see what happens in nature.</td>
</tr>
<tr>
<td></td>
<td>EMP9J</td>
<td>Science helps answer questions about how something works.</td>
</tr>
<tr>
<td></td>
<td>EMP14B</td>
<td>Scientists infer what they think is happening from what they already know.</td>
</tr>
<tr>
<td><strong>Theory-laden</strong></td>
<td>THL6C</td>
<td>Scientists use different ways to test their hypotheses.</td>
</tr>
<tr>
<td></td>
<td>THL8F</td>
<td>Scientific questions are answered by observing things.</td>
</tr>
<tr>
<td></td>
<td>THL8K</td>
<td>Scientists use what they found in the past to help explain their new findings.</td>
</tr>
<tr>
<td></td>
<td>THL9K</td>
<td>Theories can change when new evidence is found.</td>
</tr>
<tr>
<td></td>
<td>THL12K</td>
<td>Scientists create different types of experiments to answer their questions.</td>
</tr>
<tr>
<td></td>
<td>THL14E</td>
<td>If we do the same experiments many times, we may get different results.</td>
</tr>
<tr>
<td><strong>Certainty</strong></td>
<td>CER6J</td>
<td>Trying things out helps scientists think of new ideas.</td>
</tr>
<tr>
<td></td>
<td>CER6H</td>
<td>A lot of data is needed to decide if a hypothesis is true.</td>
</tr>
<tr>
<td></td>
<td>CER7H</td>
<td>Two scientists can disagree, but both can have good ideas.</td>
</tr>
<tr>
<td></td>
<td>CER8L</td>
<td>When scientists have a good idea, they continue to try to make it better.</td>
</tr>
<tr>
<td></td>
<td>CER8M</td>
<td>New theories in science should only be accepted when there is a lot of evidence to support them.</td>
</tr>
<tr>
<td><strong>Inventive</strong></td>
<td>INV6D</td>
<td>You have to be creative to work in science.</td>
</tr>
<tr>
<td></td>
<td>INV8G</td>
<td>To explain their results, scientists need to be creative.</td>
</tr>
<tr>
<td></td>
<td>INV12L</td>
<td>Although science is based on facts, scientists do need a good imagination.</td>
</tr>
<tr>
<td></td>
<td>INV12M</td>
<td>A good imagination is needed to make predictions about what will happen in an experiment.</td>
</tr>
<tr>
<td></td>
<td>INV12N</td>
<td>A good imagination is needed to create the best experiment to test an idea.</td>
</tr>
<tr>
<td><strong>Socially &amp; Culturally Embedded</strong></td>
<td>SCE13C</td>
<td>The country a scientist comes from influences how they understand the results of an experiment.</td>
</tr>
<tr>
<td></td>
<td>SCE13F</td>
<td>A scientist’s beliefs may change how they do their work.</td>
</tr>
<tr>
<td></td>
<td>SCE13H</td>
<td>Where scientists come from may lead to different answers to the same question.</td>
</tr>
<tr>
<td></td>
<td>SCE13I</td>
<td>How scientists see the world is influenced by the culture they grew up in.</td>
</tr>
<tr>
<td></td>
<td>SCE13K</td>
<td>Where scientists live may affect what they are allowed to work on.</td>
</tr>
</tbody>
</table>

1. Items for this dissertation are based on Year 1, Year 2 and Year 3 data combined from the ER Project (Horwitz & O’Dwyer, 2012);
2. EMP8D highlighted in bold font is from the following source: Conley, Pintrich, Vekiri, & Harrison (2004);
3. THL8F highlighted in bold font is from the following source: Moore & Foy (1997).
These three theoretical representations are portrayed in Figure 2.2 and discussed below.

One approach supposes that the NOS construct is unidimensional with homogeneous items (denoted by boxed X’s in Figure 2.2) lying on one continuum; items relate to the same underlying latent trait (denoted by the encircled θ) and provide student ability estimates of their understanding of NOS (denoted by the boxed NOS). This theoretical representation is commonly assessed using either a rating scale or partial credit unidimensional Rasch model given that the NOS construct is operationalized using a Likert rating scale. In the ER study, a partial credit model (Masters, 1982; Wright & Masters, 1982) was used to develop the NOSI-E (Peoples, O’Dwyer, Shields, & Wang, in review) and is used here. The NOS construct could also be represented by separate continuums, each measuring the different dimensions (Empirical, Theory-laden, Certainty, Inventive and Socially and Culturally Embedded) of the NOS construct. The NOS construct in this scenario is not considered global but composed of homogeneous dimensions that are affected directly by NOS latent traits in differentiated ways (Fusco & Dickes, 2006). Each constitutive domain is distinct from each other (i.e., each domain represents a separate latent ability or trait) and, although the items of each domain lie on a continuum, there is no common metric among them. This representation is known as the consecutive approach (Figure 2.2) to modeling the NOS construct; it is assessed using separate unidimensional partial credit Rasch models for each dimension or subscale.

The final theoretical representation of the NOS data assessed in this thesis is the multidimensional approach whereby there are two types of effects between the latent
ability for each dimension and the observed responses. As shown in Figure 2.2, there is a direct effect (represented by the straight lines) between the latent variables (domain $\theta_s$) on the observed responses. There is also, however, an indirect effect on the observed responses through the correlation of the domain latent traits with each other (as represented by the curved lines). These correlations, corrected for measurement error, are estimated by the model (Wolfe & Singh, 2011) and can be compared to the attenuated correlations between dimensions derived from the consecutive approach (Briggs & Wilson, 2003). Thus, this theoretical representation models that students’ use more than one latent ability to respond to items within each of the domains. Students who have a high latent ability on one domain of NOS e.g., the Certainty domain (direct effect) are able to make use of this knowledge of the Certainty domain to help to respond to items in other domains such as items of the Theory-laden domain (the indirect effect). The three theoretical representations were tested empirically in this study using the Rasch models outlined in the methodology section in Chapter 3.

In reporting assessment results, practitioners often will combine multiple subcomponents of a test to provide educators with a single score. Similarly, in other instances, scores are often disaggregated to supply educators with scores on each content domain of the test. Briggs and Wilson (2003, p. 97) state that these practices are a “departure from the measurement ideal of simple unidimensionality”. In each case, practitioners are violating the basic tenet of sound measurement, namely to measure unidimensional constructs so that sound and reliable interpretations and explanations can be furnished and acted upon.
If a construct is truly multidimensional in nature and is modeled using a multidimensional model, researchers will be able to obtain more differential information about their sample than if they ignored the construct’s dimensional structure and analyzed the data using a unidimensional model. Using the NOS construct as an example, if student data were analyzed with a unidimensional model, the researcher would obtain one composite score (with an associated standard error) for each student. However, if the internal structure of NOS was in fact multidimensional in nature, and the more complex multidimensional model was used to analyze the student data, the researcher would obtain five ability scores (and associated standard errors), one for each dimension (Empirical, Inventive etc.), for each student. This would provide researchers with a better assessment of students’ understanding of NOS, especially if students’ ability profiles differ across the five dimensions of the test. Cheng, Wang and Ho (2009) describe this as “providing more discrete differentiation between person measures” (p. 383). By using the unidimensional approach, this interpretable, “discrete” information is essentially ignored (Briggs & Wilson, 2003; Kennedy & Draney, 2009).

Liu, Wilson and Paek (2008) employed a multidimensional Rasch model, for example, to explore the dimensionality of OECD’s 2003 mathematics PISA assessment data and understand gender differences in mathematics achievement across four content domains. By being able to model the multidimensional nature of the mathematics data, Liu, Wilson and Paek (2008) were able to show that the magnitude of gender differences across these four domains differed with boys having higher, on average, ability estimates in the Space and Shape domain than girls but were, on average, of equal ability in the
Quantity domain. The authors emphasize “the importance of contextualized investigations of math gender issues” (Liu, Wilson, & Paek, 2008, p. 30) and how the use of multidimensional Rasch analyses enabled more refined diagnostic information to be given to teachers so that they will be able to “cater to the differential learning needs of boys and girls” (p. 30). Kennedy and Draney (2009) similarly used the multidimensional Rasch model to provide classroom teachers with individual and class based student performance on multiple, related variables. Teachers were able to use this “rich” (p. 1) formative assessment information to understand their students’ strengths and weaknesses and adjust their practice accordingly (Kennedy & Draney, 2009). The type and extent of the information provided by each model (unidimensional, consecutive and multidimensional) is one of the criteria used in this study to assess and compare the models in the determination of the most appropriate measure of student understanding of NOS. In summary, the Rasch model is particularly flexible and suited for establishing the dimensionality of a scale (Briggs & Wilson, 2003; Thomas, 2004; Wolfe, Hickey, & Kindfield, 2009) and was used in this research to establish the internal structure of the NOS construct.

Conclusion

An instrument, if it is to be used for science education research requires two essentials; it should be predicated on a sound theoretical framework and be rigorous in its ability to examine and explain the relationship between the theoretical frame and the researchers’ findings resulting from their investigations. This chapter presented a review of the literature that premised the theoretical framework of the NOSI-E instrument
designed to measure elementary student understanding of the NOS construct. Integral to
developing a new instrument is to have an understanding of the current status of students’
understandings of the NOS construct in order that the instrument is appropriately targeted
at the students being assessed. The literature review summarized past research studies
that described students’ NOS understandings.

The NOSI-E was developed on Lederman, Abd-El-Khalick, Bell and Schwartz’s
(2002) formative theoretical framework for NOS which has been used extensively in
science research to assess students’ understanding of NOS (Ryder, Leach, & Driver,
1999; Moss, Abrams, & Robb, 2001; Thoermer & Sodian, 2002; Khisfe & Abd-El-
Khalick, 2002; Bell, Blair, Crawford, & Lederman, 2003; Sandoval & Morrison, 2003;
Kawasaki, Herrenkohl, & Yeary, 2004; Akerson & Abd-El-Khalick, 2005; Akerson &
Volrich, 2006; Smith & Wenk, 2006; Khisfe & Lederman, 2006; Khisfe, 2008; Akerson
& Donnelly, 2010; Akerson, Buck, & Quigley, 2011; Akerson, Buck, Donnelly,
Nargund-Joshi, & Weiland, 2011; Wu & Wu, 2011) and in studies to examine the
relationship between students’ NOS understanding and exposure to different classroom
pedagogies (Tsai, 1998a; Smith, Maclin, Houghton, & Hennessey, 2000) and between
their NOS understanding and science content knowledge (Lombozo, Thanukos, &
Weisberg’s, 2008). Each of the five domains (Empirical, Theory-laden, Certainty,
Inventive and Socially & Culturally Embedded) that compose the NOS construct was
described to provide an in-depth presentation of the theoretical framework for the NOS
construct. As a result, by using the literature review, it was shown that the NOSI-E
instrument meets the first criterion necessary for use in science education research; namely, it is predicated on a sound theoretical framework.

Two types of studies (implicit and explicit) have been used to measure student understanding of the NOS construct. The implicit studies measured student NOS understanding during their regular science classroom instruction and activities; the researchers in these studies were trying to determine if students improved their NOS understanding by taking part in these activities without being exposed to any “explicit” NOS instruction. In contrast, in the explicit studies, students were taught NOS during their science instruction and activities and assessed to determine if this more explicit approach improved students’ NOS understanding. Both sets of studies predominantly used qualitative research to describe students’ NOS understanding. The implicit studies revealed that, in general, students have relatively limited or a naïve understanding of NOS and researchers of these studies concluded that it is largely erroneous to assume that students can improve their NOS understanding through taking part in regular science activities. In contrast, students when taught NOS explicitly, for the most part, improved their NOS understanding; this finding was evident even in young elementary students (Akerson, Buck, & Quigley, 2011; Akerson, Buck, Donnelly, Nargund-Joshi, & Weiland, 2011). The difference in instructional practice between the two types of studies was thought to play an impactful role in the results of these studies. Instruction in the implicit studies was largely teacher-led and traditional. In contrast, instruction in the explicit studies was predominantly constructivist in nature where students were actively involved in their learning activities. Findings from this literature review were used to provide
supporting validity evidence for the results of this dissertation that are discussed in Chapter 4.

To meet the second utility criterion (i.e., the instrument can explain the relationship between the theoretical frame and researchers’ findings resulting from their investigations), the instrument has to be built on sound measurement principles and provide the researcher with the information needed to explain their results as they relate to the latent construct. The literature review highlighted the utility of using the Rasch model in scale construction and science education research. If the data meet the assumptions of the model, a Rasch-based scale is particularly effective in measuring interval-level change within a population; as such, it should prove invaluable in studies targeted at understanding the relationship between student science content knowledge; their learning environment and their understanding of the nature of the scientific endeavor.

This chapter summarized the Evolution Readiness project (Horwitz & O’Dwyer, 2012) which provided the context for this dissertation and described prior development and validation research used to develop the items for the NOSI-E instrument (Peoples, O’Dwyer, Shields, & Wang, in review). The summary of the Evolution Readiness project outlined the data resources and research design upon which this study is predicated. The description of prior NOSI-E development work (which focused on content, substantive and generalizability validity aspects of construct validity) is essential to understanding the focus and purpose of this dissertation. The purpose of this dissertation is to compare and assess three Rasch models that represent three different theoretical positions for the
internal structure of the NOS construct. This dissertation therefore provides essential
evidence for the structural validity aspect of the NOS construct. An explanation of these
three theoretical internal representations of the NOS construct concludes this chapter.

When combined with the empirical evidence from prior NOSI-E’s development
and validation activities (Peoples, O’Dwyer, Shields, & Wang, in review), a validity
argument justifying the instrument’s use in science education research can be made.
Chapter 3 outlines the validity argument proposed for the study and provides the
methodology and associated evidence needed to support the validity argument. The goal
is to provide science education researchers with a measurement tool that best represents
the NOS construct and one that meets all of the assumptions of the Rasch model. If it
succeeds, researchers will be able to reliably use estimates from this tool in their science
investigations and make appropriate inferences regarding their data.
Figure 2.2: Comparative Theoretical Representations for Nature of Scientific Practices’ (NOS) Data
(Theoretical Framework: Lederman, Abd-El-Khalick, Bell and Schwartz, 2002; Representations modeled from Briggs & Wilson, 2003)
Chapter Three: Methodology

Lederman (2007) and Akerson, Buck, Donnelly, Nargund-Joshi and Weiland (2011) recommend that future research is needed that examines the relationship between students’ views on the nature of scientific knowledge, learning environments and students’ content knowledge. Similarly, in the NRC’s (2011) new framework, the NRC explicitly states that research is needed into “how engagement in specific practices supports the development of both specific (core) ideas in science and understanding of the nature of science” (p. 13-4). They emphasize the importance of studying student learning progressions that embed core content ideas and assessing their relationship between classroom pedagogy and student understanding of the nature of scientific knowledge construction.

The primary purpose of this dissertation is to determine which of three models (unidimensional, multidimensional or consecutive) best represents the internal structure of the NOS construct. The Rasch-based NOS instrument developed will measure elementary students’ views of NOS and will be suitable for use in studies needed to examine these types of relationships and one that can examine changes over time as students’ progress in their learning and understanding. This chapter describes the methodology that was used to address the research question that was first presented in Chapter 1:

Q1: Is the internal structure of the NOS construct best represented by a unidimensional, multidimensional or consecutive Rasch model? Which of the three Rasch-based models will provide the most reliable, interpretable and
responsive Rasch-based measure that will assess elementary students’ understanding of the nature of science (NOS); and, one that can be used in science education research and teaching?

To address the research question, data from Year 1, Year 2 and Year 3 of the ER project were used. This chapter is divided into the following four distinct sections: (1) a description of the Rasch models that will be used to represent the three theoretical structures of the NOS construct discussed in Chapter 2; (2) a description of the study’s data sources; (3) a statement of the hypothesis formulated that relates to the research question; and (4) a description of the data analysis procedures, framed in the context of a validity argument, that were used to address the research question highlighted above.

The Rasch Family of Models

The Rasch model cannot be considered as just one model; there is a family of Rasch models that are designed to handle different types of data (Wright & Mok, 2000). A Rasch model (Rasch, 1960) offers the ability to place persons and items on a common scale which is linear, invariant and hierarchical (Hambleton, Swaminathan, & Rogers, 1991). Dichotomous models are used extensively in the education field to measure student responses to multiple choice items and in the medical field, for example, to test patient’s short-term memory (Wright & Stone, 1979). Extensions of the dichotomous Rasch model enable researchers to analyze ordinal categorical data (e.g., Likert scales) by applying either the Rating Scale model (RSM) or Partial Credit Rasch model (PCM); these forms of the Rasch model have been used extensively in the medical and

Another extension of the Rasch model, the Multi-dimensional Random Coefficients Multinomial Logit Model (MRCML: Adams, Wilson, & Wang, 1997) is used to model constructs that have multiple latent traits; multidimensional constructs are often found in social science and education research (Briggs & Wilson, 2003; Liu, Wilson, & Paek, 2008). As mentioned, the MRCML (Adams, Wilson, & Wang, 1997) is especially suitable for constructs that are inherently multi-faceted but whose domains are related by an underlying higher-order construct. In the following sections, the Rasch models that will be used in this study to model the internal structure of students’ NOSI-E responses are explained. First, the basic Rasch model that underpins all these models will be discussed.

**The basic Rasch model.** All these models highlighted above rely on the basic functional form of the Rasch model (Masters & Wright, 1984); the Rasch model provides a mathematical model for the probabilistic relationship between a person’s ability and the difficulty of items on a test or survey. In order to understand the psychometric models used in this dissertation, it is important to first understand the basic Rasch model. The Rasch model (Rasch, 1960) was developed for dichotomous achievement data in education. Students respond to multiple choice items; if they answer successfully (the key), they receive a score of 1; if they answer unsuccessfully (choose a distracter), they receive a score of 0. The probability of scoring a one is governed by the difference in the
student’s latent ability and the difficulty of the item: this probabilistic relationship is illustrated by Equation 1:

\[ P_{ni}(X=1) = f(\theta_n - \delta_i) \quad \text{Eq. 1} \]

Where:

\[ P_{ni}(X=1) = \text{The Probability of a Person } n \text{ achieving a score } (X) \text{ of 1 on an item } i, \text{ i.e., probability of getting the item right.} \]

\[ f(\theta_n - \delta_i) = \text{A function } (f) \text{ of the difference between a person’s ability } (\theta_n) \text{ and an item’s difficulty } (\delta_i). \]

In estimating these parameters, a statistical model is developed for Equation 1. In addition, the equation is transformed into log likelihood estimates in order to correctly model the non-linear relationship (Figure 3.1) between a person’s ability and the difficulty of an item. Therefore, the statistical Rasch model is portrayed in Equation 2:

\[ P_{ni}(X = 1/\theta_n, \delta_i) = \frac{e^{(\theta_n - \delta_i)}}{1 + e^{(\theta_n - \delta_i)}} \quad \text{Eq.2} \]

Where:

\[ P_{ni}(X = 1/\theta_n, \delta_i) = \text{The probability of a person } n \text{ achieving a score } (X) \text{ of 1 on an item } i \text{ given their ability estimate } (\theta_n) \text{ and the estimate of item difficulty } (\delta_i). \]
\[ e^{(\theta_n - \delta_i)} = \text{Naperian constant (e)} \text{ or natural log function (2.7183) raised to the power of the difference between a person’s ability (\( \theta_n \)) and item’s difficulty (\( \delta_i \)).} \]

\[ 1 + e^{(\theta_n - \delta_i)} = \text{One plus the Naperian constant (e) raised to the power of the difference between a person’s ability (\( \theta_n \)) and item’s difficulty (\( \delta_i \)).} \]

The logistics function (natural log of the odds ratio) in Equation 2 has certain beneficial properties and the function aptly models the non-linear relationship graphically portrayed in Figure 3.1 between a person’s ability level and the difficulty of an item. As an item’s difficulty becomes increasingly greater than a person’s ability level, the probability or chance of the person getting the item right tends toward zero i.e., \( \frac{e^{(-\infty)}}{1 + e^{(-\infty)}} \to 0; \)

similarly, as a person’s ability level increasingly exceeds the item difficulty level, the probability of the person getting the item right tends toward one i.e., \( \frac{e^{(-\infty)}}{1 + e^{(-\infty)}} \to 1. \)

Therefore, the model is cumulative because, as the person’s ability increases relative to the item difficulty, the probability of answering the item correctly also increases. As denoted on the graph in Figure 3.1, when a person’s ability level is equivalent to the item difficulty (\( \theta_n - \delta_i = 0 \)), the probability of getting the item right is 0.5, given that

\[ \frac{e^{(0)}}{1 + e^{(0)}} = \frac{1}{2}. \]
The denominator can be considered a normalizing factor as it ensures that the probability of getting an item right ($X=1$) plus the probability of getting an item wrong ($X=0$) sum to one i.e., $P\{X_{ni} = 1\} + P\{X_{ni} = 0\} = 1.0$ In addition, with $\theta_{n} > 0$ and $\delta_{i} > 0$, then the probability of getting an item right is between 0 and 1; that is, $0 < P\{X_{ni} = 1\} < 1$ is governed by the model in Equation. 2.

From Equation 2, it can be seen that the only influence on production of the data matrix are the person abilities ($\theta_{n}$) and the item difficulties ($\delta_{i}$) i.e., with the exception of these parameters, the $X_{ni}$’s are modeled independent of one another. The raw scores of persons (marginal sum of the rows) and the raw scores of items (marginal sum of the
columns) are sufficient statistics for estimating either the person estimates or item estimates; in the estimation procedure, one parameter is conditioned out of the other (referred to as “specific objectivity) and as a result of the partitioning of the outcome latent space (total person scores or total item scores), there is sufficient information to estimate each parameter separately. The estimation of item locations is free of the abilities of persons used to estimate these measures (“sample free”); similarly, the calibrations of the persons are free of the characteristics of the items used to estimate the calibrations (“test free”). In the unidimensional model, this means that the persons with identical raw scores on a test will have the same ability estimates (Wang, Yao, Tsai, Wang, & Hsieh, 2006). As mentioned, the sample free and test free features of the Rasch-based tests are particularly beneficial if one wants to measure the impact of an intervention pre and post and for measuring change over time (for example, students’ learning progressions).

The unit of measurement resulting from the natural log transformation of person responses results in separate ability and item difficulty estimates called logits (Ludlow & Haley, 1995); this transformation expands the theoretical ability range from negative infinity to plus infinity with most estimates falling in the range of -4 to +4 logits (Ludlow & Haley, 1995). Items can be similarly interpreted in logits with a theoretical range of negative infinity to positive infinity; items with a positive logit are, on average, more difficult to endorse than items with negative logits (Ludlow & Haley, 1995). The persons and items are placed on a common continuum (the scale metric axis of the variable map) and as such, the persons can be characterized by their location on the continuum by the
types and level of items of which they are associated. Person expected responses can be
compared to their observed responses to determine if “the logit estimate of ability
corresponding to an original raw data summary score is consistent or inconsistent with
the pattern expected for that estimate of ability” (Ludlow & Haley, 1995). As mentioned,
by taking the natural log of the odds ratio, stable replicable information about the relative
strengths of persons and items is derived with equal differences in logits translating into
equal differences in the probability of getting an item right no matter where on the scale
an item is located; this interval-level unit of measurement is a fundamental assumption of
parametric tests.

The Partial Credit model for Likert-based responses. The partial credit model
(Masters, 1982; Wright & Masters, 1982) was used in the ER study (Peoples, O’Dwyer,
Shields, & Wang, in review) to model students’ Likert responses to the NOS items as it is
appropriate for use with rating scales (Embretson & Reise, 2000). The partial credit
model (PCM) is an extension of the basic Rasch model. Four response options were
provided for students who completed the NOS survey; the four response options were:
“disagree a lot”; “disagree a little”; “agree a little” and “agree a lot”. Assuming the data
fit the model, these ordered ratings (scored zero for “disagree a lot” to three for “agree a
lot”) are transformed using the PCM and result in an interval-level scale (Wright & Mok,
2000). In all models represented in this study, a higher overall score is associated with
greater understanding of the latent trait; i.e., a higher score signifies that students have a
more informed view of the NOS.
In using a rating scale to operationalize the NOS construct, student understanding is also defined between two adjacent categories of an item e.g., between “disagree a little” (scored 1) and “agree a little” (scored 2). This provides the qualitative and quantitative interpretation of their NOS understanding (Wright & Mok, 2000). The threshold or step between two categories is the point at which the probability of the student choosing the next category up (in this example, choosing a 2 instead of a 1) is higher than the probability of the student choosing the previous category (Wright & Mok, 2000). Similar to the overall scale, at the item level, students who score higher (endorse a two instead of a one on the category rating scale) are considered to have a more informed view of the NOS content associated with the item.

The step function in the PCM is allowed to vary across survey items. In other words, the PCM makes no assumptions about the relative difficulties of each threshold within an item (Embretson & Reise, 2000). The PCM model is summarized in Equation 3:

\[
\ln \left( \frac{P_{n_{ik}}}{P_{n_{i(k-1)}}} \right) = \theta_n - \delta_i - \tau_{ik}
\]

Eq. 3 (partial credit model)

Where:

- \( P_{n_{ik}} \) = probability of a response in category \( k \) of item \( i \) for person \( n \)
- \( P_{n_{i(k-1)}} \) = probability of a response in category \( k - 1 \) of item \( i \) for person \( n \)
\[
\ln\left( \frac{P_{nk}}{P_{n(k-1)}} \right) = \text{log odds of the probability of a person’s response in category } k \text{ of item } i \text{ compared to category } k - 1
\]

\[\theta_n = \text{person’s ability estimate on the latent trait}\]

\[\delta_i = \text{item } i \text{'s overall difficulty}\]

\[\tau_{ik} = \text{the } k\text{th threshold difficulty for item } i\]

The PCM reports threshold parameters for each item in the scale with the number reported for each item equal to one less than the number of response categories in the Likert scale.

Conquest software can fit partial credit Rasch-based models that are composed of up to fifteen dimensions (Wu, Adams, Wilson, & Haldane, 2007). In this dissertation, Conquest software (Wu, Adams, Wilson, & Haldane, 2007) was used to fit students’ NOSI-E item responses to one dimension (unidimensional form) and to five dimensions (multidimensional form) of the Multidimensional Random Coefficients Multinomial Logit Model (MRCLM). The syntax for these models can be found in Appendix 2.

**The Multi-dimensional Random Coefficients Multinomial Logit model.** The Multidimensional Random Coefficient Logit Model (MRCLM) specifies the functional form for the items of the NOSI-E. The MRCLM is portrayed in Equation 4 and Equation 5. The principal distinction between the unidimensional form and the multidimensional form of the MRCLM is the number of dimensions specified by the researcher. In this dissertation, a unidimensional MRCLM was compared to a five dimensional MRCLM. In
the case of the unidimensional form of the MRCLM, all items were assigned to a single
dimension where \( \theta_{nd} \) is the latent ability \( \theta_n \) of student \( n \) on dimension \( d \) (\( d \) is
assigned a value of one in the unidimensional form of the model). In contrast, the
between-items multidimensional specification of the MRCLM places NOSI-E items on
five separate but correlated dimensions. This between-items model is similarly
represented in Equation 4 but extends the unidimensional form by adding \( D \) latent ability
traits \( \theta_{nd} \) to account for students’ NOSI-E responses. Using notation from Adams,
Wilson, & Wang (1997) and Briggs & Wilson (2003), the MRCML model is expressed in
Equation 4 and Equation 5 below:

\[
\ln\left( \frac{P_{nik}}{P_{n(k-1)k}} \right) = \theta_{nd} - \delta_{ik} \tag{Eq. 4}
\]

\[
\ln\left( \frac{P_{nik}}{P_{n(k-1)k}} \right) = b_{ik}'\theta_n + a_{ik}'\xi \tag{Eq. 5}
\]

Where:

\( P_{nik} \) = probability of a response in category \( k \) of item \( i \) for person \( n \)

\( P_{n(k-1)k} \) = probability of a response in category \( k-1 \) of item \( i \) for person \( n \)

\( b_{ik}' \) = score vector given to category \( k \) of item \( i \) across \( D \) latent traits.

\( a_{ik}' \) = design vector given to category \( k \) of item \( i \) that describes the linear
relationship among the elements of \( \xi \)
\[ \theta_{nd} \] = person n’s level on D latent traits \((\theta_{ni}, \ldots, \theta_{nid})\) which represents a random sample from a population with a multivariate normal distribution, \(g(\theta_n; \Sigma; \mu)\)

Where: \(\mu\) = mean vector

\(\Sigma\) = variance-covariance matrix for \(g\)

\(\xi\) = response pattern vector of difficulty parameters that describe the items.

Similar to the simple PCM shown in Equation 3, the items’ response structures in the MRCML are assumed to be ordinal with monotonically increasing response categories; these categories are indexed by \(k\). Each item, \(i\), relates to either one domain \(d\) (unidimensional form) or to five dimensions of the NOS construct \((d = 5;\) multidimensional form). Students denoted by \(n\) respond to each item \(i\) and the log odds of the probability of student’s response in category \(k\) of item \(i\) \((P_{ni(k)})\) compared to category \(k-1\) \((P_{ni(k-1)})\) is modeled as a linear function of their latent ability on the dimension(s) \((\theta_{nd})\) and the relative difficulty of category \(k\) \((\delta_{ik})\). The MRCML assumes that one or a set of D traits \((\theta_{ni}, \ldots, \theta_{nid})\) underlies the person responses. The MRCML employs a design matrix and scoring matrix to properly specify the functional form of the model.

The \(b_{ik}\) and \(a_{ik}\) components denoted in Equation 5 are not model parameters, but are instead specified by the researcher to represent the functional form of the model which maps the items to their hypothesized dimension(s), D, and difficulty vector, \(\xi\) (scoring matrix). Depending upon the specification, it assumes that one or a set of D traits
underlies the person responses. The design vector \( a_{ik} \) describes the linear relationship between items on the response vector \( \xi \) (Figure 3.2). The item and category parameter represented by \( \delta_{ik} \) in the unidimensional form is a scalar value; in the multidimensional form, the item and category parameters are subsumed in the vector parameter \( \xi \) (a D by 1 column vector with D representing the number of hypothesized dimensions in a given instrument). The scores across D dimension(s) are collected into the column vector \( b_{id} \), and then collected into the sub-matrix for item \( i \) and lastly into the scoring matrix B (Figure 3.3) for the whole test. With the mean \( \mu \) constrained to zero, the identification of parameters \( \xi \) and \( \sum \) (variance-covariance matrix of person level latent trait on D dimensions) can occur. Once this calibration of the instrument is performed, maximum likelihood estimates of the individual persons can be obtained from the mean response vector. Each person’s position in the D dimensional latent space is measured by a profile of latent trait estimates (vector: \( \theta_{n1} \ldots \theta_{nD} \)).

For the multidimensional form of the MRCML model, Lederman’s (Lederman, 1992; Lederman, Abd-El-Khalick, Bell, & Schwartz, 2002; Lederman, 2007) theoretical framework governing the NOS construct guided how the design and score matrices were set up. Each item was developed to measure students’ ability on only one subscale. In the between-item multidimensional model, the dimensions are related but the items are only related to one dimension with none common across dimensions. The between-item multidimensional model was therefore ideal to model Lederman’s theoretical framework for the NOS construct and was used to assign items to each of the five NOS dimensions.
in Conquest’s computer program (Wu, Adams, Wilson, & Haldane, 2007). The alternative, a within-item multidimensional model (and its associated design and scoring matrices), is appropriate if more than one latent ability is needed to answer an item. This type of multidimensional model is not aligned to Lederman’s theoretical framework (Lederman, 1992; Lederman, Abd-El-Khalick, Bell, & Schwartz, 2002; Lederman, 2007) and was not considered in this dissertation.

Figure 3.2 and Figure 3.3 portrays the design and scoring matrices, respectively, for a between-item multidimensional form of the NOS construct. To keep the representation of the matrices manageable, the model is based on two dichotomous items per NOS dimension. The design matrix provides the association between items (rows) and the model parameters (columns). In this exemplar, ten parameters are estimated, one for each item. This results in a design matrix of ten columns (Figure 3.2). The scoring matrix (Figure 3.3) which represents the association between items (rows) and dimensions (columns) has one column for each of the five NOS dimensions. In a dichotomous model, items are usually scored zero or one with zero acting as the reference category.

Figure 3.3 reflects that the first two items (four rows) are scored on dimension 1; the next two items mapped and scored on dimension 2 etc., with the last two items scored on dimension 5. The flexibility of the design and scoring matrices enables the MRCML to accommodate different types of data; dichotomous data (shown here) and Likert-scale data (Briggs & Wilson, 2003). In this study the multi-dimensional model accommodates
the PCM and this will be compared to the results from the PCM used for the unidimensional and consecutive models.

In layman’s terms, the between-item multidimensional model hypothesizes that the NOS construct domains are correlated and a high ability estimate on one domain increases the probability of having a high ability estimate on another domain (practically, a student can use their knowledge of one domain to help respond to items from another domain). Cheng, Wang and Ho (2009) suggest that the correlation between the subscales
provides “collateral information” (p. 374) about the persons and this ‘borrowed’ or additional information helps increase the precision of item and person measures; the authors view this additional information as similar to when researchers use person demographics and contextual educational background information to improve parameter estimates such as in value-added models. By relinquishing the orthogonality constraint imposed by the consecutive approach, the multidimensional model provides more precise parameter estimates for students’ scores for each of the five distinct ability domains than would be obtained using the consecutive approach.

The correlations between latent traits are often of interest to researchers and are helpful in evaluating the structural validity aspect of constructs (Cheng, Wang, & Ho, 2009). Using the consecutive approach, person estimates are derived for each dimension or subscale calibration process and then the Pearson correlations are computed; this process introduces measurement error and the resulting correlations are attenuated. In contrast, the multidimensional approach simultaneously calibrates all dimensions, adjusts for measurement error and produces an accurate estimate of the correlations between dimensions as the correlations are disattenuated. The multidimensional approach therefore can provide the researcher with more accurate information and support for the validity claims for the dimensional structure of their construct (Liu, Wilson, & Paek, 2008; Cheng, Wang, & Ho, 2009).

Wright and Stone (1979, p.133) posit that the “best test” is one that provides the most precise estimates “in the region within which measurements are expected to occur”. To produce the smallest standard errors of measurement for person estimates in a test, the
difficulty of test items should be targeted at the center of the person distribution; be of a range to adequately cover the target distribution and the test should have a sufficient number of items to provide the desired precision for the test (Wright & Stone, 1979). Researchers cite that the drawback of the consecutive approach is that it reduces coverage, does not use all the available data and the smaller number of items used to form unidimensional subscales results in larger person estimates for each domain (Adams, Wilson, & Wang, 1997; Briggs & Wilson, 2003; Wang, Yao, Tsai, Wang, & Hsieh, 2006; Cheng, Wang & Ho, 2009). The benefit of the multidimensional approach is that it models and uses the interrelationship between the construct dimensions to improve the reliability of students’ estimated scores for each latent trait (Wang, Yao, Tsai, Wang, & Hsieh, 2006; Cheng, Wang, & Ho, 2009) and “explicitly recognizes the test developer’s intended structure” (Adams, Wilson, & Wang, 1997, p. 11). With the increased accuracy of the parameter estimates, researchers have found the estimates produced by the multidimensional model to be more reliable than those produced through the consecutive approach and nears the reliability of those produced by the composite unidimensional approach (Briggs & Wilson, 2003; Allen & Wilson, 2006; Wang, Yao, Tsai, Wang, & Hsieh, 2006; Cheng, Wang, & Ho, 2009).

In the multidimensional model, the raw scores of each subscale are no longer sufficient statistics as two persons with the same raw scores may not necessarily have the same ability estimate for the latent trait of the subscale. This is a consequence of allowing the subscales to correlate with the ability estimate of a person on one subscale being dependent on not just their raw score for that subscale but also on their raw scores for the
other subscales (Wang, Yao, Tsai, Wang, & Hsieh, 2006). Therefore, in the MRCML model, the sufficient statistic for the person measures is the vector of raw scores (in this study, all raw scores over the five NOS subscales); the only time when students will have the same person measures is when they have the identical pattern of raw scores across the five NOS domains (Wang, Yao, Tsai, Wang, & Hsieh, 2006).

For a measure to be unidimensional, the items should be independent and explained by one latent trait. The performance on any one item should not be dependent on any others in a survey or test. Therefore, an important assumption of the Rasch model is one of item local independence. Local independence reasons: conditional on a person’s ability level (locale), the person’s responses to any two items of the survey or test are independent i.e., within the subpopulation of persons at the same location on the latent trait, item responses are statistically independent (Van Alphen, Halfens, Hasman, & Imbos, 1994). The dimensionality of a survey or test is equivalent to the number of latent traits required to achieve local independence. Mis-specification of a model as unidimensional when the items, in truth, rely on multiple latent traits will violate the local independence assumption and lead to biased item parameter estimates and standard errors that are too small (Reckase, 1985). Local independence and the unidimensionality of a scale can be tested empirically and is important to establish if a researcher is claiming that their scale measures a single underlying latent trait.

The data analyses procedure section presents a detailed account of how the dimensional structure of the NOS construct was assessed using the models discussed in this section. This assessment was guided by Messick’s unified validity conceptual
framework (Messick, 1995). First the data sources and hypothesis underpinning the analyses is provided.

**Data Sources**

The development of the NOSI-E was biphasic: the first phase used the pilot and Year 1 data to inform item ideation and scale development efforts. The second phase used the combined data from Year 2 and Year 3, anchored on Year 1’s metric, to produce the final items for the scale. A multi-dimensional Rasch model in particular requires a sufficiently large sample in order to provide precise item and person parameter estimates for each dimension. To address this reality, in this dissertation, all the data (Year 1, Year 2 and Year 3) will, for the first time, be combined to assess which Rasch model offers the best representation for the internal structure of the NOS construct.

**Sample characteristics, sample size and sample parameter estimates.** As highlighted in Chapter 2, the data included in this dissertation were collected over a period of three years (spring 2008 to spring 2011) as part of an NSF-funded project titled, “Evolution Readiness: A Modeling Approach” (ER project). Data were obtained in Year 1 from students within the classrooms of ten participating elementary teachers at the end of their regular 6-week life science units (this cohort of students provided the baseline data for the three measures of the study). In Year 2 and Year 3 of the cohort-based research project, new cohorts of students within these same ten classrooms were exposed to the evolution curriculum intervention and were assessed on the NOS measure. All students for this project were from the target elementary grade, that is, fourth grade. The
students numbered 115, 163 and 154 in Year 1, Year 2 and Year 3, respectively.

Summating the N’s across the three years provides a total possible sample size of 432 students for the study; the profile of the sample, broken down by cohort, is shown in Table 3.1.

Table 3.1

Profile of Samples used in the Study

<table>
<thead>
<tr>
<th>Samples</th>
<th>Number of Students</th>
<th>Number of Schools</th>
<th>Number of States</th>
<th>Number of Classrooms</th>
<th>Grade</th>
<th>% Male</th>
<th>% ELL</th>
</tr>
</thead>
<tbody>
<tr>
<td>YEAR 1</td>
<td>115</td>
<td>4</td>
<td>3</td>
<td>10</td>
<td>100%</td>
<td>55.6</td>
<td>NA</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>G4</td>
<td></td>
<td></td>
</tr>
<tr>
<td>YEAR 2</td>
<td>163</td>
<td>4</td>
<td>3</td>
<td>10</td>
<td>100%</td>
<td>50.0</td>
<td>NA</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>G4</td>
<td></td>
<td></td>
</tr>
<tr>
<td>YEAR 3</td>
<td>154</td>
<td>4</td>
<td>3</td>
<td>10</td>
<td>100%</td>
<td>48.1</td>
<td>35.7</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>G4</td>
<td></td>
<td></td>
</tr>
<tr>
<td>TOTAL</td>
<td>432</td>
<td>4</td>
<td>3</td>
<td>10</td>
<td>100%</td>
<td>50.7</td>
<td>---</td>
</tr>
</tbody>
</table>

The ER project took place in three states; one Southern, one Northeastern and one from the Midwest. To ensure the confidentiality of the respondents’ data, limited demographic information was collected. The students were, on average, 50.7% male; in Year 3, 35.7% of the students were English Language Learners. To provide additional context for the sample, aggregate characteristics of respondents’ schools are summarized. One hundred percent of the sample resided in urbanized clusters (populations of greater than 2,500 people but fewer than 50,000 people). The students attend schools whose racial profiles differ: two schools have a White population of between 60 and 70%; one
school has a White population of 30% with the remaining school having a White population of less than 2%. Students came from schools with a relatively low socio-economic status with three of the schools having between 45 and 60% of their student body eligible for free or reduced price lunch; in the remaining school, 94% of the student body was, on average, economically disadvantaged.

The research in this dissertation is based on the combined data from Year 1 (N = 115), Year 2 (N = 163) and Year 3 (N = 154) of the ER project. The invariance of the sixteen items common to the three “test” administrations were invariant indicating that the responses from the three years could justifiably be combined. Combining the data from the three cohorts was required in order to ensure a sufficient sample size for the analyses. As Liu (2010, p. 51) posits, the “issue of required minimal sample size is essentially an issue of standard error of measures (Rasch person and item parameter estimates).” For high-stakes decisions, a standard error (SE) of 0.15 or below is deemed acceptable with a SE of 0.25 considered adequate for low-stakes decisions (Liu, 2010). Using Wright’s (1977, p. 223) formula for computing sample sizes (N), needed for a desired level of precision (N = 6/SE^2), a sample size of 432 will provide standard errors for parameter estimates of 0.12, one suitable for high-stakes decisions. Wright’s formula is, however, based on dichotomous data. Wolfe and Smith (2007b, p. 210) recommend that there are at least 10 observations per response category in rating scales to ensure stable parameter estimates. In the partial credit model used here, items delimit their own rating scale. Linacre (2002), therefore, suggests that to ensure stability of estimates and minimally achieve 10 observations per category across the scale, at least 25*(m+1) to as
many as 100*(m + 1) respondents are needed (where m is the number of steps in the rating scale). There are 3 steps in the NOS partial credit rating scale which, based on Linacre’s (2002) suggested sample size criteria, indicated a sample size of between 100 and 400 respondents was likely needed to produce stable estimates. By combining the data from the three cohorts, the upper benchmark was exceeded and provides support that the parameter estimates used to assess the internal structure of the NOS construct will be reliable.

Item and person parameter estimates produced by the MRCML were used to compare and assess the internal structure of the NOS construct. Marginal Maximum Likelihood (MML) and expectation-maximization methodology is used by Conquest’s software (Wu, Adams, Wilson, & Haldane, 2007) to estimate parameters for each of the models compared. Conquest’s default estimation algorithms and assumed population distribution (multivariate normal) were used to compare parameter estimates from the unidimensional form and multidimensional forms of the MRCML. The dimensional structure specified for the items determines how the maximum likelihood function estimates the parameters. The unidimensional form maximizes the likelihood function using fixed quadrature points; in contrast, a Monte Carlo method, in which the quadrature points are adaptively readjusted based on prior estimates, is used by Conquest when the number of dimensions being modeled exceeds three (Adams, Wilson, & Wang, 1997; Wu, Adams, Wilson, & Haldane, 2007). Once the parameter estimates are computed, individual expected a posteriori (EAP) person estimates are calculated by deriving five plausible values from the marginal posterior distribution. In addition, Conquest provides
estimates of population means, variances and covariances for each dimension (Wu, Adams, Wilson, & Haldane, 2007). These estimates provided the fundamental evidence needed to assess the structural and external validity aspects of the NOS construct and, by corollary, the hypothesis related to the research question of this thesis.

**Hypothesis Related to Research Question 1**

One hypothesis, related to the research question, was tested in this thesis; the hypothesis was:

*The internal structure of the NOS construct is best represented by the multidimensional Rasch model. In comparison to the unidimensional and consecutive models, scores from the multidimensional model are more reliable, interpretable and suitable for use in science education research.*

**Data Analyses Procedures**

To create a reliable measure of student understanding of the NOS and address the hypothesis presented was a complex undertaking and involved a sequential step by step process. This process was framed by a validity argument designed to evaluate the hypothesis. This validity argument is summarized in Figure 3.4. To justify the claim that the multidimensional model best represents the NOS construct, four sets of empirical analyses were required. These analyses and associated claims build upon each other to provide evidence that the scores produced by the multidimensional model were the most
reliable, interpretable and suitable for use in science education research. Messick (1995, p.747) posits that “empirical evaluation is meant to convey that the validation process is scientific as well as rhetorical and requires both evidence and argument”.

At each step of building the validity argument, propositions are constructed and evidence is provided to support the inferences made; the arrows in Figure 3.4 denote that the inferences and supporting evidence are unidirectional with each step forming the foundation for the next argument made with the ultimate goal of providing enough inferential evidence to support the use of NOS dimension scores in science education research. The type of validity evidence and related analyses needed to sustain the validity argument that the multidimensional Rasch model best represents the internal structure of the NOS construct and provides scores that support the instrument’s intended use in science education research is summarized in Figure 3.4 and outlined below. Critical to building the validity argument is providing evidence from the literature that supports and justifies any claims made about this study’s findings. Using evidence from the literature is highlighted by including it in Figure 3.4 at each stage. However, appropriately, the findings from the literature will only be reported in Chapter 4 when the results are discussed.
CONCLUSION: The internal structure of the NOS construct is best represented by the multidimensional Rasch model. The scores provided by the model are reliable, interpretable and suitable for use in science education research and teaching.

Analyses 4: External Validity Evidence
- Responsiveness – Variable Maps & Number of Person Strata.
- Multi-Level Regression Model – NOSI-E and CIER Scores.
- Multi-Level Regression Model – NOSI-E and ESSCES Scores.
- Literature Resources.

Analyses 3: Structural Validity Evidence
- Dimensionality Analyses – Rasch Goodness of Fit.
- Dimensionality Analyses – Confirmatory Factor Analyses.
- Rasch Subscale Correlations – Convergent Validity.
- Discrepant Case Analyses – Person Estimates.
- Literature Resources.

Analyses 2: Generalizability Evidence
- Item Invariance – Delta Correlations across Models.
- Differential Item Functioning – Gender.
- Reliability – Person Separation (Spearman-Brown Prophecy).
- Precision of person estimates comparison.
- Literature Resources.

Analyses 1: Content & Substantive Validity Evidence
- Item Technical Quality – Model Fit.
- Rating Scale Functioning – Category Functioning.
- Item Difficulty Hierarchy – Theoretical Expected Ordering.
- Literature Resources.

A. The multidimensional approach to modeling the internal structure of the NOS construct competes with the unidimensional and consecutive approaches. The analyses specified above will determine if the multidimensional approach best represents the internal structure of the NOS construct.

B. Based on Lederman’s theoretical framework, the NOS construct is composed of five inter-related domains/dimensions: Empirical; Theory-laden; Certainty; Inventive and Social and Culturally Embedded.

C. Based on Lederman’s theoretical framework, each item developed is related to only one domain.

Source: Validity argument model based on Chapelle & Jamieson (2010, p. 10)

Figure 3.4
Validity argument for the NOS construct
Premise of Validity Argument. The hypothesis that scores produced by the multidimensional Rasch model are more reliable, interpretable and suitable for use in science education research than scores produced by the unidimensional model or the consecutive model premises the validity argument outlined in Figure 3.4. These three models, unidimensional, consecutive and multidimensional models, each represent an alternative internal structure for the NOS construct. However, based on Lederman’s theoretical conceptualization for NOS (Lederman, 1992; Lederman, Abd-El-Khalick, Bell, & Schwartz, 2002; Lederman, 2007), it is hypothesized that the NOS construct is composed of five inter-related but separate dimensions, namely, Empirical, Theory-laden, Certainty, Inventive and Socially and Culturally Embedded which is best represented by the multidimensional Rasch model.

Analyses 1: Content and Substantive Validity Evidence. Before the structural validity aspect of the NOS construct could be assessed, the content and substantive validity aspects of the NOS construct needed to be examined. To support the interpretation and inferences made concerning the dimensionality of the NOSI-E and to address the research question and its related hypothesis, it was important to first establish that the content and substantive aspects of NOS construct validity reported and justified during the scale’s development (Peoples, O’Dwyer, Shields, & Wang, in review) were still supported regardless of the Rasch model used to represent the internal structure of the NOS construct. With the exception of the first four tasks under the content validation column in Table 2.2, these two validity aspects were assessed for the unidimensional, consecutive and multidimensional Rasch models using all three years of the ER data.
Specifically, to justify and make the claim that the multidimensional model best represents the NOS construct, the evaluation of this claim begins with assessing the content (technical quality of items) and substantive validity aspects (rating scale functioning and expected item hierarchy) of the NOS construct. These analyses help address the research question by providing evidence for the interpretability of the scores across the three Rasch models. The (a) technical quality of the items; (b) item rating scale category functioning and (c) expected hierarchical ordering of the items was compared and evaluated across the three Rasch models. The analyses conducted for these two validity aspects were:

(a) Fit statistics (mean square and t-statistics), comparing the observed and expected response patterns, were evaluated to measure the extent that items conform to the Rasch model (Wright & Mok, 2000; Smith, 2002) in each of the three representations. An item was deemed misfitting in this study if the infit or outfit mean square error was greater than 1.3 or less than 0.7. These misfit criteria are accepted practice (Wright & Linacre, 1994) and used widely (Wang, Yao, Tsai, Wang, & Hsieh, 2006; Bond & Fox, 2007; Liu, Wilson, & Paek, 2008; Cheng, Wang, & Ho, 2009). Convention also indicates that an item is misfitting if standardized t-statistics are greater than 2.0 or less than -2.0; however, these fit statistics are susceptible to sample size (Linacre, 2003) with samples over 250 flagging items with mean squares of less than 1.3 and greater than 0.7 as significantly misfitting. Both fit statistics are reported in
this dissertation but more credibility is placed on the mean square error statistics, given the sample size of 432 in this study.

(b) The functioning of the rating scales was compared across the three Rasch models to assess the substantive aspect of NOS construct validity. The expected performance of a well-functioning item is that the steps (thresholds) are monotonic; that is, increasing levels of the latent trait are associated with endorsement of more affirmative categories. Evidence of this for the sample was derived by examining the reported average of measure difference (ability estimate for person – item difficulty) for each category to determine if it increases monotonically up the rating scale i.e., from category 0, 1, 2 to category 3. In general, observations in higher categories are derived from higher measures; if this does not occur, the meaning of the rating scale is doubtful and its use called into question (Linacre, 2002). This essential expectation should hold regardless of the Rasch model fitted to the data.

The average of the measures highlighted above pertains to the sample under analysis. To obtain evidence that the rating scale could be used for future samples and have inferential interpretability, item category delta parameters were examined. The deltas are the points at which two adjacent probability curves intersect and these deltas should increase monotonically up the rating scale. The deltas were assessed to determine whether an increasing level of the latent trait within respondents translated into an increased probability of the respondents being observed in higher categories of the
rating scale. Disordering of the category deltas suggests that a higher category is not being used as frequently (i.e., low probability of observance) when compared to a lower category or is indicative of problems with the average measures with lower categories being chosen by respondents with higher measures. Disordinality reduces the interpretability of the resulting person measures as it likely indicates that “a concept is poorly defined in the minds of the respondents” (Linacre, 2002, p. 95).

(c) The average population parameter estimates and average item difficulty parameters (deltas) produced by Conquest (Wu, Adams, Wilson, & Haldane, 2007) were used to assess if the items were behaving as anticipated in terms of expected item difficulty hierarchy. The average population parameter estimates, when possible, were used to ascertain if the expected ordering of the dimensions conformed to theoretical expectations on the continuum. This analysis is not possible for the unidimensional approach as it only produces one population mean for the NOSI-E. Similarly, average item difficulty (average of item deltas) combined with a qualitative content assessment was used to determine if the item difficulty hierarchy within each dimension(s) met theoretical expectations across the three Rasch models.

Confirmation that (a) the technical quality, (b) the rating scale functioning, and (c) the expected item difficulty hierarchy of the NOS items is comparable across Rasch models provides evidence to support any inferences made from using the scores derived
from each of the Rasch models. Disconfirmation of either the content or substantive validity evidence in any one of the models would suggest that scores from the model in question were not suited for use in science education research as score interpretations could be called into question. Before addressing the structural validity aspect of the NOS construct (at the heart of this dissertation), another validity aspect was examined. It is important to establish that the scores derived from each Rasch model are generalizable and reliable; this will provide support for the use of the scores across multiple contexts and help determine if scores from the multidimensional model are sufficiently reliable to be used by science education researchers. The generalizability aspect of NOS construct validity is addressed by the next set of analyses.

**Analyses 2: Generalizability Evidence.** The concept of generalizability in this dissertation is limited in that it was focused on the internal stability of the measure and not on the external stability of the measure. Generalizability theory (Crocker & Algina, 2006, p. 234) assumes that “each measurement taken from an individual represents a random sample form the universe of possible measurements that could have been obtained”. It was beyond the scope of this dissertation to assess whether facets such as mode of administration and time of data collection could influence student responses. The rationale for this set of analyses was to ensure that each NOSI-E item retained its meaning and quality regardless of the Rasch model used; and to assess the reliability of the NOSI-E items across these models. These analyses provide evidence to address the research question and related hypothesis of the study in that they will help determine
which model provides the most reliable and interpretable measure of elementary 
students’ NOS understanding.

Item invariance across multiple contexts is an important assumption of the Rasch 
model (Wolfe & Smith, 2007b; Bond & Fox, 2007) and ensures that the items used to 
measure a construct have the same meaning and interpretation across subgroups and 
contexts. To support the hypothesis that the multidimensional model is the best model to 
represent the NOS construct, evidence is required to show that the items are invariant 
across contexts when the items are configured within this internal structure. The 
generalizability of the measure holds if the item invariance assumption is met and 
provides support for the construct validity of the measure. Item invariance was examined 
by (a) using Pearson correlation analyses to establish the invariance of item parameter 
estimates across the three models, (b) using gender as a grouping variable in Differential 
Item Functioning (DIF) analyses (c) using reliability analyses to assess the level of error 
variance in each model and (d) by assessing the stability of person measures. The 
analyses conducted were:

(a) The correlational analyses were designed to establish the invariance of item 
difficulties across the three models (unidimensional, consecutive and 
multidimensional); Pearson correlations were derived for the three sets of item 
difficulty parameter estimates. Determining that the items were invariant 
across models was particularly important to establish between the consecutive 
and multidimensional approaches. With the exception of the correlations 
between the five domains under the multidimensional approach, the make-up
and parameterization of the two models were identical. The expectation is that the items deltas would achieve near one-to-one correspondence.

By corollary, despite the difference in parameterization of the models, the unidimensional items should also be largely invariant when compared to the items of the consecutive and multidimensional models. The expectation was that the item difficulty estimates should be moderately to highly correlated (> 0.7) between the unidimensional model and consecutive or multidimensional models thereby supporting the internal consistency of the scale(s). Establishing item invariance across these three models should enable proper and valid conclusions to be made on which model best represents the internal structure of the NOS construct in the next set of analyses.

(b) In the context of Rasch modeling, DIF analyses investigate whether two or more groups that are of equal latent trait abilities have the same probability of providing a “correct” response to an item. If the probabilities differ, the item content may provide an advantage to one group leading to item bias. Hambleton (2006, p. S184) stated that “the power to detect DIF with any of the DIF statistics increases with sample size” with many of these flagged items revealing a “trivial” level of DIF. In this study, an item had to have a difference of > 0.5 logits in item endorsement between groups (holding all else constant) and this difference had to be statistically significant (after performing a Bonferroni adjustment for multiple comparisons) to be considered as a candidate for DIF. It was important to establish that
irrespective of the model used, DIF was not present. Gender was used as grouping variable; no other grouping variable were available.

(c) The reliability analyses help to establish the internal consistency and stability of the instrument(s) across the three Rasch models. Error variance can undermine score interpretation and the plausibility of inferences related to these measures (Wolfe & Smith, 2007b). In the Rasch model, the internal consistency expected \textit{a posteriori} (EAP) reliability coefficient, the person separation reliability, measures the ratio of the variance in latent (true) person measures to the estimated person measures and was used to compare the degree of error variance of the three models. Low reliability ($< 0.7$) would indicate that the items did not separate persons well making it hard to differentiate and make inferences on individual person differences. Wang, Yao, Tsai, Wang and Hsieh, 2006 (2006) determined that the increased reliability of the multidimensional approach over the consecutive approach was equivalent to an increased test length (using Spearman-Brown Prophecy formula) of 103\% (from 6 to 12.2 items) for the psychological sub-scale of the WHOQOL-BREF and of 114\% for the physical subscale. Using the Spearman-Brown Prophecy formula, a similar analysis was performed over and above the simple comparison of expected \textit{a posteriori} (EAP) reliability coefficients to determine which model is more reliable and efficient.
(d) Reliability indices are a good measure of the stability of group-based measures but are of limited use when applied to individual measures (Wolfe, 2007b). Each of the three Rasch models can provide a direct measure of error variance for each person’s ability and item’s difficulty (Schumacker & Smith, 2007). The precision of the person measures is assessed using the standard error of estimation ($SE(\hat{\theta})$: the standard deviation of true person ability estimates) estimates and can help compare the reliability of models. The amount of information in a test is inversely related to the precision of ability estimates at each location (ability level) on the continuum (Hambleton, Swaminathan, & Rogers, 1991); this is denoted in Equation 6. In Rasch methodology, the standard error of estimate, SE, varies with ability level, with smaller SE values associated with more precision. In general, there is less information and therefore less precision at the extremes of estimating the underlying latent trait with more precise estimates where the person abilities and item difficulties match (Hambleton, Swaminathan, & Rogers, 1991). The standard errors across the sample ability distribution will be compared across the models.

$$SE(\hat{\theta}) = \frac{1}{\sqrt{I(\theta)}}$$

where:

$SE(\hat{\theta})$ is the standard error of estimate, and

$I(\theta)$ is the information function.
In the Rasch framework, smaller standard errors are associated with more stable parameter estimates; a comparison of the stability of the parameter estimates across the three Rasch models was therefore used as evidence to support the argument as to whether the multidimensional model best represents the internal structure of the NOS construct.

The next two sets of analyses are at the heart of the dissertation; these analyses establish the structural (Analyses 3) and external validity (Analyses 4) aspects of the NOS construct.

**Analyses 3: Structural Validity Evidence.** To address the research question, dimensionality analyses were used to provide essential evidence of the internal structure of the NOS construct. The hypothesis related to the research question proposes that consistent with Lederman’s theoretical conceptualization of the NOS construct, the responses are best represented and explained by the multidimensional Rasch model. Lederman’s theoretical framework posits that five inter-related domains (Empirical, Theory-laden, Certainty, Inventive and Socially and Culturally Embedded) combine to form the NOS construct. This hypothesis, however, competes with two plausible, rival hypotheses: (1) the NOS construct is unidimensional, and (2) the NOS construct is best represented by five independent unidimensional sub-scales, namely the consecutive approach. Disconfirming plausible, rival hypotheses is essential to building and supporting the validity argument summarized in Figure 3.4 for the use of the multidimensional approach to represent the internal structure of the NOS construct.
Dimensionality analyses were therefore used to make explicit the internal structure of the scores produced by the NOSI-E instrument using two complementary methodologies. These analyses included (a) a comparison of goodness of model fit statistics using Rasch methodology; (b) an assessment of output and goodness of model fit statistics using confirmatory factor analyses (CFA); (c) an evaluation of Rasch-based subscale correlation analyses and (d) an appraisal of discrepant cases using the multidimensional subscale estimates. These analyses are described below:

(a) A summary goodness-of-fit statistic can be used to compare two of the three Rasch models to help determine which model best fits the NOS data. The model fit of all three models cannot be contrasted directly. The unidimensional and multidimensional approaches are hierarchical, nested models and can be compared using the likelihood ratio statistic $G^2$ (Wu, Adams, Wilson, & Haldane, 2007). The likelihood ratio test ($G^2$) is based asymptotically on the chi-square distribution with the difference in deviance between two models compared to the critical chi-square value (degrees of freedom are based on the difference in the number of parameters estimated by the models). The model with the lower deviance value is considered the better fitting model. The consecutive approach is not nested with either the unidimensional model or the multidimensional model so a goodness-of-fit comparison using the $G^2$ statistic cannot take place. To compare the consecutive model to the unidimensional and multi-dimensional models, evidence from the CFA and subscale correlation analyses were also relied
upon. In this manner, a body of evidence is used to make a decision on the suitability of the consecutive approach to represent the internal structure of the NOS construct.

(b) To support the Rasch analyses, the three theoretical representations of the internal structure of the NOS scale were examined using confirmatory factor analyses (CFA); this was performed using MPLUS (Muthén & Muthén, 2010). The syntax for the three CFA models is found in Appendix 3. In CFA, items that share variance and covariances are hypothesized to form a dimension. In each instance, a theoretical model was outlined and tested to determine if the data fit the model. These analyses are based on ordinal measures (Likert responses); as a result, the CFA use the correlation matrix structure to assess the relationship between the variables of the model. These polychoric correlations are used to determine if the relationships between the observed variables are explained by an underlying continuous latent variable (Byrne, 2012).

The unidimensional (1 factor; 28 items) model is nested within the five factor multidimensional model structure (Empirical; Theory-Laden; Certainty; Creativity and Imagination; and, Socially and Culturally Embedded). When compared to the five-factor model, the one-factor model is the more restrictive (nested) model. This means that if the inter-correlations of the five dimensions in the multidimensional model are constrained to one, the unidimensional model is derived (Liu, Wilson, &
Paek, 2008). Theoretically, the multidimensional construct will be composed of five first-order factors with the latent construct, NOS, explaining the polychoric correlations among the first-order factors. In contrast, in the one-factor model, the latent NOS construct explains the observed responses of each item of the 28 item unidimensional model. These nested models can be compared for goodness of fit using MPLUS’s DIFFTEST. Because the data are ordinal in nature, a Weighted Least Square Mean Variance (WLSMV) estimator is the default estimator for producing parameter estimates. This estimator corrects for the categorical nature of the data and the likely violation of the normality assumption required of CFA (Byrne, 2012). As a result, the chi-square statistic is scaled to take these factors into account and this scaled value is used in the CFA DIFFTEST by MPLUS to determine if the model fit between two nested models is significantly different. Similar to the Rasch methodology, the consecutive model is not a nested model and five separate CFA analyses were run to determine if the items form a factor. These five models cannot be compared directly with the unidimensional or multidimensional models, but their fit statistics can be used to assess each model in turn.

The criteria used to compare the fit of the CFA models were taken from Schumacker and Lomax (2004, p. 82, Table 5.1). The chi square goodness-of-fit test was used to assess each model in turn and, as mentioned, in comparing nested model. A non-significant chi-square suggests good
model fit with the sample correlation matrix reproducing the hypothesized correlation matrix. This statistic is susceptible to large sample sizes (Schumacker & Lomax, 2004; Byrne, 2012) resulting in Type I errors or false positives. As a result, other criteria will also be used in this study. The Root Mean Square Error of Approximation (RMSEA) is an alternative and most often cited fit statistic that measures whether the hypothesized structure recovers the sample data. The RMSEA is a standardized global measure of fit that is parsimonious as it favors models with fewer parameters. A confidence interval around the point estimate of the RMSEA is also reported and useful in determining the precision of error approximation. Schumacker and Lomax (2004, p. 82) report that a value of less than 0.05 signifies a close model fit. In addition, the Tucker-Lewis Index and Comparative Fit Index will be examined; values of above 0.95 indicate excellent model fit (Schumacker & Lomax, 2004, p. 82) with a value of 1.00 indicating perfect fit.

In each analysis, modification indices were examined to determine if the model had been mis-specified. For example, in the five-factor model, if items specified to load onto one factor cross load onto another factor, this may suggest that there are not five separate dimensions but four. Therefore, the modification indices can be used to ensure that the models are properly specified and can be compared.

(c) The Rasch-based correlations between subscores (e.g., between Empirical and Theory-laden scores) provided by the consecutive and multidimensional
approaches were compared. Researchers have determined that the consecutive approach underestimates the correlation between subscores (Brigg & Wilson, 2003; Cheng, Wang, & Ho, 2009). If the correlations found using the consecutive approach are substantially less than those found using the multidimensional approach, this would provide supporting evidence for the use of the multidimensional approach to model the internal structure of the NOS construct as the estimates derived from this approach are more reliable. Alternatively, if the two models were comparable and the correlations were low within each approach, this would provide support for using the consecutive approach to model the NOS construct.

(d) An analysis of discrepant cases (Briggs & Wilson, 2003; Allen & Wilson, 2006) was undertaken using standardized person estimates from the multidimensional model. Person ability estimates from each dimension can potentially be treated as a “general” measure of NOS and these estimates were compared across dimensions. The two dimensions with the highest score correlation (Empirical and Certainty) and the two with the lowest score correlation (Inventive and Theory-laden) were examined for the percent of students whose estimates differed by more than one standard deviation (the percent of cases whose difference in ability estimates, $\theta_{dj} - \theta_{dk}$, was greater than one standard deviation).

In addition, using the sum of squares discrepancy indicator for each student (Briggs & Wilson, 2003; Allen & Wilson, 2006), the number of
discrepant cases was also compared across all five dimensions simultaneously. The sums of squares discrepancy indicator, $DI_p$, is computed by the following formula $\sum_{d=1}^{5} (\bar{\theta} - \theta_d)^2$ where $\bar{\theta}$ is the average case estimate on the 28 item unidimensional scale and $\theta_d$ is the case estimate on each dimension, $d$. Briggs and Wilson’s (2003) and Allen and Wilson’s (2006) set arbitrary standards of 0.5 and 1.0, respectively for calculating the percentage of students whose person ability estimates were consider discrepant. Briggs and Wilson (2003, p.96) suggest that discrepant cases are indicative of “students in the sample for whom dimensional estimates might reveal differing stories about underlying student ability.” The sums of squares discrepancy indicator was calculated for students in this study and the percentage of students at these two cut points is reported in Chapter 4.

In validating their multidimensional scales, Boman, Curtis, Furlong and Smith (2006) and Wolfe and Singh (2011) found that the results from Rasch-based and CFA methodologies provided comparable confirmatory evidence of the dimensional structure and fit of their constructs and led to similar conclusions on the appropriate use of their scales. The results from these analyses along with the other analyses (Analyses 3a, 3b and 3d) combine to provide the evidence needed to determine the optimum representation of the internal structure of the NOS construct. Scores from each of the three Rasch models were then used to provide evidence for the external validity aspect of the NOS construct. For an instrument to be of use in the field, it has to have the ability to measure change in students’
understanding of the construct (Wolfe, 2007b) i.e., it has to be responsive. In addition, in the next section, external validity evidence is used to examine if the scores from the optimum Rasch model can be used to predict science achievement and students’ perceptions of their classroom learning environment.

**Analyses 4: External Validity Evidence.** External validity evidence was used to address the issue of how responsive the NOSI-E is in measuring change in student understanding of the NOS construct. The responsiveness of an instrument refers to “the degree to which an instrument is capable of detecting changes in person measures following an intervention that is assumed to impact the target construct” (Wolfe & Smith, 2007b, p. 222) and provides “evidence relating to the external aspect of validity” (Wolfe & Smith, 2007b, p.223). The responsiveness of the NOSI-E across the three Rasch models was assessed using the variable maps and the person strata index.

To further support the external validity aspect of the NOSI-E construct, student parameter estimates produced by the preferred Rasch model were assessed to determine if they were predictive of elementary students’ performance in science and were related to their views of their science classroom learning environment. Messick (1995, p.746) poses that the “meaning of scores is substantiated externally by appraising the degree to which empirical relationships with other measures – or the lack thereof – are consistent with that meaning”. Other concurrent data collected over the three years of the ER project enabled the empirical relationship between NOS construct score(s) and CIER scores; and between NOS construct scores and ESSCES scores to be assessed.
To assess the impact of the evolution readiness intervention, student content knowledge was measured using the CIER. The CIER was administered to students in each of the three years of the ER project. Student responses were placed on a Rasch scale using a partial credit model (Masters, 1982; Wright & Masters, 1982); the maximum likelihood estimates (MLEs) derived were used in these analyses. As mentioned in Chapter 2, researchers hypothesize that a better understanding of the nature of science will enhance student learning of science content (Smith, Maclin, Houghton, & Hennessey, 2000; Sandoval, 2005; NRC, 2011) and help boost student science achievement overall. This study examines if the NOSI-E scores from the preferred Rasch model are predictive of student achievement on the CIER content knowledge test. If NOSI-E scores are predictive of CIER scores, this provides support for the external validity aspect of NOS construct validity.

In conjunction with the administration of the CIER and NOSI-E, the ESSCES was administered to students. The ESSCES measures students’ perceptions of constructivist practices within the elementary science classroom. The responses were similarly modeled using a Rasch partial credit model (Masters, 1982; Wright & Masters, 1982) and the MLEs derived were also used in these analyses. Tsai (1998a) found that students who preferred a constructivist environment possessed more informed epistemic NOS knowledge; similarly, Smith, Maclin, Houghton and Hennessey (2000) found that students who were exposed to constructivist teaching practices had more informed conceptions of NOS when compared to students who were taught in a more traditional, teacher-led environment. If the NOSI-E scores are predictive of ESSCES scores, this
would provide further supporting evidence for the external validity aspect of the NOS construct.

The three sets of analyses used to assess the external validity aspect of construct validity are outlined below:

(a) The person strata index was calculated to assess the responsiveness of the NOS instrument. The person strata index, $H$, provides the number of statistically distinct ability groups whose centers of score distributions are separated by at least three standard errors of measurement within the sample (Wright & Masters, 2002). Variable maps for each model were produced to provide qualitative evidence of the coverage of the item calibrations for the person distributions for each model. The variable maps help visualize the meaning behind the number provided by the person strata index calculation. This calculation is explained in the next paragraph.

In the Rasch framework, measurement error is estimated for each person enabling disattenuated reliabilities to be calculated. The person separation reliability, $R$, is used first to calculate the person separation index which, in turn, is used to calculate the person strata index. The person separation index, $G$, measures the ratio of the variance in latent person measures to the estimated person measures (person separation reliability) in standard error units. This can be calculated using the following formula from Schumacker and Smith (2007, p. 399): $G = [R/(1 - R)]^{1/2}$ where $R$ is the person separation reliability. According to the formula, first determined by Wright and Masters (2002, p. 888) and used by
Schumacker and Smith (2007), the person strata index is calculated using the formula $H = \frac{4G + 1}{3}$ where $H$ is the number of distinct person strata and $G$ is the person separation index. The person strata index was compared across the three Rasch models. To support the conclusion that the multidimensional Rasch model best represents the internal structure of the NOS construct, the person strata index of each of the five dimensions needs to be of sufficient magnitude (>3.0). This would indicate that each subscale would be able to detect change in the construct.

(b) Multilevel regression models (MLM) were used to assess whether NOSI-E scores(s) predicted student performance on the CIER content knowledge test. Hierarchical Linear Modeling software was used for the analyses (Raudenbush, Bryk, & Congdon, 2004). To account for the hierarchical structure of the ER student data (students are nested within cohorts within teachers), multi-level regression analyses were performed. A primary assumption of MLM is that observations and their associated errors are independent of each other; in the ER project, this assumption cannot be met as students (level-1) are grouped within cohorts (level-2) and further grouped within teachers’ classrooms (level-3). As such, students grouped together in a cohort and/or teachers’ classroom behave or perform more similarly to each other than those from other cohorts and/or teachers’ classrooms resulting in an error structure that is correlated. If observations are not independent, the standard errors of the regression coefficients will be attenuated which leads to inflated t-values and type-1 errors. By
performing a MLM, this error structure is accounted for and the regression
coefficient errors are appropriately estimated. These analyses helped provide
evidence (albeit of a limited nature) for the external validity aspect of the NOS
construct.

(c) Multi-level regression models were also performed to determine if NOSI-E scores
predict students’ perceptions of the constructivist nature of their science
classroom learning environment. Again, these analyses were performed to provide
evidence for the external validity aspect of the NOS construct.

Validity Argument Conclusion. Optimally, the most useful information derived
from the data would be to have reliable student estimates for each of the five NOS
dimensions; this information it is argued would result from the multidimensional
approach to modeling NOS responses. These scores could then be used to provide
individual differential profiles of student understanding of the NOS thereby improving
the diagnostic information provided to educators and researchers. These estimates could
also be used to examine their relationship with other variables such as student
achievement to inform science education research. The validity argument in Figure 3.4
was constructed to guide the validation activities and analyses needed to justify the use of
NOSI-E scores in science education research. The tenability and utility of the scores
hinges on the inferences and evidence outlined in the argument. If the validity argument
is substantiated through these analyses, then the conclusion of the validity argument is
justified and the scores from the NOSI-E have application within science education research and within the elementary school classroom.

**Conclusion**

The analyses and methodology presented in this chapter were proposed to provide the evidence needed to assess which theoretical representation of the internal structure of the NOS data leads to a theoretically-grounded, reliable, interpretable and responsive Rasch-based measure of elementary students’ understanding of NOS. In addition, the MLMs outlined were designed to determine if scores from the measurement model have the potential to help science researchers understand the complex relationship between student science content knowledge and NOS, and between students’ classroom learning environments and NOS. Elucidating these types of relationships could provide science education researchers with important information to support student learning progressions in the core ideas and scientific practices contained in the new science frameworks (NRC, 2011). In Chapter 4, the results of these analyses, along with supporting evidence, from the literature will be presented. In Chapter 5, the significance of the results, in the context of the validity argument, will be discussed.
Chapter Four: Results and Discussion

Chapter 4 presents the results of the four sets of analyses discussed in Chapter 3 and outlined in Figure 3.4. Each set of analyses provides validity evidence used to test the research hypothesis and build support for the claim that the multidimensional Rasch model best represents the theoretical internal model for the NOS construct. The first set of analyses provides content and substantive validity evidence to support this claim. The second set of analyses compares generalizability validity evidence for each of the three Rasch models: both sets of analyses are used to address this dissertation’s research question. If the first and second set of analyses deliver comparable results across the three Rasch models, a legitimate argument can be made that it is appropriate to continue with the third and fourth sets of analyses that are at the heart of this dissertation; namely, to establish the dimensional structure of the NOS construct and to show that measures from the optimum NOS construct model relate as expected to other measures that were administered during the Evolution Readiness (ER) project.

The third set of analyses was also used to address the research question and its related hypothesis. Results are presented for the dimensionality analyses using two separate methodological techniques; both methodologies (Rasch-based modeling and Confirmatory Factor Analyses) were used as confirmatory tests of the internal structure of the NOS construct. The fourth set of analyses was designed to address the external validity evidence for the NOS construct with the goal of supporting the use of the NOSI-E scores in science education research.
Combined these four sets of analyses were used to build and support the validity argument presented in Figure 3.4, that is, the internal structure of the NOS construct is best represented by the multidimensional Rasch model and the scores derived by the model are sufficiently reliable, interpretable and suitable for use in science education research. In addition, at each stage, it was important to buttress the inferences and claims made about this study’s findings with relevant evidence from the literature. In this manner, literature resources are used as another measure of substantiating each validity aspect of the NOS construct. The results of each set of analyses follows. These analyses are based on three cohorts of data combined from the ER project, resulting in a total N of 432 students.

**Results from Analyses 1: Content and Substantive Validity Evidence**

In Chapter 3, a validity argument was proposed that the best representation of the internal structure of the NOS construct is modeled using the multidimensional Rasch approach. The argument begins with providing content and substantive validity evidence for this claim through a set of analyses; these analyses are summarized in Figure 4.1.

<table>
<thead>
<tr>
<th>Analyses 1: Content &amp; Substantive Validity Evidence</th>
</tr>
</thead>
<tbody>
<tr>
<td>a. Item Technical Quality – Model Fit.</td>
</tr>
<tr>
<td>b. Rating Scale Functioning – Category Functioning.</td>
</tr>
<tr>
<td>c. Item Difficulty Hierarchy – Theoretical Expected Ordering.</td>
</tr>
<tr>
<td>d. Literature Resources.</td>
</tr>
</tbody>
</table>

*Figure 4.1. Summary of content & substantive validity analyses*
To support this argument, it was first important to establish that (a) the technical quality of the items; (b) the item rating scale functioning and (c) the expected item difficulty hierarchy across the observable variables for the three models investigated were, at a minimum, comparable and overall psychometrically sound. Support for the findings was enhanced by relating the findings to relevant literature resources. The results and evaluation of these analyses follow.

**Item Technical Quality.** Table 4.1 and Table 4.2 compare the outfit and infit mean-square (MNSQ) error and standardized t statistics for the three Rasch models, respectively. The item content prompts highlighted in this section are summarized in Table 2.4. The outfit MNSQ error is unweighted and is computed by taking an average of the squared residuals (observed – expected). In contrast, the infit MNSQ error is calculated using a weighted average of the squared residuals: less weight is given to responses at the tail end of the distribution with more weight given to those responses proximal to the distribution. Both outfit and infit statistics were used to assess model fit across the three models. As mentioned, more credence is given to the MNSQ error fit statistics than the standardized t-statistics as they are more stable for large sample sizes (> 250) than the t-statistic; specifically, the t-statistic flags items as misfitting because they inflate the misfit in proportion to sample size (Linacre, 2003; Smith, Rush, Fallowfield, Velikova, & Sharpe, 2008). An item was deemed misfitting in this study if the absolute outfit or infit MNSQ error was greater than 1.3 or less than 0.7.

The absolute MNSQ error outfit statistics (Table 4.1) are comparable across all three Rasch models; the absolute MNSQs were all between 0.7 and 1.3. All three models
identified THL12K (Scientists create different types of experiments to answer their
questions) as the most over-fitting item (an item with negative values indicates that there
is less variation in the data than explained by the model). The over-fit was relatively
slight and all THL12K outfit MNSQs of the three models met the criteria established for
a well-fitting item. Both the unidimensional and consecutive models identified SCE13K
(Where scientists live may affect what they are allowed to work on) as the most under-
fitting or noisy item (an item with a positive fit value indicates there was more variation
in the data than explained by the model). In contrast, the multidimensional model
indicated that the random error or noise was greater for CER14A (In science, finding out
that a hypothesis is NOT correct is as important as finding out that a hypothesis IS
correct.). However, the random error for CER14A was well within the bounds to be
considered a well-fitting item and overall the fit statistics of the multidimensional model
conform to those of the unidimensional and consecutive models. Given the sample size of
432, the standardized t-statistics also substantiate that the items are well-fitting across all
three Rasch models.

In Table 4.2, the infit statistics are excellent with the absolute value of the infit
MNSQ error statistic between 0.8 and 1.2 for all items across all three models. Similarly,
with the exception of item SCE13K for the unidimensional model, the standardized t-
statistics were excellent with all absolute values between 0 and 2 across the three Rasch
models. Similar to the finding for the outfit statistics, item THL12K was the most over-
fitting item with item SCE13K the most under-fitting item for the unidimensional and
consecutive models. In contrast to the outfit statistical analyses, item THL14E (If we do
the same experiments many times, we may get different results.) was the most underfitting item (MNSQ of 1.16; T of 1.4) for the multidimensional model. The fit statistics for THL14E are, however, still excellent and comparable to those of the other two models.

In summary, the technical quality of the items was comparable across Rasch models. There is, therefore, no impediment to continuing with the analyses outlined in Figure 3.4 as the data support the content validity aspect of NOS construct validity for each of three models.

**Rating Scale Functioning.** The functioning of the rating scale was compared across the three Rasch models. It was assessed by examining the monotonicity of (1) the average difference measures \( (\theta_n - \delta_i) \) across each category for each item; and (2) the item delta parameters for each item. Table 4.3 reports the average of the difference measures for all 28 items across the three models. Table 4.4 portrays the category functioning of items using the delta parameters. The average measures are an important diagnostic tool and empirical indicator of how well the rating scale is functioning across categories (Linacre, 2002). The results pertaining to the average measures reveal that the rating scale category functioning is good for all items and comparable across Rasch models with the average score associated with each category increasing monotonically as one moves up the rating scale from 0 to 3. Students who have, on average, a greater understanding of the concepts underpinning the NOS construct respond, on average,
Table 4.1

Comparison of Outfit Mean Square Error (MNSQ) and Standardized T Statistics

<table>
<thead>
<tr>
<th>Item</th>
<th>Unidimensional</th>
<th></th>
<th>Consecutive</th>
<th></th>
<th>Multidimensional</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>MNSQ</td>
<td>T</td>
<td>MNSQ</td>
<td>T</td>
<td>MNSQ</td>
<td>T</td>
</tr>
<tr>
<td>EMP8D</td>
<td>0.93</td>
<td>-1.1</td>
<td>1.03</td>
<td>0.5</td>
<td>1.04</td>
<td>0.6</td>
</tr>
<tr>
<td>EMP8I</td>
<td>0.97</td>
<td>-0.4</td>
<td>0.97</td>
<td>-0.4</td>
<td>1.03</td>
<td>0.4</td>
</tr>
<tr>
<td>EMP9A</td>
<td>0.88</td>
<td>-1.8</td>
<td>0.97</td>
<td>-0.4</td>
<td>0.96</td>
<td>-0.6</td>
</tr>
<tr>
<td>EMP9I</td>
<td>0.99</td>
<td>-0.1</td>
<td>1.03</td>
<td>0.4</td>
<td>1.06</td>
<td>0.9</td>
</tr>
<tr>
<td>EMP9J</td>
<td>0.86</td>
<td>-2.2</td>
<td>0.95</td>
<td>-0.7</td>
<td>0.90</td>
<td>-1.6</td>
</tr>
<tr>
<td>EMP14B</td>
<td>0.90</td>
<td>-0.8</td>
<td>1.03</td>
<td>0.3</td>
<td>1.01</td>
<td>0.1</td>
</tr>
<tr>
<td>INV6D</td>
<td>1.03</td>
<td>0.5</td>
<td>0.99</td>
<td>-0.1</td>
<td>0.96</td>
<td>-0.6</td>
</tr>
<tr>
<td>INV8G</td>
<td>1.01</td>
<td>0.2</td>
<td>0.99</td>
<td>-0.1</td>
<td>0.95</td>
<td>-0.8</td>
</tr>
<tr>
<td>INV12L</td>
<td>1.00</td>
<td>0.0</td>
<td>1.06</td>
<td>0.7</td>
<td>1.04</td>
<td>0.6</td>
</tr>
<tr>
<td>INV12M</td>
<td>1.02</td>
<td>0.3</td>
<td>1.06</td>
<td>0.8</td>
<td>1.08</td>
<td>1.1</td>
</tr>
<tr>
<td>INV12N</td>
<td>0.95</td>
<td>-0.6</td>
<td>1.01</td>
<td>0.1</td>
<td>0.92</td>
<td>-1.0</td>
</tr>
<tr>
<td>THL6C</td>
<td>1.01</td>
<td>0.1</td>
<td>0.94</td>
<td>-0.8</td>
<td>0.96</td>
<td>-0.5</td>
</tr>
<tr>
<td>THL8F</td>
<td>1.02</td>
<td>0.3</td>
<td>1.05</td>
<td>0.7</td>
<td>0.99</td>
<td>-0.2</td>
</tr>
<tr>
<td>THL8K</td>
<td>1.03</td>
<td>0.4</td>
<td>1.07</td>
<td>0.9</td>
<td>1.11</td>
<td>1.4</td>
</tr>
<tr>
<td>THL9K</td>
<td>1.01</td>
<td>0.2</td>
<td>0.98</td>
<td>-0.2</td>
<td>1.00</td>
<td>-0.0</td>
</tr>
<tr>
<td>THL12K</td>
<td>0.79</td>
<td>-2.9</td>
<td>0.87</td>
<td>-1.7</td>
<td>0.81</td>
<td>-2.6</td>
</tr>
<tr>
<td>THL14E</td>
<td>1.13</td>
<td>1.0</td>
<td>1.05</td>
<td>0.4</td>
<td>1.22</td>
<td>1.7</td>
</tr>
<tr>
<td>CER6H</td>
<td>0.95</td>
<td>-0.7</td>
<td>0.95</td>
<td>-0.7</td>
<td>0.98</td>
<td>-0.3</td>
</tr>
<tr>
<td>CER6J</td>
<td>0.95</td>
<td>-0.6</td>
<td>0.95</td>
<td>-0.8</td>
<td>0.89</td>
<td>-1.7</td>
</tr>
<tr>
<td>CER7H</td>
<td>1.08</td>
<td>1.2</td>
<td>1.01</td>
<td>0.2</td>
<td>1.03</td>
<td>0.5</td>
</tr>
<tr>
<td>CER8L</td>
<td>0.99</td>
<td>-0.1</td>
<td>0.96</td>
<td>-0.4</td>
<td>1.02</td>
<td>0.2</td>
</tr>
<tr>
<td>CER8M</td>
<td>1.05</td>
<td>0.7</td>
<td>1.09</td>
<td>1.1</td>
<td>1.09</td>
<td>1.1</td>
</tr>
<tr>
<td>CER14A</td>
<td>1.11</td>
<td>0.9</td>
<td>1.09</td>
<td>0.8</td>
<td>1.15</td>
<td>1.2</td>
</tr>
<tr>
<td>SCE13C</td>
<td>1.11</td>
<td>1.6</td>
<td>1.02</td>
<td>0.3</td>
<td>1.00</td>
<td>-0.0</td>
</tr>
<tr>
<td>SCE13F</td>
<td>0.95</td>
<td>-0.7</td>
<td>0.99</td>
<td>-0.1</td>
<td>0.91</td>
<td>-1.4</td>
</tr>
<tr>
<td>SCE13H</td>
<td>1.01</td>
<td>0.2</td>
<td>0.96</td>
<td>-0.5</td>
<td>0.96</td>
<td>-0.5</td>
</tr>
<tr>
<td>SCE13I</td>
<td>1.06</td>
<td>0.9</td>
<td>1.03</td>
<td>0.4</td>
<td>1.02</td>
<td>0.3</td>
</tr>
<tr>
<td>SCE13K</td>
<td>1.17</td>
<td>2.0</td>
<td>1.12</td>
<td>1.4</td>
<td>1.08</td>
<td>1.0</td>
</tr>
</tbody>
</table>
Table 4.2

Comparison of Infit Mean Square Error (MNSQ) and Standardized T Statistics (T)

<table>
<thead>
<tr>
<th>Item</th>
<th>Unidimensional</th>
<th>Consecutive</th>
<th>Multidimensional</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>MNSQ</td>
<td>T</td>
<td>MNSQ</td>
</tr>
<tr>
<td>EMP8D</td>
<td>0.95</td>
<td>-0.7</td>
<td>1.04</td>
</tr>
<tr>
<td>EMP8I</td>
<td>1.00</td>
<td>0.0</td>
<td>1.00</td>
</tr>
<tr>
<td>EMP9A</td>
<td>0.94</td>
<td>-0.7</td>
<td>1.01</td>
</tr>
<tr>
<td>EMP9I</td>
<td>1.00</td>
<td>0.1</td>
<td>1.05</td>
</tr>
<tr>
<td>EMP9J</td>
<td>0.89</td>
<td>-1.6</td>
<td>0.95</td>
</tr>
<tr>
<td>EMP14B</td>
<td>0.93</td>
<td>-0.5</td>
<td>1.04</td>
</tr>
<tr>
<td>INV6D</td>
<td>1.04</td>
<td>0.7</td>
<td>1.00</td>
</tr>
<tr>
<td>INV8G</td>
<td>1.01</td>
<td>0.2</td>
<td>0.98</td>
</tr>
<tr>
<td>INV12L</td>
<td>0.99</td>
<td>-0.2</td>
<td>1.03</td>
</tr>
<tr>
<td>INV12M</td>
<td>1.05</td>
<td>0.6</td>
<td>1.06</td>
</tr>
<tr>
<td>INV12N</td>
<td>0.95</td>
<td>-0.7</td>
<td>1.02</td>
</tr>
<tr>
<td>THL6C</td>
<td>1.02</td>
<td>0.2</td>
<td>0.95</td>
</tr>
<tr>
<td>THL8F</td>
<td>1.02</td>
<td>0.3</td>
<td>1.04</td>
</tr>
<tr>
<td>THL8K</td>
<td>0.98</td>
<td>-0.2</td>
<td>1.03</td>
</tr>
<tr>
<td>THL9K</td>
<td>1.02</td>
<td>0.3</td>
<td>0.98</td>
</tr>
<tr>
<td>THL12K</td>
<td>0.87</td>
<td>-1.4</td>
<td>0.90</td>
</tr>
<tr>
<td>THL14E</td>
<td>1.12</td>
<td>1.1</td>
<td>1.07</td>
</tr>
<tr>
<td>CER6H</td>
<td>0.98</td>
<td>-0.3</td>
<td>0.95</td>
</tr>
<tr>
<td>CER6J</td>
<td>0.95</td>
<td>-0.7</td>
<td>0.96</td>
</tr>
<tr>
<td>CER7H</td>
<td>1.02</td>
<td>0.2</td>
<td>1.00</td>
</tr>
<tr>
<td>CER8L</td>
<td>0.94</td>
<td>-0.6</td>
<td>0.96</td>
</tr>
<tr>
<td>CER8M</td>
<td>1.06</td>
<td>0.9</td>
<td>1.10</td>
</tr>
<tr>
<td>CER14A</td>
<td>1.11</td>
<td>1.0</td>
<td>1.09</td>
</tr>
<tr>
<td>SCE13C</td>
<td>1.09</td>
<td>1.5</td>
<td>1.01</td>
</tr>
<tr>
<td>SCE13F</td>
<td>0.95</td>
<td>-0.8</td>
<td>1.00</td>
</tr>
<tr>
<td>SCE13H</td>
<td>1.02</td>
<td>0.4</td>
<td>0.96</td>
</tr>
<tr>
<td>SCE13I</td>
<td>1.06</td>
<td>1.0</td>
<td>1.02</td>
</tr>
<tr>
<td>SCE13K</td>
<td>1.16</td>
<td>2.3</td>
<td>1.10</td>
</tr>
</tbody>
</table>
within the higher response categories. For this sample, the meaning and interpretability of the rating scale is consistent across all items and Rasch models and the rating scale is being used as intended by the scale developers.

The Rasch model is a stochastic model with responses modeled by the partial credit function (Equation 3: Unidimensional; Equations 4 and 5: Multidimensional). This function models students’ abilities in the probabilistic relationship to the difficulty of each item; the expectation is that the delta parameters (intersection of adjacent probability curves) should increase monotonically (Linacre, 2002; Boman, Curtis, Furlong, & Smith, 2006). The results, shown in Table 4.4, indicate that there are three items whose delta parameters are disordinal. These items, EMP9a (Scientists explain how something works), THL8K (Scientists use what they found in the past to help explain their new findings) and CER7H (Two scientists can disagree, but both can have good ideas) were disordinal across all Rasch models. Adams (personal communication, September 10, 2012, p. 5) indicates that when the delta parameters are disordinal, the “continuum can no longer be divided into regions in which each of the categories takes it turn as being the single most probable response.” Regardless of the Rasch model, for these three items, a higher level of the latent trait is not consistently associated with higher rating scale categories. Two factors can contribute to this, (1) the log-ratio of the frequency of neighboring categories and (2) the average measures of the respondents opting for each category (Linacre, 2002). The results in Table 4.3 above have shown that, as expected, the average observed measures increase monotonically. Therefore, the likely cause of disordinality is related to the frequency of use of item categories. Students do not appear
to use category 1 (disagree a little) as expected. Relative to the lower category (zero), there is an unexpected lower frequency of observed responses in this category resulting in the disordinal nature of these items.

The data presented in this section are consistent across Rasch models. The same three items with disordinal delta thresholds exhibit this behavior in all three models. Given that the sample was a convenience sample, it is premature to make definitive decisions about the suitability of keeping these items in the scale. For the ER sample, the average of the observed responses for these three items were monotonically increasing as one moved up the rating scale indicating that the rating scale, at least for this sample, was functioning appropriately. Overall, the data indicate that the rating scale functioning is equivalent across Rasch models and provides evidence for the substantive validity aspect of the NOS construct. The empirical evidence provided in this section still support the use of any of the three internal representations of the NOS construct.
### Table 4.3

**Rating Scale Functioning: Monotonicity of Average Measures**

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>EMP</td>
<td>0</td>
<td>0.10</td>
<td>-0.12</td>
<td>0.15</td>
<td>THL</td>
<td>0</td>
<td>-0.03</td>
<td>-0.19</td>
<td>0.47</td>
<td>CER</td>
<td>0</td>
<td>0.08</td>
<td>0.16</td>
<td>0.41</td>
</tr>
<tr>
<td>8D</td>
<td>1</td>
<td>0.30</td>
<td>0.44</td>
<td>0.50</td>
<td>6C</td>
<td>1</td>
<td>0.42</td>
<td>0.54</td>
<td>0.54</td>
<td>6H</td>
<td>1</td>
<td>0.36</td>
<td>0.50</td>
<td>0.46</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>0.79</td>
<td>1.13</td>
<td>1.12</td>
<td></td>
<td>2</td>
<td>0.62</td>
<td>0.94</td>
<td>0.96</td>
<td></td>
<td>2</td>
<td>0.81</td>
<td>1.06</td>
<td>1.04</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>1.14</td>
<td>1.57</td>
<td>1.62</td>
<td></td>
<td>3</td>
<td>1.10</td>
<td>1.58</td>
<td>1.66</td>
<td></td>
<td>3</td>
<td>1.16</td>
<td>1.47</td>
<td>1.58</td>
</tr>
<tr>
<td>EMP</td>
<td>0</td>
<td>0.00</td>
<td>-0.11</td>
<td>-0.06</td>
<td>THL</td>
<td>0</td>
<td>0.36</td>
<td>0.39</td>
<td>0.58</td>
<td>CER</td>
<td>0</td>
<td>-0.30</td>
<td>-0.07</td>
<td>-0.20</td>
</tr>
<tr>
<td>8I</td>
<td>1</td>
<td>0.40</td>
<td>0.45</td>
<td>0.52</td>
<td>8F</td>
<td>1</td>
<td>0.46</td>
<td>0.76</td>
<td>0.64</td>
<td>6J</td>
<td>1</td>
<td>0.31</td>
<td>0.49</td>
<td>0.49</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>0.85</td>
<td>1.06</td>
<td>1.11</td>
<td></td>
<td>2</td>
<td>0.76</td>
<td>1.15</td>
<td>1.16</td>
<td></td>
<td>2</td>
<td>0.73</td>
<td>0.88</td>
<td>0.95</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>1.11</td>
<td>1.66</td>
<td>1.68</td>
<td></td>
<td>3</td>
<td>1.19</td>
<td>1.66</td>
<td>1.79</td>
<td></td>
<td>3</td>
<td>1.08</td>
<td>1.41</td>
<td>1.45</td>
</tr>
<tr>
<td>EMP</td>
<td>0</td>
<td>-0.06</td>
<td>-0.20</td>
<td>-0.06</td>
<td>THL</td>
<td>0</td>
<td>0.10</td>
<td>0.12</td>
<td>0.27</td>
<td>CER</td>
<td>0</td>
<td>-0.26</td>
<td>0.01</td>
<td>-0.15</td>
</tr>
<tr>
<td>9A</td>
<td>1</td>
<td>0.31</td>
<td>0.45</td>
<td>0.44</td>
<td>8K</td>
<td>1</td>
<td>0.37</td>
<td>0.54</td>
<td>0.64</td>
<td>7H</td>
<td>1</td>
<td>0.55</td>
<td>0.45</td>
<td>0.64</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>0.68</td>
<td>0.82</td>
<td>0.91</td>
<td></td>
<td>2</td>
<td>0.73</td>
<td>1.11</td>
<td>1.15</td>
<td></td>
<td>2</td>
<td>0.73</td>
<td>0.88</td>
<td>0.88</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>1.13</td>
<td>1.62</td>
<td>1.64</td>
<td></td>
<td>3</td>
<td>1.19</td>
<td>1.66</td>
<td>1.79</td>
<td></td>
<td>3</td>
<td>1.05</td>
<td>1.41</td>
<td>1.46</td>
</tr>
<tr>
<td>EMP</td>
<td>0</td>
<td>0.17</td>
<td>-0.01</td>
<td>0.20</td>
<td>THL</td>
<td>0</td>
<td>-0.06</td>
<td>-0.06</td>
<td>0.24</td>
<td>CER</td>
<td>0</td>
<td>-0.20</td>
<td>-0.02</td>
<td>0.00</td>
</tr>
<tr>
<td>9I</td>
<td>1</td>
<td>0.53</td>
<td>0.55</td>
<td>0.65</td>
<td>9K</td>
<td>1</td>
<td>0.45</td>
<td>0.88</td>
<td>0.78</td>
<td>8L</td>
<td>1</td>
<td>0.43</td>
<td>0.62</td>
<td>0.45</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>0.81</td>
<td>1.13</td>
<td>1.16</td>
<td></td>
<td>2</td>
<td>0.78</td>
<td>1.08</td>
<td>1.17</td>
<td></td>
<td>2</td>
<td>0.62</td>
<td>0.76</td>
<td>0.83</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>1.21</td>
<td>1.79</td>
<td>1.85</td>
<td></td>
<td>3</td>
<td>1.17</td>
<td>1.77</td>
<td>1.74</td>
<td></td>
<td>3</td>
<td>1.11</td>
<td>1.39</td>
<td>1.46</td>
</tr>
<tr>
<td>EMP</td>
<td>0</td>
<td>-0.21</td>
<td>-0.17</td>
<td>-0.30</td>
<td>THL</td>
<td>0</td>
<td>-0.18</td>
<td>-0.02</td>
<td>-0.20</td>
<td>CER</td>
<td>0</td>
<td>0.17</td>
<td>0.45</td>
<td>0.44</td>
</tr>
<tr>
<td>9J</td>
<td>1</td>
<td>0.33</td>
<td>0.34</td>
<td>0.37</td>
<td>12K</td>
<td>1</td>
<td>0.09</td>
<td>0.35</td>
<td>0.33</td>
<td>8M</td>
<td>1</td>
<td>0.45</td>
<td>0.60</td>
<td>0.62</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>0.72</td>
<td>0.98</td>
<td>0.98</td>
<td></td>
<td>2</td>
<td>0.60</td>
<td>1.04</td>
<td>0.99</td>
<td></td>
<td>2</td>
<td>0.93</td>
<td>1.22</td>
<td>1.19</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>1.18</td>
<td>1.65</td>
<td>1.74</td>
<td></td>
<td>3</td>
<td>1.09</td>
<td>1.54</td>
<td>1.61</td>
<td></td>
<td>3</td>
<td>1.17</td>
<td>1.40</td>
<td>1.51</td>
</tr>
<tr>
<td>EMP</td>
<td>0</td>
<td>0.14</td>
<td>0.11</td>
<td>0.17</td>
<td>THL</td>
<td>0</td>
<td>0.30</td>
<td>0.16</td>
<td>0.32</td>
<td>CER</td>
<td>0</td>
<td>0.24</td>
<td>0.41</td>
<td>0.32</td>
</tr>
<tr>
<td>14B</td>
<td>1</td>
<td>0.29</td>
<td>0.20</td>
<td>0.40</td>
<td>14E</td>
<td>1</td>
<td>0.44</td>
<td>0.71</td>
<td>0.71</td>
<td>14A</td>
<td>1</td>
<td>0.60</td>
<td>0.75</td>
<td>0.75</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>0.76</td>
<td>1.12</td>
<td>1.06</td>
<td></td>
<td>2</td>
<td>0.88</td>
<td>1.27</td>
<td>1.32</td>
<td></td>
<td>2</td>
<td>0.75</td>
<td>1.21</td>
<td>1.01</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>1.21</td>
<td>1.64</td>
<td>1.71</td>
<td></td>
<td>3</td>
<td>1.08</td>
<td>1.63</td>
<td>1.70</td>
<td></td>
<td>3</td>
<td>1.20</td>
<td>1.42</td>
<td>1.57</td>
</tr>
</tbody>
</table>
Table 4.3

*Rating Scale Functioning: Monotonicity of Average Measures (continued)*

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>INV</td>
<td>0</td>
<td>0.36</td>
<td>-0.28</td>
<td>-0.29</td>
<td>SCE</td>
<td>0</td>
<td>0.49</td>
<td>-0.26</td>
<td>-0.21</td>
</tr>
<tr>
<td>6D</td>
<td>1</td>
<td>0.56</td>
<td>0.30</td>
<td>0.34</td>
<td>13C</td>
<td>1</td>
<td>0.54</td>
<td>0.11</td>
<td>0.11</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>0.91</td>
<td>1.00</td>
<td>0.82</td>
<td>2</td>
<td>0.88</td>
<td>0.47</td>
<td>0.48</td>
<td></td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>1.23</td>
<td>1.53</td>
<td>1.50</td>
<td>3</td>
<td>1.36</td>
<td>1.04</td>
<td>1.00</td>
<td></td>
</tr>
<tr>
<td>INV</td>
<td>0</td>
<td>0.28</td>
<td>-0.28</td>
<td>-0.23</td>
<td>SCE</td>
<td>0</td>
<td>0.32</td>
<td>-0.55</td>
<td>-0.37</td>
</tr>
<tr>
<td>8G</td>
<td>1</td>
<td>0.55</td>
<td>0.16</td>
<td>0.19</td>
<td>13F</td>
<td>1</td>
<td>0.34</td>
<td>0.02</td>
<td>0.03</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>0.80</td>
<td>0.79</td>
<td>0.71</td>
<td>2</td>
<td>0.83</td>
<td>0.42</td>
<td>0.39</td>
<td></td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>1.30</td>
<td>1.66</td>
<td>1.48</td>
<td>3</td>
<td>1.30</td>
<td>0.88</td>
<td>0.89</td>
<td></td>
</tr>
<tr>
<td>INV</td>
<td>0</td>
<td>0.37</td>
<td>-0.42</td>
<td>-0.27</td>
<td>SCE</td>
<td>0</td>
<td>0.25</td>
<td>-0.44</td>
<td>-0.35</td>
</tr>
<tr>
<td>12L</td>
<td>1</td>
<td>0.46</td>
<td>0.16</td>
<td>0.16</td>
<td>13H</td>
<td>1</td>
<td>0.53</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>0.77</td>
<td>0.64</td>
<td>0.62</td>
<td>2</td>
<td>0.88</td>
<td>0.47</td>
<td>0.46</td>
<td></td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>1.24</td>
<td>1.50</td>
<td>1.39</td>
<td>3</td>
<td>1.27</td>
<td>0.93</td>
<td>0.97</td>
<td></td>
</tr>
<tr>
<td>INV</td>
<td>0</td>
<td>0.01</td>
<td>-0.48</td>
<td>-0.66</td>
<td>SCE</td>
<td>0</td>
<td>0.37</td>
<td>-0.21</td>
<td>-0.18</td>
</tr>
<tr>
<td>12M</td>
<td>1</td>
<td>0.47</td>
<td>0.03</td>
<td>0.15</td>
<td>13I</td>
<td>1</td>
<td>0.62</td>
<td>0.17</td>
<td>0.16</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>0.87</td>
<td>0.77</td>
<td>0.77</td>
<td>2</td>
<td>0.91</td>
<td>0.54</td>
<td>0.49</td>
<td></td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>1.12</td>
<td>1.30</td>
<td>1.18</td>
<td>3</td>
<td>1.36</td>
<td>0.93</td>
<td>0.99</td>
<td></td>
</tr>
<tr>
<td>INV</td>
<td>0</td>
<td>0.30</td>
<td>-0.45</td>
<td>-0.36</td>
<td>SCE</td>
<td>0</td>
<td>0.46</td>
<td>-0.13</td>
<td>-0.10</td>
</tr>
<tr>
<td>12N</td>
<td>1</td>
<td>0.37</td>
<td>0.09</td>
<td>0.01</td>
<td>13K</td>
<td>1</td>
<td>0.60</td>
<td>0.22</td>
<td>0.26</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>0.81</td>
<td>0.82</td>
<td>0.81</td>
<td>2</td>
<td>0.86</td>
<td>0.47</td>
<td>0.46</td>
<td></td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>1.33</td>
<td>1.50</td>
<td>1.39</td>
<td>3</td>
<td>1.24</td>
<td>0.92</td>
<td>0.95</td>
<td></td>
</tr>
</tbody>
</table>

*Legend:* Cat. – Response Category; Uni. – Unidimensional Model; Con. – Consecutive Model; Mlt. – Multidimensional Model
Table 4.4

*Rating Scale Functioning: Monotonicity of Delta Parameters*

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>EMP</td>
<td>0</td>
<td>NA</td>
<td>NA</td>
<td>THL</td>
<td>0</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
<td>CER</td>
<td>0</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
</tr>
<tr>
<td>8D</td>
<td>0/1</td>
<td>-0.68</td>
<td>-0.77</td>
<td>-0.78</td>
<td>6C</td>
<td>0/1</td>
<td>-1.44</td>
<td>-1.41</td>
<td>-1.47</td>
<td>6H</td>
<td>0/1</td>
<td>-1.59</td>
<td>-1.55</td>
<td>-1.53</td>
</tr>
<tr>
<td></td>
<td>1/2</td>
<td>-0.16</td>
<td>-0.05</td>
<td>-0.04</td>
<td>1/2</td>
<td>-0.90</td>
<td>-0.69</td>
<td>-0.63</td>
<td>1/2</td>
<td>0.18</td>
<td>0.17</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>2/3</td>
<td>0.42</td>
<td>0.78</td>
<td>0.80</td>
<td>2/3</td>
<td>0.38</td>
<td>0.78</td>
<td>0.79</td>
<td>2/3</td>
<td>1.03</td>
<td>1.02</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>EMP</td>
<td>0</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
<td>THL</td>
<td>0</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
<td>CER</td>
<td>0</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
</tr>
<tr>
<td>8I</td>
<td>0/1</td>
<td>-0.66</td>
<td>-0.75</td>
<td>-0.77</td>
<td>8F</td>
<td>0/1</td>
<td>-1.05</td>
<td>-0.95</td>
<td>-0.92</td>
<td>6J</td>
<td>0/1</td>
<td>-1.11</td>
<td>-1.10</td>
<td>-1.07</td>
</tr>
<tr>
<td></td>
<td>1/2</td>
<td>-0.55</td>
<td>-0.44</td>
<td>-0.42</td>
<td>1/2</td>
<td>-0.22</td>
<td>0.06</td>
<td>0.10</td>
<td>1/2</td>
<td>-0.60</td>
<td>-0.45</td>
<td>-0.45</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>2/3</td>
<td>0.82</td>
<td>1.19</td>
<td>1.22</td>
<td>2/3</td>
<td>0.96</td>
<td>1.40</td>
<td>1.40</td>
<td>2/3</td>
<td>0.26</td>
<td>0.56</td>
<td>0.54</td>
<td></td>
<td></td>
</tr>
<tr>
<td>EMP</td>
<td>0</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
<td>THL</td>
<td>0</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
<td>CER</td>
<td>0</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
</tr>
<tr>
<td>9A</td>
<td>0/1</td>
<td>-0.33</td>
<td>-0.46</td>
<td>-0.48</td>
<td>8K</td>
<td>0/1</td>
<td>-0.46</td>
<td>-0.39</td>
<td>-0.40</td>
<td>7H</td>
<td>0/1</td>
<td>-0.64</td>
<td>-0.63</td>
<td>-0.61</td>
</tr>
<tr>
<td></td>
<td>1/2</td>
<td>-0.81</td>
<td>-0.74</td>
<td>-0.73</td>
<td>1/2</td>
<td>-1.06</td>
<td>-0.79</td>
<td>-0.76</td>
<td>1/2</td>
<td>-0.80</td>
<td>-0.65</td>
<td>-0.65</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>2/3</td>
<td>0.35</td>
<td>0.68</td>
<td>0.70</td>
<td>2/3</td>
<td>0.84</td>
<td>1.29</td>
<td>1.27</td>
<td>2/3</td>
<td>0.29</td>
<td>0.58</td>
<td>0.57</td>
<td></td>
<td></td>
</tr>
<tr>
<td>EMP</td>
<td>0</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
<td>THL</td>
<td>0</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
<td>CER</td>
<td>0</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
</tr>
<tr>
<td>9I</td>
<td>0/1</td>
<td>-1.07</td>
<td>-1.13</td>
<td>-1.15</td>
<td>9K</td>
<td>0/1</td>
<td>-1.12</td>
<td>-1.03</td>
<td>-1.04</td>
<td>8L</td>
<td>0/1</td>
<td>-0.97</td>
<td>-1.00</td>
<td>-0.97</td>
</tr>
<tr>
<td></td>
<td>1/2</td>
<td>0.00</td>
<td>0.15</td>
<td>0.18</td>
<td>1/2</td>
<td>-0.34</td>
<td>-0.01</td>
<td>-0.02</td>
<td>1/2</td>
<td>-0.74</td>
<td>-0.61</td>
<td>-0.61</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>2/3</td>
<td>1.16</td>
<td>1.57</td>
<td>1.61</td>
<td>2/3</td>
<td>1.03</td>
<td>1.51</td>
<td>1.48</td>
<td>2/3</td>
<td>0.36</td>
<td>0.63</td>
<td>0.64</td>
<td></td>
<td></td>
</tr>
<tr>
<td>EMP</td>
<td>0</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
<td>THL</td>
<td>0</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
<td>CER</td>
<td>0</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
</tr>
<tr>
<td>9J</td>
<td>0/1</td>
<td>-1.09</td>
<td>-1.21</td>
<td>-1.25</td>
<td>12K</td>
<td>0/1</td>
<td>-1.07</td>
<td>-1.05</td>
<td>-1.12</td>
<td>8M</td>
<td>0/1</td>
<td>-0.69</td>
<td>-0.63</td>
<td>-0.61</td>
</tr>
<tr>
<td></td>
<td>1/2</td>
<td>-0.62</td>
<td>-0.53</td>
<td>-0.52</td>
<td>1/2</td>
<td>-0.85</td>
<td>-0.63</td>
<td>-0.60</td>
<td>1/2</td>
<td>0.31</td>
<td>0.53</td>
<td>0.51</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>2/3</td>
<td>0.72</td>
<td>1.07</td>
<td>1.09</td>
<td>2/3</td>
<td>0.12</td>
<td>0.53</td>
<td>0.52</td>
<td>2/3</td>
<td>1.02</td>
<td>1.35</td>
<td>1.34</td>
<td></td>
<td></td>
</tr>
<tr>
<td>EMP</td>
<td>0</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
<td>THL</td>
<td>0</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
<td>CER</td>
<td>0</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
</tr>
<tr>
<td>14B</td>
<td>0/1</td>
<td>-0.81</td>
<td>-0.88</td>
<td>-0.94</td>
<td>14E</td>
<td>0/1</td>
<td>-0.60</td>
<td>-0.47</td>
<td>-0.48</td>
<td>14A</td>
<td>0/1</td>
<td>-0.40</td>
<td>-0.28</td>
<td>-0.30</td>
</tr>
<tr>
<td></td>
<td>1/2</td>
<td>-0.35</td>
<td>-0.18</td>
<td>-0.24</td>
<td>1/2</td>
<td>0.22</td>
<td>0.53</td>
<td>0.55</td>
<td>1/2</td>
<td>0.44</td>
<td>0.67</td>
<td>0.64</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>2/3</td>
<td>1.27</td>
<td>1.70</td>
<td>1.71</td>
<td>2/3</td>
<td>0.89</td>
<td>1.33</td>
<td>1.33</td>
<td>2/3</td>
<td>1.04</td>
<td>1.35</td>
<td>1.37</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Table 4.4

*Rating Scale Functioning: Monotonicity of Delta Parameters (continued)*

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>INV</td>
<td>0</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
<td>SCE</td>
<td>0</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
</tr>
<tr>
<td>6D</td>
<td>0/1</td>
<td>-0.19</td>
<td>-0.66</td>
<td>-0.63</td>
<td>13C</td>
<td>0/1</td>
<td>-0.41</td>
<td>-0.90</td>
<td>-0.90</td>
</tr>
<tr>
<td></td>
<td>1/2</td>
<td>0.43</td>
<td>0.24</td>
<td>0.23</td>
<td></td>
<td>1/2</td>
<td>0.29</td>
<td>-0.16</td>
<td>-0.16</td>
</tr>
<tr>
<td></td>
<td>2/3</td>
<td>1.07</td>
<td>1.17</td>
<td>1.13</td>
<td></td>
<td>2/3</td>
<td>1.67</td>
<td>1.26</td>
<td>1.28</td>
</tr>
<tr>
<td>INV</td>
<td>0</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
<td>SCE</td>
<td>0</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
</tr>
<tr>
<td>8G</td>
<td>0/1</td>
<td>-0.65</td>
<td>-1.19</td>
<td>-1.14</td>
<td>13F</td>
<td>0/1</td>
<td>-0.55</td>
<td>-1.09</td>
<td>-1.06</td>
</tr>
<tr>
<td></td>
<td>1/2</td>
<td>0.01</td>
<td>-0.24</td>
<td>-0.24</td>
<td></td>
<td>1/2</td>
<td>-0.06</td>
<td>-0.53</td>
<td>-0.52</td>
</tr>
<tr>
<td></td>
<td>2/3</td>
<td>1.34</td>
<td>1.42</td>
<td>1.38</td>
<td></td>
<td>2/3</td>
<td>1.16</td>
<td>0.74</td>
<td>0.75</td>
</tr>
<tr>
<td>INV</td>
<td>0</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
<td>SCE</td>
<td>0</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
</tr>
<tr>
<td>12L</td>
<td>0/1</td>
<td>-0.51</td>
<td>-1.04</td>
<td>-0.99</td>
<td>13H</td>
<td>0/1</td>
<td>-0.58</td>
<td>-1.11</td>
<td>-1.08</td>
</tr>
<tr>
<td></td>
<td>1/2</td>
<td>-0.08</td>
<td>-0.35</td>
<td>-0.35</td>
<td></td>
<td>1/2</td>
<td>-0.20</td>
<td>-0.67</td>
<td>-0.66</td>
</tr>
<tr>
<td></td>
<td>2/3</td>
<td>1.08</td>
<td>1.11</td>
<td>1.09</td>
<td></td>
<td>2/3</td>
<td>1.69</td>
<td>1.27</td>
<td>1.29</td>
</tr>
<tr>
<td>INV</td>
<td>0</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
<td>SCE</td>
<td>0</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
</tr>
<tr>
<td>12M</td>
<td>0/1</td>
<td>-0.55</td>
<td>-1.11</td>
<td>-1.05</td>
<td>13I</td>
<td>0/1</td>
<td>-0.13</td>
<td>-0.66</td>
<td>-0.63</td>
</tr>
<tr>
<td></td>
<td>1/2</td>
<td>-0.43</td>
<td>-0.73</td>
<td>-0.73</td>
<td></td>
<td>1/2</td>
<td>0.35</td>
<td>-0.08</td>
<td>-0.10</td>
</tr>
<tr>
<td></td>
<td>2/3</td>
<td>1.12</td>
<td>1.15</td>
<td>1.11</td>
<td></td>
<td>2/3</td>
<td>1.67</td>
<td>1.27</td>
<td>1.28</td>
</tr>
<tr>
<td>INV</td>
<td>0</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
<td>SCE</td>
<td>0</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
</tr>
<tr>
<td>12N</td>
<td>0/1</td>
<td>-0.25</td>
<td>-0.73</td>
<td>-0.71</td>
<td>13K</td>
<td>0/1</td>
<td>0.07</td>
<td>-0.30</td>
<td>-0.40</td>
</tr>
<tr>
<td></td>
<td>1/2</td>
<td>-0.11</td>
<td>-0.35</td>
<td>-0.36</td>
<td></td>
<td>1/2</td>
<td>0.42</td>
<td>0.01</td>
<td>-0.02</td>
</tr>
<tr>
<td></td>
<td>2/3</td>
<td>1.24</td>
<td>1.30</td>
<td>1.27</td>
<td></td>
<td>2/3</td>
<td>1.31</td>
<td>0.94</td>
<td>0.92</td>
</tr>
</tbody>
</table>

**Legend:** Cat. – Response Category; Uni. – Unidimensional Model; Con. – Consecutive Model; Mlt. – Multidimensional Model

1. Disordinal item delta parameters are in bolded font.
**Expected Item Difficulty Hierarchy.** Evidence relating to the substantive aspect of construct validity was provided by examining the ordering of average population parameters across the five NOS dimensions. If the order of dimension means conform to those expected by the theoretical framework built for the NOS construct, support is provided for this aspect of construct validity.

**Subscale Hierarchy.**

When using the unidimensional model, only one population parameter estimate is reported; this estimate provides the population mean for the 28 item NOSI-E scale. It is therefore not possible to estimate the means of the five dimensions using this model. Table 4.5 represents the means and standard deviations (S.D.) for the five NOS dimensions using the consecutive and multidimensional approaches; the mean and standard deviation for the unidimensional approach is provided for context. The means are ordered from most difficult to least difficult in the table.

In relative order of difficulty, students find concepts measured by the Inventive (INV) and Socially and Culturally Embedded (SCE) dimensions, on average, more difficult to understand than concepts measured by the Certainty (CER), Theory-laden (THL) and Empirical (EMP) dimensions. The mean of the ability estimates ranged from +0.42 (SCE) to +1.26 (THL) logits; these two dimensions, on average, anchor the difficult and easier ends of the latent NOS continuum, respectively. The estimates for the consecutive and multidimensional approaches are almost identical; all bivariate correlations between the estimates of the two models were 0.99 or above (data not shown).
This ordering of average dimension difficulties parallels the findings of several qualitative studies (Smith, Maclin, Houghton, & Hennessey, 2000; Moss, Abrams, & Robb, 2001; Bell, Blair, Crawford, & Lederman, 2003; Akerson & Abd-El-Khalick, 2005; Akerson & Donnelly, 2010; Akerson, Buck, Donnelly, Nargund-Joshi, & Weiland, 2011) and of quantitative studies (Tsai & Liu, 2005; Huang, Tsai, & Chang, 2005; Kang, Scharmann, & Noh, 2005). In these studies, the INV and SCE concepts were, on average, hardest for students to understand with the tentative (CER), theory-laden and empirical nature of science conceptually easier for students to understand. These results suggest that the ordering of average dimension difficulties conform to theoretical expectations as all of these studies were premised on Lederman’s theoretical framework for the NOS construct.
To provide additional evidence that the item hierarchy agrees with theoretical expectations for the NOS construct across the three models, an assessment of the ordering of item difficulties (average delta) within each dimension was undertaken; this was supported by a qualitative assessment of the content embedded in the items. Table 4.6 compares the deltas for the 28 items across the five dimensions for the three models. Dimensions in the table are in order of most difficult (SCE) to least difficult (THL); for the most part, items within each dimension are similarly in order of most difficult (positive logits) to least difficult (negative logits). The unidimensional model item difficulty means were used to order the items; items from the consecutive and multidimensional models that do not conform to same pattern as the unidimensional model are flagged.

With few exceptions, the ordered pattern of item difficulties within each dimension is similar across the three Rasch models (Table 4.6). In the SCE dimension, item SCE13K (Where scientists live may affect what they are allowed to work on) was the most difficult item for students when the consecutive model was used to analyze the data; in the unidimensional and multidimensional, it was the second most difficult. In the Empirical dimension, item EMP14B (Scientists infer what they think is happening from what they already know), was the most difficult item for students under the unidimensional and consecutive approaches but ranked second most difficult in the multidimensional approaches. Overall, however, there was no substantial difference in the ordering of average deltas across the three models.
Qualitative analyses of the content of these dimensional items were conducted to establish that the difficulty hierarchy of items met our *a priori* assumptions and conform to Lederman’s theoretical framework (Lederman, 1992; Lederman, Abd-El-Khalick, Bell, & Schwartz, 2002; Lederman, 2007) for the NOS construct. The item prompts are reported in Table 4.6. There are few item-level (none that are Rasch-based) quantitative studies of elementary students’ NOS views which can provide supporting evidence for the item hierarchy at the subscale level evident in this study. Most supporting evidence is provided by Smith, Maclin, Houghton and Hennessey’s (2000) seminal research on sixth grade students’ epistemologies of science. Where possible, other supporting evidence is provided but the discussion is largely driven by theoretical considerations.

Students, on average, found it harder to understand item SCE13I (How scientists see the world is influenced by the culture they grew up in) and item SCE13K (Where scientists live may affect what they are allowed to work on) than item SCE13F (A scientists beliefs may change how they do their work). The most difficult SCE items are related to the influence of culture on scientists with the least difficult item of this dimension pertaining to scientists’ individual beliefs and how these beliefs may change their perspectives. Novice respondents in Lederman, Abd-El-Khalick, Bell and Schwartz’s study (2002), as a group, had no perception of the importance of the role that culture plays on scientific knowledge construction; this finding provides only partial support that the items within the SCE domain are ordered appropriately and theoretically grounded.
For the Inventive domains students found item INV6D (You have to be creative to work in science) and item INV12N (A good imagination is needed to create the best experiment to test an idea) more difficult to understand than item INV12M (A good imagination is needed to make predictions about what will happen in an experiment). This ordering of items seems intuitive. In the classroom, students are often asked to predict what will happen in an experiment and through this experience, likely realize the role imagination plays in this process. Given the criticism that students within K-12 classrooms are often spoon-fed formulaic experiments to perform (Carey & Smith, 1993; Driver, Asoko, Leach, Mortimer, & Scott, 1994; Smith, Maclin, Houghton, & Hennessey, 2000; Kawasaki, Herrenkohl, & Leary; 2004; Duschl, 2008) that lead to “foregone conclusions” (Apedoe & Ford, 2010, p. 168), it is predictable that students may not understand that a good imagination is also needed to create the best experiment to test out scientists’ ideas and the role of creativity in science. Therefore, the ordering of items within the Inventive domain appears theoretically grounded.

Students found it easier, on average, to understand that “trying things out helps scientists think of new ideas” (CER6J) but, on average, they had more trouble endorsing the intentionally more difficult concepts in item CER8M (New theories in science should only be accepted when there is a lot of evidence to support them) and in item CER14A (In science, finding out that a hypothesis is NOT correct is as important as finding out that a hypothesis IS correct). A higher percentage of students taught within Smith, Maclin, Houghton and Hennessey’s (2000, p. 374) constructivist classroom were able to understand that scientific change involves the development of ideas (61%) but fewer
were aware that change processes in science involves “complex evidence” (39%) and the search for “better explanations” (33%). Students within the comparison classroom (teacher-led pedagogy) exhibited similar patterns for these concepts but at a much lower level of frequency and predominantly endorsed that scientist would likely keep or abandon an idea based on a single experiment (59%). These data provide support for the hierarchy of item difficulty evident in the Inventive dimension found in this study and suggests that the Inventive domain is theoretically grounded.

The role of inference in science was hard for students to understand with item EMP14B (Scientist infer what they think is happening from what they already know) being the hardest item within the Empirical dimension to understand. In contrast, students found it relatively easier to understand that scientists use explanation to describe how something works (EMP9A) and that science helps answer questions about how something works (EMP9J). Supporting evidence that this sequence of item difficulties within the Empirical domain is theoretically grounded is also provided by Smith, Maclin, Houghton and Hennessey’s seminal research (2000). Smith, Maclin, Houghton and Hennessey found that students’ basic (Level 1) understanding of the nature of science was that science was concerned with “seeing if something works” (p. 418) or “finding an answer” (p. 418). Students with higher levels of understanding were able to comprehend that scientist use what they already know to develop ideas (Level 2). However, no students were able to understand the use of “indirect evidence” (p. 419) i.e., inferential evidence in their hypothesis testing (Level 3). This chronology of understanding in Smith, Maclin, Houghton and Hennessey’s study is mirrored by the item hierarchy
evident in the Empirical dimension of this study which suggests that the Empirical dimension is theoretically grounded.

Even though items that make up the Theory-laden dimension were the easiest to understand on average when compared to items of other dimensions, students still had a harder time understanding some items relative to others in the subscale. Item THL14E (If we do the same experiments many times, we may get different results) was conceptually harder to understand than THL12K (Scientists use different types of experiments to answer their questions) and THL6C (Scientists use different ways to test their hypotheses). The ordering of items from relatively easy to relatively hard appear to correspond to epistemologies that move from those that are relatively knowledge unproblematic to those that are knowledge problematic (Carey & Smith, 1993; Smith, Maclin, Houghton, & Hennessey, 2000).

Individuals with knowledge problematic epistemologies believe that science is objective, involves concrete prescribed procedures and activities that are in pursuit of absolute truths. At the other end of the continuum are individuals with knowledge problematic epistemologies who believe that science is a subjective, conjectural theory-driven process that pursues explanations for observable and unobservable entities. Students within the ER project and within Carey and Smith’s study (1993) and within Smith, Maclin, Houghton and Hennessey’s study (2000) appear to have a relatively easier time understanding that science involves trying things out (knowledge problematic). This involves using different types of experiments to test out hypotheses and answer questions. Students, however, in each of the studies and within the ER project had a
relatively harder time understanding the role of prior knowledge and coherence of evidence needed to explain their new scientific findings (intermediate level epistemologies). The hardest item on the Theory-laden dimension taps into knowledge problematic epistemologies and recognizes that experiments can lead to conflicting results which may lead to theories being revised. The positioning of THL8F (Scientific questions are answered by observing things) on the Theory-laden continuum conflicts with this explanation of item hierarchy within this subscale. The content of THL8K would suggest that it would be the easiest item for students to understand given that it taps into knowledge unproblematic epistemologies where young students often “think knowledge as stemming directly from sensory experience” (NRC, 2007, p. 175). This item’s difficulty was invariant over the three years of the study and in the pilot of the instrument so this finding is curious. Overall, the difficulty of the items on the Theory-laden continuum form a logical progression of student understanding of the concepts embedded within this NOS domain.

The item difficulty hierarchy of the NOS dimensions and within each dimension is largely invariant across the three Rasch models. The item difficulty hierarchy, for the most part, conforms to theoretical expectations for student NOS understanding of the concepts measured by the items. Combined with the item technical quality and the rating scale functioning, these results provide supporting evidence for the content and substantive aspect of NOS construct validity. The evidence on the integrity of the observable variables is congruent across the three Rasch models used to represent the internal structure of the NOS construct.
### Table 4.6

**Comparison of Item Difficulty Hierarchy (average deltas)**

<table>
<thead>
<tr>
<th>Item</th>
<th>Item Prompt</th>
<th>Unidimensional</th>
<th>Consecutive*</th>
<th>Multidimen. **</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Est.</td>
<td>S.E.</td>
<td>Est.</td>
</tr>
<tr>
<td>SCE13I</td>
<td>How scientists see the world is influenced by the culture they grew up in.</td>
<td>0.63</td>
<td>0.04</td>
<td>0.17</td>
</tr>
<tr>
<td>SCE13K</td>
<td>Where scientists live may affect what they are allowed to work on.</td>
<td>0.60</td>
<td>0.25</td>
<td>0.22</td>
</tr>
<tr>
<td>SCE13C</td>
<td>The country a scientist comes from influences how they understand the results of an experiment.</td>
<td>0.52</td>
<td>0.04</td>
<td>0.07</td>
</tr>
<tr>
<td>SCE13H</td>
<td>Where scientists come from may lead to different answers to the same question.</td>
<td>0.31</td>
<td>0.04</td>
<td>-0.17</td>
</tr>
<tr>
<td>SCE13F</td>
<td>A scientist’s beliefs may change how they do their work.</td>
<td>0.18</td>
<td>0.04</td>
<td>-0.29</td>
</tr>
<tr>
<td>INV6D</td>
<td>You have to be creative to work in science.</td>
<td>0.44</td>
<td>0.04</td>
<td>0.25</td>
</tr>
<tr>
<td>INV12N</td>
<td>A good imagination is needed to create the best experiment to test an idea.</td>
<td>0.29</td>
<td>0.05</td>
<td>0.08</td>
</tr>
<tr>
<td>INV8G</td>
<td>To explain their results, scientists need to be creative.</td>
<td>0.23</td>
<td>0.04</td>
<td>-0.01</td>
</tr>
<tr>
<td>INV12L</td>
<td>Although science is based on facts, scientists do need a good imagination.</td>
<td>0.16</td>
<td>0.05</td>
<td>-0.09</td>
</tr>
<tr>
<td>INV12M</td>
<td>A good imagination is needed to make predictions about what will happen in an experiment.</td>
<td>0.05</td>
<td>0.05</td>
<td>-0.23</td>
</tr>
<tr>
<td>CER14A</td>
<td>In science, finding out that a hypothesis is NOT correct is as important as finding out that a hypothesis IS correct.</td>
<td>0.36</td>
<td>0.05</td>
<td>0.58</td>
</tr>
<tr>
<td>CER8M</td>
<td>New theories in science should only be accepted when there is a lot of evidence to support them.</td>
<td>0.21</td>
<td>0.05</td>
<td>0.42</td>
</tr>
<tr>
<td>CER6H</td>
<td>A lot of data is needed to decide if a hypothesis is true.</td>
<td>-0.30</td>
<td>0.05</td>
<td>-0.11</td>
</tr>
<tr>
<td>CER7H</td>
<td>Two scientists can disagree, but both can have good ideas.</td>
<td>-0.39</td>
<td>0.05</td>
<td>-0.23</td>
</tr>
<tr>
<td>CER8L</td>
<td>When scientists have a good idea, they continue to try to make it better.</td>
<td>-0.45</td>
<td>0.05</td>
<td>-0.33</td>
</tr>
<tr>
<td>CER6J</td>
<td>Trying things out helps scientists think of new ideas.</td>
<td>-0.49</td>
<td>0.05</td>
<td>-0.33</td>
</tr>
<tr>
<td>EMP14B</td>
<td>Scientists infer what they think is happening from what they already know.</td>
<td>0.04</td>
<td>0.06</td>
<td>0.21</td>
</tr>
<tr>
<td>EMP9I</td>
<td>Experiments are used to see what happens in nature.</td>
<td>0.03</td>
<td>0.05</td>
<td>0.20</td>
</tr>
<tr>
<td>EMP8I</td>
<td>Science describes what happens in nature.</td>
<td>-0.13</td>
<td>0.05</td>
<td>0.00</td>
</tr>
<tr>
<td>EMP8D†</td>
<td>A good way to know if something is true is to do an experiment.</td>
<td>-0.14</td>
<td>0.04</td>
<td>-0.01</td>
</tr>
<tr>
<td>EMP9A</td>
<td>Scientists explain how something works.</td>
<td>-0.26</td>
<td>0.05</td>
<td>-0.17</td>
</tr>
<tr>
<td>EMP9J</td>
<td>Science helps answer questions about how something works.</td>
<td>-0.33</td>
<td>0.05</td>
<td>-0.22</td>
</tr>
<tr>
<td>THL14E</td>
<td>If we do the same experiments many times, we may get different results.</td>
<td>0.17</td>
<td>0.05</td>
<td>0.46</td>
</tr>
<tr>
<td>THL8F‡</td>
<td>Scientific questions are answered by observing things.</td>
<td>-0.11</td>
<td>0.05</td>
<td>0.17</td>
</tr>
<tr>
<td>THL9K</td>
<td>Theories can change when new evidence is found.</td>
<td>-0.14</td>
<td>0.05</td>
<td>0.16</td>
</tr>
<tr>
<td>THL8K</td>
<td>Scientists use what they found in the past to help explain their new findings.</td>
<td>-0.23</td>
<td>0.05</td>
<td>0.04</td>
</tr>
<tr>
<td>THL12K</td>
<td>Scientists create different types of experiments to test their questions.</td>
<td>-0.60</td>
<td>0.05</td>
<td>-0.38</td>
</tr>
<tr>
<td>THL6C</td>
<td>Scientists use different ways to test their hypotheses.</td>
<td>-0.65</td>
<td>0.05</td>
<td>-0.44</td>
</tr>
</tbody>
</table>

**Legend:** Est. Estimate S.E. Standard Error; Multidimen. Multidimensional

* SCE13K ordering out of sequence for Consecutive model. ** EMP9I ordering out of sequence for Multidimensional model.
The next step in the validity argument (Figure 3.4) was to ensure that the observed scores from the NOSI-E are generalizable to the population under study. These analyses are provided in the next section.

**Results from Analyses 2: Generalizability Validity Evidence**

Once the content and substantive aspects of construct validity have been established, the expectation is that the observed scores from an instrument will be meaningful and stable across different contexts and support any inferences made based on these scores (Messick, 1995; Wolfe, 2007b). It is important to provide evidence that this aspect of NOS construct validity is met regardless of the model used to represent the internal structure of the NOS construct. In Figure 4.2, a summary of the analyses needed to support the generalizability of the NOSI-E scores and the validity argument outlined in Figure 3.4 is provided.

**Analyses 2: Generalizability Evidence**

- b. Differential Item Functioning – Gender.
- c. Reliability – Person Separation (Spearman-Brown Prophecy).
- d. Precision of person estimates comparison.
- e. Literature Resources.

*Figure 4.2. Summary of generalizability validity analyses*

These analyses are composed of (a) assessing item invariance across models; (b) evaluating Differential Item Functioning (DIF); (c) comparing the reliability of person estimates and (d) an appraisal of the precision of person estimates from the three Rasch
models. Again, literature resources are reviewed and cited to provide support for the
claims and inferences made for this set of analyses.

**Item Invariance across Models.** Correlational analyses were performed to
establish the invariance of item difficulty parameters (average of item deltas) across the
three Rasch models. The overlay plot in Figure 4.3 shows the positive association
between item deltas between each Rasch model pairing.

**Figure 4.3.** Item invariance across the three Rasch models

- Uni. Con. R = 0.66
- Uni. Mlt. R = 0.66
- Con. Mlt. R = 1.00

**Legend:** Uni. Unidimensional; Con. Consecutive; Mlt. Multidimensional
As expected, there was a virtual one-to-one positive correlation ($R = 0.997$) between the Consecutive (Con.) and Multidimensional (Mlt.) model delta parameters. These models are based on the same items and constrain the same five items for parameter identification purposes for each dimension. The Unidimensional (Uni.) is based on all 28 items and constrains one item for parameter identification purposes; as a result, the relationships between the Unidimensional model and the Consecutive and Multidimensional models were weaker but still positive (0.66). The item deltas are reasonably invariant across the three Rasch models and support the internal consistency of the scale. Item invariance was also examined using gender as the grouping variable in a Differential Item Functioning (DIF) analysis.

**Differential Item Functioning.** Items should have the same meaning regardless of the group being assessed. A DIF analysis was performed to determine if male and female students of equal ability have the same probability of providing a “correct” response across all three Rasch models. It is important to establish that if a mean difference in ability between groups is significant (or not), that this is not due to bias in item content that favors one group over the other. Before the DIF analyses were run, student maximum likelihood estimates from each Rasch model were exported to SPSS and independent t-tests were performed to determine if the mean difference between male and females were significant. To adjust for multiple comparisons a p-value equal to $0.0045 (\alpha/11)$ was used to assess significance. Table 4.7 provides the mean difference and significance between males and females. The mean difference between males and
females was not significant for the unidimensional model (-0.07) or for any of the five
dimensions of the consecutive and multidimensional models (Table 4.7). This finding is
similar to studies who found no difference in student NOS understanding across gender
(Conley, Pintrich, Vekiri, & Harrison, 2004; Dogan & Abd-El-Khalick, 2008) but
conflicts with other studies who reported male students have, on average, more informed
understandings of NOS (Huang, Tsai, & Chang, 2005; Tsai & Liu, 2005).

Table 4.7

<table>
<thead>
<tr>
<th>Mean Difference in Ability Estimates by Gender (Male –Female) ¹</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Unidimensional</strong></td>
</tr>
<tr>
<td><strong>Mean Difference</strong></td>
</tr>
<tr>
<td>Empirical</td>
</tr>
<tr>
<td>Inventive</td>
</tr>
<tr>
<td>Theory-laden</td>
</tr>
<tr>
<td>Certainty</td>
</tr>
<tr>
<td>Soc. &amp; Cult. Emb.</td>
</tr>
</tbody>
</table>

Legend: Soc. & Cult. Emb: Socially & Culturally Embedded. ¹ Based on 206 Males and 198 Females.

The data in Table 4.8 were used to determine if item DIF was present between males and
females across the three Rasch models; an item was considered to exhibit DIF if the
difference in deltas were greater than 0.5 logits.

Using the standard of a difference of 0.5 logits between deltas, there were no
items in any of the three Rasch models that exhibited DIF. The biggest difference in delta
was related to the multidimensional model; the mean delta difference for item INV12L
(Although science is based on facts, scientists do need a good need imagination) was
0.44. This difference is, however, below the established criterion for DIF of 0.5 logits.
For each of the three Rasch models, the items were substantially invariant across gender
groups indicating that both groups have similar probabilities of producing a given
response.

These results provide supporting evidence that regardless of the Rasch model
used, the meaning and interpretability of the items (generalizability aspect of construct
validity) is sustained across measurement models. The estimates produced by the models
could potentially be used to measure change on the NOS variable(s). As part of assessing
the generalizability aspect of NOS construct validity, reliability analyses were performed
to compare the internal consistency across the three Rasch models. These results are
presented in the next section.
Table 4.8: 
*Differential Item Functioning across Rasch Models-Gender*

<table>
<thead>
<tr>
<th>Item</th>
<th>Unidimensional Delta Difference</th>
<th>Consecutive Delta Difference</th>
<th>Multidimensional Delta Difference</th>
</tr>
</thead>
<tbody>
<tr>
<td>EMP8D</td>
<td>0.094</td>
<td>0.074</td>
<td>0.160</td>
</tr>
<tr>
<td>EMP8I</td>
<td>0.058</td>
<td>0.086</td>
<td>0.058</td>
</tr>
<tr>
<td>EMP9A</td>
<td>0.080</td>
<td>0.106</td>
<td>0.024</td>
</tr>
<tr>
<td>EMP9I</td>
<td>0.124</td>
<td>0.096</td>
<td>0.198</td>
</tr>
<tr>
<td>EMP9J</td>
<td>0.056</td>
<td>0.080</td>
<td>0.258</td>
</tr>
<tr>
<td>EMP14B</td>
<td>0.056</td>
<td>0.100</td>
<td>0.066</td>
</tr>
<tr>
<td>INV6D</td>
<td>0.024</td>
<td>0.018</td>
<td>0.092</td>
</tr>
<tr>
<td>INV8G</td>
<td>0.080</td>
<td>0.070</td>
<td>0.168</td>
</tr>
<tr>
<td>INV12L</td>
<td>0.198</td>
<td>0.216</td>
<td>0.444</td>
</tr>
<tr>
<td>INV12M</td>
<td>0.122</td>
<td>0.112</td>
<td>0.096</td>
</tr>
<tr>
<td>INV12N</td>
<td>0.016</td>
<td>0.016</td>
<td>0.090</td>
</tr>
<tr>
<td>THL6C</td>
<td>0.158</td>
<td>0.064</td>
<td>0.240</td>
</tr>
<tr>
<td>THL8F</td>
<td>0.006</td>
<td>0.090</td>
<td>0.136</td>
</tr>
<tr>
<td>THL8K</td>
<td>0.172</td>
<td>0.076</td>
<td>0.328</td>
</tr>
<tr>
<td>THL9K</td>
<td>0.082</td>
<td>0.188</td>
<td>0.320</td>
</tr>
<tr>
<td>THI2K</td>
<td>0.030</td>
<td>0.062</td>
<td>0.146</td>
</tr>
<tr>
<td>THL14E</td>
<td>0.308</td>
<td>0.200</td>
<td>0.032</td>
</tr>
<tr>
<td>CER6H</td>
<td>0.078</td>
<td>0.062</td>
<td>0.292</td>
</tr>
<tr>
<td>CER6J</td>
<td>0.026</td>
<td>0.002</td>
<td>0.128</td>
</tr>
<tr>
<td>CER7H</td>
<td>0.036</td>
<td>0.058</td>
<td>0.026</td>
</tr>
<tr>
<td>CER8L</td>
<td>0.110</td>
<td>0.070</td>
<td>0.010</td>
</tr>
<tr>
<td>CER8M</td>
<td>0.090</td>
<td>0.048</td>
<td>0.150</td>
</tr>
<tr>
<td>CER14A</td>
<td>0.076</td>
<td>0.126</td>
<td>0.002</td>
</tr>
<tr>
<td>SCE13C</td>
<td>0.062</td>
<td>0.022</td>
<td>0.006</td>
</tr>
<tr>
<td>SCE13F</td>
<td>0.124</td>
<td>0.070</td>
<td>0.102</td>
</tr>
<tr>
<td>SCE13H</td>
<td>0.038</td>
<td>0.086</td>
<td>0.074</td>
</tr>
<tr>
<td>SCE13I</td>
<td>0.074</td>
<td>0.012</td>
<td>0.018</td>
</tr>
<tr>
<td>SCE13K</td>
<td>0.042</td>
<td>0.018</td>
<td>0.016</td>
</tr>
</tbody>
</table>
Reliability of Person Estimates. Reliability analyses were used to establish the stability and internal consistency of the instrument across the three Rasch models. The instrument reliabilities for the five dimensions of the NOS construct were compared for the consecutive and multidimensional models; the results are shown in Table 4.9 The reliability of the unidimensional approach is also provided for comparative purposes.

Table 4.9
Comparison of Dimension Reliabilities across Rasch Models

<table>
<thead>
<tr>
<th>Unidimensional</th>
<th>Dimension</th>
<th>Number of Items</th>
<th>Consecutive</th>
<th>Multidimensional</th>
<th>Increase in Test Length</th>
</tr>
</thead>
<tbody>
<tr>
<td>N/A</td>
<td>Empirical (EMP)</td>
<td>6</td>
<td>0.66</td>
<td>0.78</td>
<td>83%</td>
</tr>
<tr>
<td>N/A</td>
<td>Inventive (INV)</td>
<td>5</td>
<td>0.67</td>
<td>0.73</td>
<td>33%</td>
</tr>
<tr>
<td>N/A</td>
<td>Theory-Laden (THL)</td>
<td>6</td>
<td>0.56</td>
<td>0.73</td>
<td>112%</td>
</tr>
<tr>
<td>N/A</td>
<td>Certainty (CER)</td>
<td>6</td>
<td>0.58</td>
<td>0.73</td>
<td>96%</td>
</tr>
<tr>
<td>N/A</td>
<td>Socially &amp; Culturally Embedded. (SCE)</td>
<td>5</td>
<td>0.62</td>
<td>0.74</td>
<td>74%</td>
</tr>
<tr>
<td>0.84</td>
<td>NOSI-E</td>
<td>28</td>
<td>N/A</td>
<td>N/A</td>
<td>N/A</td>
</tr>
</tbody>
</table>

Legend: N/A Not Applicable

As shown in Table 4.9, the dimension reliabilities using the consecutive approach are lower than those using the multidimensional approach. When compared to the unidimensional reliability estimate, the loss in reliability is lessened when the multidimensional model is applied. Each dimension under the consecutive approach is
treated as a separate unidimensional model; as a result, the standard error of measurement for person estimates used to compute the reliability of each dimension is relatively large, when compared to the unidimensional approach due to the smaller number of items used to measure the latent construct. The reliabilities of the consecutive approach range from 0.56 for the THL dimension to 0.67 for the INV dimension (Table 4.9). The reliabilities for these same dimensions increase to 0.73 for both the THL dimension and INV dimension when the multidimensional approach is applied. Similarly, the dimension reliabilities increase for the remaining three dimensions when the multidimensional approach is used to model the data (Table 4.9).

Using Spearman Brown’s prophecy formula, $N = \frac{\rho_{xx'}(1-\rho_{xx'})}{\rho_{xx'}(1-\rho_{xx'})}$ where $\rho_{xx'}$ is the consecutive reliability and $\rho_{xx'}$ is the multidimensional reliability for each dimension, the increase in dimension (test) length, $N$, needed for the consecutive approach to yield the same reliability as the multidimensional can be derived. For example, an increase of 112% (equivalent to a ~13 item subscale) is needed for the reliability of the THL dimension under the consecutive approach to equal the reliability of the THL dimension using the multidimensional approach. Similarly, the test length increase required for the consecutive dimensions to match that of the multidimensional approach for the EMP, INV, CER and SCE dimensions are 83% (~11 item subscale), 33% (~7 items) 96% (~12 items), and 74% (~9 items), respectively.

Under the consecutive approach, the total number of “test” items would have to increase from 28 items to 52 items to achieve the same dimensional reliabilities as the multidimensional approach. By relinquishing the orthogonality constraint on the
dimensions and allowing them to be inter-related; the reliability of the person estimates
of the multidimensional approach nears the reliability of the unidimensional model
(0.84). The reliabilities for each dimension under the multidimensional approach
exceeded 0.7. The enhanced reliabilities of the multidimensional approach result from a
reduction in error due to the randomness of responses to each dimensions’ items. The
information of student responses to other dimensional items is used to reduce the level of
random error (Allen & Wilson, 2006). These results suggest the relative efficiency of the
multidimensional approach when compared to the consecutive approach. The results also
highlight that the multidimensional approach provides researchers with more information
on their respondents as it provides scores of reasonable reliability for each of the five
dimensions as opposed to only obtaining one composite score that results from using the
unidimensional model. In summary, these results are consistent with prior investigations
by Briggs and Wilson (2003); Allen and Wilson (2006); Wang, Yao, Tsai, Wang and
Hsieh (2006); and Cheng, Wang and Ho (2009) who reported that the reliabilities of their
subscales using the multidimensional approach were higher than those using the
consecutive approach and neared those of the unidimensional approach. The results also
support the claim that the multidimensional model can produce scores that are reliable
and suitable for use in science education research.

**Precision of Person Estimates.** These analyses compare the precision of person
estimates along the latent ability continuum across the three Rasch models. The
comparison is performed for each dimension of the NOS construct. As mentioned,
smaller standard errors are associated with more stable parameter estimates; the most
information, and by corollary most precision, in general is found where the items and person abilities match up well on the continuum. Evident from the five graphs in Figure 4.4, is that the standard errors for the unidimensional model are universally smaller across the entire latent ability continuum than those derived from the consecutive and multidimensional models; the precision pattern of the latter two models are concurrent. These results are unsurprising given that the unidimensional ability estimates are based on 28 items; this compares to 5 or 6 items for the consecutive and multidimensional approaches.

Across the models, the most precision occurs approximately between -1.0 logits and +1.0 logits. However, the distribution of ability estimates is revealing. If one was to only report the composite (unidimensional) score, a loss of information about student abilities is apparent from the graphs in Figure 4.4. The estimates from the unidimensional model, for the most part, are above -0.5 logits. However, apparent from each dimensions’ graph, there are students who struggle to comprehend concepts measured by each subscale with many students’ abilities falling below -0.5 logits. Therefore, these data suggest that important interpretable or diagnostic information is lost when the multidimensionality of a construct is ignored. These findings are investigated further in the next section when the results from an analysis of discrepant cases are provided.

The previous analyses (Analyses 2c) indicated that the reliability of person estimates derived from the multidimensional model neared those of the composite unidimensional model whereas the reliability of the consecutive model estimates were by comparison relatively low. The multidimensional model compensates for the lower
reliability of the smaller subscales by using the information (student abilities) from all the subscales to provide estimates of sufficient reliability for any one subscale. These multidimensional subscale scores are of adequate reliability to be used in science education research.

One of the main conclusions from these analyses is that the NOSI-E scores from the consecutive approach are not sufficiently reliable to warrant their use across different contexts. All consecutive subscale scores’ reliabilities were below 0.7 which indicates that the meaning, interpretation and inferences made using the scores are not likely generalizable beyond the sample. Across the three Rasch models, items were largely invariant and did not exhibit DIF which suggests at least for the two representations (unidimensional and multidimensional) of the NOS construct, the NOSI-E observed scores are of sufficient reliability and “reflect the expected scores” (Chapelle & Jamieson, 2010) across tasks and contexts. To progress the validity argument (Figure 3.4) further and to answer the research question and its related hypothesis, dimensional and other analyses were performed to support the claim that the multidimensional approach best represents the internal structure of the NOS construct. The results of these analyses are provided in the next section.
Figure 4.4. Comparison of the precision of ability estimates across the latent ability continuum.
Results from Analyses 3: Structural Validity Evidence

These analyses provide evidence for the structural validity aspect of the NOS construct. The research question posited is designed to determine which of the three possible representations (unidimensional, consecutive, and multidimensional) of the internal structure of the NOS is the best one to reflect this structure. The structure of a construct should reflect and be consistent with the theoretical framework envisioned by the instrument developers (Wolfe, 2007b). The premise of the validity argument outlined in Figure 3.4 posits that the NOS construct is best represented by the multidimensional Rasch model because it is this model that is consistent with Lederman’s theoretical framework of the NOS construct i.e., it is composed of five inter-related dimensions. The analyses to investigate this proposition are summarized in Figure 4.5. As always, the findings and conclusions from these analyses are supported and justified by providing evidence from literature resources.

### Analyses 3: Structural Validity Evidence

- b. Dimensionality Analyses – Confirmatory Factor Analyses.
- c. Rasch Subscale Correlations – Convergent Validity.
- d. Discrepant Case Analyses – Person Estimates.
- e. Literature Resources.

*Figure 4.5. Summary of structural validity analyses*

Two distinct but confirmatory methodologies were used to compare the three representations of the internal structure of the NOS construct: (a) Rasch-based methodology and (b) Confirmatory Factory Analyses (CFA) methodology. Further
structural validity evidence resulted from (c) Rasch subscale correlational analyses (restricted to consecutive and multidimensional approaches) and (d) discrepant case analyses of subscale measures (restricted to the multidimensional model). These results follow.

**Dimensionality Analyses using Rasch-based Goodness-of-Fit Statistics.** The model fit of all three models cannot be contrasted directly. The composite unidimensional and multidimensional approaches are nested (hierarchical) and can be compared using the likelihood ratio statistic $G^2$. In contrast, the consecutive approach is not hierarchical to either the unidimensional or multidimensional models and is not therefore a part of these analyses. The results of these analyses are shown in Table 4.10.

Table 4.10

*Dimensionality Analyses: Rasch Model Fit Comparison*

<table>
<thead>
<tr>
<th>Model</th>
<th>Deviance ($G^2$)</th>
<th>Number of Parameters</th>
</tr>
</thead>
<tbody>
<tr>
<td>Unidimensional</td>
<td>21,704</td>
<td>85</td>
</tr>
<tr>
<td>Multidimensional</td>
<td>21,481</td>
<td>99</td>
</tr>
<tr>
<td>Difference</td>
<td>223</td>
<td>14</td>
</tr>
<tr>
<td>Chi-square Test</td>
<td>$\chi^2_{crit.}(0.001, 14) = 36.1$</td>
<td></td>
</tr>
</tbody>
</table>

The difference in deviance between the unidimensional and multidimensional model was 223 with 14 degrees of freedom. The results, portrayed in Table 4.10, indicate that this
difference is significant at the $p < 0.01$ level of confidence. When compared to the unidimensional model, the multidimensional model for the NOSI-E data is a significantly better fit and is a more accurate representation of the internal structure of the NOS construct. Supporting evidence for this conclusion was provided by the next analyses performed; CFA was used to perform dimensionality analyses of the three representations of the NOS construct.

**Dimensionality Analyses using CFA Goodness of Fit Statistics.** CFAs of the three theoretical representations for the internal structure of the NOS construct were performed separately and a comparison of the model fit statistics, when available, was assessed. The CFA DIFFTEST provided by MPLUS (Muthén & Muthén, 2010) was used to compare fit statistics and help determine if the model fit differed significantly between a one-factor (unidimensional) model and a five-factor (multidimensional) model. Other statistics (RMSEA, CFI and TLI) explained in Chapter 3 were used to assess all models; these goodness-of-fit statistics were key to assessing the consecutive factors (the DIFFTEST is not applicable to these models as the consecutive factors are not submodels of either the one-factor unidimensional model or the five-factor multidimensional model). The fit statistics are presented in Table 4.11 with the results of the CFA DIFFTEST between the unidimensional (one factor) and multidimensional (five factor) presented in Table 4.12.

In Table 4.11, the chi-square test of model fit ($\chi^2 = 676.1$, 350 d.f.) was significant for the one-factor unidimensional model suggesting that the model was a poor fit for the data. However, the chi-square goodness-of-fit test is sensitive to sample size
(Schumacker & Lomax, 2004; Byrne, 2012) and other evidence is needed to ensure this conclusion. When using Schumacker and Lomax’s conventions for acceptable fit (Schumacker & Lomax, 2004, p. 82), the other fit statistics (RMSEA = 0.049; CFI = 0.89; TLI = 0.88) indicate that the one-factor model is reasonably well fitting with the fit statistics close to meeting the respective conventional cutoffs. The point estimate of RMSEA of 0.046 is below 0.05, the accepted standard for a well-fitting model. In addition, the upper bound of the 90% confidence interval (0.052) is only just above 0.05 further indicating that the model is fairly close-fitting. In addition, the 90% confidence interval of the one-factor model is narrow (± 0.005) providing a high degree of certainty that the estimate of the error of approximation is accurate and can help with the model’s evaluation. Although, the CFI (0.89) and TLI (0.88) indices are not above 0.95 which is the cutoff for a well-fitting model, they are close to the cutoff of 0.9 for a reasonably well-fitting model. Overall, the unidimensional one-factor model is a reasonably well-fitting model of the observed data.

The fit statistics (Table 4.11) of the multidimensional five-factor model were: $\chi^2 = 461.1$ (340 d.f.); RMSEA = 0.029; CFI = 0.96 and TLI = 0.96. Similar, to the one-factor model, the chi-square test is significant indicating poor model fit; the validity of this test is, however, called into question given the sample size used was reasonably large (N =432). The chi-square likelihood ratio test is sensitive to sample size and researchers have found that despite minor differences between a sample’s covariance matrix and its hypothesized structure, the model is often rejected when a large sample is involved (Byrne, 2012) Again, other fit statistics were used to assess the five-factor model fit.
Table 4.11

*Dimensionality Analyses: CFA Model Fit Comparison*

<table>
<thead>
<tr>
<th>Statistic</th>
<th>One Factor (Unidimensional)</th>
<th>Consecutive Factors</th>
<th>Five Factor (Multi-dimensional)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>EMP</td>
<td>INV</td>
<td>THL</td>
</tr>
<tr>
<td>$\chi^2$ test, model fit (d.f.)</td>
<td>676.1** (350)</td>
<td>36.3** (9)</td>
<td>15.3** (5)</td>
</tr>
<tr>
<td>RMSEA</td>
<td>0.046</td>
<td>0.084</td>
<td>0.069</td>
</tr>
<tr>
<td>RMSEA 90% CI (Lower)</td>
<td>0.041</td>
<td>0.056</td>
<td>0.031</td>
</tr>
<tr>
<td>RMSEA 90% CI (Upper)</td>
<td>0.052</td>
<td>0.113</td>
<td>0.110</td>
</tr>
<tr>
<td>CFI</td>
<td>0.89</td>
<td>0.95</td>
<td>0.98</td>
</tr>
<tr>
<td>TLI</td>
<td>0.88</td>
<td>0.91</td>
<td>0.96</td>
</tr>
<tr>
<td>No. of Free Parameters</td>
<td>112</td>
<td>24</td>
<td>20</td>
</tr>
</tbody>
</table>

**Legend:** RMSEA: Root Mean Square Error of Approximation; CI: Confidence Interval; CFI: Comparative Fit Index; TLI: Tucker-Lewis Index. ** $p < 0.001$.

1. The one-factor CFA model is equivalent conceptually to the unidimensional Rasch model. The five consecutive factors are equivalent conceptually to the five unidimensional Rasch consecutive models. The five-factor model is equivalent conceptually to the multidimensional Rasch model.

The point estimate of the RMSEA was 0.029, well below the cutoff of 0.05.

Similar to the unidimensional one-factor model, the 90% confidence interval of the RMSEA point estimate is narrow (± 0.007) indicating that the point estimate is precise. In contrast to the one factor model, the upper bound of the RMSEA (0.035) for the five-factor model is well below the conventional accepted cutoff of 0.05 suggesting that the multidimensional five-factor model is a better fitting model for the observed data than the
unidimensional one-factor model (this contrast is tested statistically below). Specifically, the point estimate of 0.029 and the precision of this estimate for the five-factor model indicate that the multidimensional model explains the relationship among the variables very well. Supporting this conclusion, are the results of the CFI and TLI indices; both indices are above 0.95 indicating that the five-factor structural model for the NOS construct fits the observed data closely.

The fit statistics for the five consecutive factors provide conflicting information. Interestingly, the chi-square test of model fit value for the Theory-laden and Socially & Culturally Embedded factors were not significant, indicating good model fit. Similarly, the point estimate of the RMSEA for these two factors were 0.042 and 0.000, respectively supporting that each factor explains the relationship among their respective items well. The upper bound of the 90% confidence interval for the Theory-laden factor (0.075) is greater than the conventional cutoff of 0.05. The precision of the Theory-laden RMSEA point estimate is also relatively wide (± 0.042) indicating less confidence in assessing the degree of fit of the model. The CFI (0.98) and TLI (0.97) indices for the Theory-laden factor each exceed the conventional cutoff of 0.95, again suggesting good model fit. The 90% confidence interval RMSEA upper bound of the Socially & Culturally Embedded factor is below 0.05 again signifying good model fit. When combined with the perfect CFI (1.0) and TLI (1.0) indices, the consecutive factor model fits the SCE data very well. Overall, the consecutive factor models for the Theory-laden and Socially & Culturally Embedded fit the data well.
In contrast to the Theory-laden and Socially & Culturally Embedded consecutive factors, the model fit for the Empirical, Inventive and Certainty factors are each less conclusive. The point estimates of the RMSEA for each is above 0.05 indicating relatively poor model fit. The upper bound of the RMSEA 90% confidence interval for the Empirical, Inventive and Certainty factors are 0.113, 0.110 and 0.093, respectively (Table 4.11). These values indicate poor model fit. The results from the CFI indices for these three consecutive factors conflict with the results of the RMSEA fit statistics. All three CFI indices are above 0.95 indicating the factor models capture the observed data well. The TLI index for the Inventive factor (0.97) is also above 0.95 supporting a well-fitting model. The TLI indices for the Empirical (0.91) and Certainty (0.92) factors are below 0.95 but still above 0.9 suggesting that the models are reasonably well fitting. Overall, the five consecutive factor models are reasonably well-fitting and indicate that five independent latent traits may plausibly represent the internal structure of the NOS construct.

MPLUS’s (Muthén & Muthén, 2010) CFA DIFFTEST was used to assess whether the multidimensional five factor model improved the model fit significantly when compared to the unidimensional one factor model. The results in Table 4.12 show that the difference value of 156.0 between the two models was significant ($p < 0.0001$) indicating that the multidimensional model is a better fitting model than the unidimensional model. These data provide confirming evidence that the multidimensional model better represents the internal structure of the NOS construct when compared to the unidimensional model.
Table 4.12

*Dimensionality Analyses: CFA DIFFTEST for 1-Factor vs. 5-Factor Model*

<table>
<thead>
<tr>
<th>Difference (One Factor Model – Five Factor Model)</th>
<th>( \chi^2 ) Value(^1)</th>
<th>Difference in Number of Modeled Parameters</th>
</tr>
</thead>
<tbody>
<tr>
<td>( p )-value of Chi-square Test for Difference Testing</td>
<td>( \chi^2 )</td>
<td>156.0</td>
</tr>
</tbody>
</table>

\(^1\) MPLUS uses a Weighted Least Square Mean Variance (MLSMV) Estimator in modeling categorical data; as a result the DIFFTEST procedure adjusts the \( \chi^2 \) statistic to reflect that this difference value is not distributed as a chi-square.

The decision as to whether the multidimensional five factor model is better fitting than the five independent consecutive one factor models has to be grounded in the theoretical framework for the NOS construct. It is unsurprising that the five consecutive factors have reasonable model fit as they each form a principal element of the NOS construct (McComas, 1998). However, Lederman’s (Lederman, 1992; Lederman, Abd-El-Khalick, Bell, & Schwartz, 2002; Lederman, 2007) conceptualization of the NOS construct is that these principal elements are inter-related and co-jointly form the basis of the NOS construct. The correlation and covariance estimates (shown in Table 4.13) of the five-factor CFA provide supporting evidence that the each factor is related to the other NOS factors.

Correlations ranged from a low of 0.50 between Theory-laden and Inventive factors to a high of 0.81 between the Empirical and Certainty factors. These relatively high correlations of the structural components of the CFA model indicate that the five latent factors are related to each other with the NOS construct postulated to be the
underlying variable that explains the responses to each observed factor item on the NOSI-E test.

These correlations, for the most part, parallel those reported by Lin and Tsai’s (2008) confirmatory factor analyses of students’ scientific epistemological views (SEV). Lin and Tsai measured five factors of SEV, four of which (Inventive and Creative (IC); Theory-laden (TL); Changing and Tentative (CT) and Cultural Impacts (CU)) correspond to the factors measured in this study. As an example of the agreement between the two studies, Lin and Tsai (2008) reported a correlation of 0.51 between their CT and IC factors; the correlation between the Certainty and Inventive factors reported in this study

Table 4.13

CFA Correlations and Covariances between NOS dimensions (5-Factor Model)

<table>
<thead>
<tr>
<th></th>
<th>Empirical</th>
<th>Inventive</th>
<th>Theory-laden</th>
<th>Certainty</th>
<th>Socially &amp; Culturally Embedded</th>
</tr>
</thead>
<tbody>
<tr>
<td>Empirical</td>
<td>1.0</td>
<td>0.25</td>
<td>0.27</td>
<td>0.29</td>
<td>0.21</td>
</tr>
<tr>
<td>Inventive</td>
<td>0.65</td>
<td>1.0</td>
<td>0.18</td>
<td>0.22</td>
<td>0.23</td>
</tr>
<tr>
<td>Theory-laden</td>
<td>0.79</td>
<td>0.50</td>
<td>1.0</td>
<td>0.26</td>
<td>0.20</td>
</tr>
<tr>
<td>Certainty</td>
<td>0.81</td>
<td>0.57</td>
<td>0.78</td>
<td>1.0</td>
<td>0.18</td>
</tr>
<tr>
<td>Socially &amp; Culturally Embedded</td>
<td>0.67</td>
<td>0.68</td>
<td>0.68</td>
<td>0.60</td>
<td>1.0</td>
</tr>
</tbody>
</table>

Legend: Factor Correlations are below the diagonal; Factor Covariances are above the diagonal.
is 0.57. With the exception of the correlations concerning the Theory-laden domain, the correlations in this study are of the same magnitude as those reported by Lin and Tsai (2008). In this study, the correlations between Theory-laden and any of the other three corresponding factors were higher than those reported by Lin and Tsai (2008). For example, in this study, the correlation between the Theory-laden and Certainty factors was 0.78; this compares to 0.30 in Lin and Tsai’s study. This difference may be due to Lin and Tsai’s use of only 3 items to measure the Theory-laden dimension; in contrast, six items were used in this study. Overall, the CFA analyses reported here are comparable to Lin and Tsai’s previous CFA analyses; this provides further support that the five factor multidimensional model best represents the internal structure of the NOS construct.

Based on the model fit results and these reported correlations, the multidimensional five-factor model appears the best fitting model to represent the internal structure of the NOS construct. In Figure 4.6, the five-factor model for the NOS construct is shown; all item parameter estimates (factor loadings) were significant ($p < 0.001$) as were the covariances between factors. With the exception of SCE13K (0.49), all standardized factor loadings were positive, above 0.50 and differed significantly from zero. The residual variance for SCE13K is relatively high (0.76). Because categorical data is being used, the interpretation of this residual variance differs from one used with continuous variables. This value represents the “proportion of variance in the underlying continuous and latent aspect” of SCE13K that cannot be explained by the SCE factor of the hypothesized model (Byrne, 2012, p. 144). The factor loadings are, however, reasonably substantial and indicate that the underlying factors explain the association
between the observed variables well. The moderate to large correlations highlighted in Figure 4.6 corroborate Lederman’s theoretical framework for the NOS construct (Lederman, 2007) and provide disconfirming evidence that the consecutive approach is representative of the internal structure of the construct. To ensure that the five-factor multidimensional model was not mis-specified, the modification indices available in MPLUS were requested and evaluated.

**Modification Indices.** Modification indices were requested for all of the three modeling approaches. No modification indices (MI) were reported for the unidimensional one-factor model indicating that the model parameters were not mis-specified. Of note, in the consecutive one-factor models, the MI for the Empirical factor suggested that residuals of items EMP9I (Experiments are used to see what happens in nature) and EMP8I (Science describes what happens in nature) were correlated likely reflecting some redundancy in item content. The MI was above 10, the standard used to indicate parameter mis-specification; however, Byrne (2012, p. 87) suggests that the decision to respecify a model should be based whether (1) the change is “substantively meaningful”; and (2) whether the “existing model exhibits adequate fit.” The MI value was 12.93 indicating that this mis-specification was relatively minor and not of sufficient magnitude to include this specification in a future model.
Legend: e and the associated number represent the residual variance for each item.

Figure 4.6. Five-factor CFA multidimensional model for NOS construct.
In the five-factor multidimensional model the MI indicated that item CER7H (Two scientists can disagree, but both can have good ideas) cross-loads on the Theory-laden factor. The MI was 10.7. This cross-loading is not unsurprising given that, conceptually, the theory-building process is subjective and contributes to the tentative or uncertain nature of science. The MI is still relatively small and the overall fit of the five-factor multidimensional model is excellent which suggests that changing the specification of the model is not warranted. Once a model is respecified, the researcher is no longer in confirmatory mode but has entered an exploratory mode (Byrne, 2012). Therefore, any changes to the model parameterization need to substantially affect the fit of the model for inclusion; respecification of item CER7H was therefore not undertaken as it was not deemed necessary.

**CFA Conclusion.**

In summary, the CFA analyses support the conclusion that the five-factor multidimensional model best represents the internal structure of the NOS construct. The RMSEA of the multidimensional model is well below 0.05, precise and its upper 90% confidence interval falls below 0.05 which all point to a well-fitting model. Similarly, the CFI and TLI both indicate a well-fitting model. When compared to the unidimensional one-factor model, the 5 factor model is a significantly better fitting model. The five-factor model with five inter-related dimensions is, in addition, theoretically grounded in Lederman’s (Lederman, 2007) conception of a multidimensional NOS construct and this conceptualization is supported here.
Comparison of Rasch Model Subscale Correlations. The Rasch-based correlations between subscale scores provided by the consecutive and multidimensional approaches were also compared. Conquest reports the variance-covariance and correlation matrices between dimensions when a multidimensional model is run; to obtain the correlations for the consecutive approach, the maximum likelihood estimates for each dimension have to be exported into SPSS and the Pearson correlations calculated. The results are shown in Table 4.14. The correlations between dimensions using the multidimensional approach are below the diagonal with the correlations between dimensions using the consecutive approach shown above the diagonal.

The correlations for the multidimensional approach range from a low of 0.51 (THL and INV) to a high of 0.84 (EMP and CER); these same pairings under the consecutive approach are 0.29 and 0.48, respectively. The strength of the relationships between dimensions under the consecutive approach is attenuated due to measurement error in the person estimates and this accounts for the lower correlations reported in Table 4.14. In contrast, the correlations between latent dimensions are simultaneously computed without measurement error by Conquest resulting in disattenuated correlations. Under the multidimensional approach, the correlations are moderate to high across all comparisons; this provides convergent validity evidence that the five dimensions are measuring a common, broader concept, namely, NOS. If the multidimensional approach was ignored, researchers may make the wrong conclusion that the subscales were not related by the underlying NOS latent trait due to the lower subscale correlations derived
from the consecutive approach. These results suggest that the multidimensional approach is a more accurate depiction of the dimensional structure of the NOS construct.

The correlations from the multidimensional model are also remarkably similar in magnitude to those reported using CFA (Table 4.13) for the five-factor model with eight out of the ten correlations only slightly higher than those from the CFA. Wolfe and Singh (2011) in their comparison of multidimensional MRCLM Rasch and CFA methodologies similarly noted that the latent correlations between dimensions were comparable in magnitude with the Rasch-based correlations only slightly higher for their simulated data. Both methodologies support that the five NOS dimensions are inter-related with the underlying NOS construct explaining the relationship between the observed variables. The moderate to large correlations reported from the two methodologies refute the argument that the dimensions are independent governed by separate NOS latent traits.

The results reported in this section again indicate that the multidimensional model is more efficient as it can use the information from the “true” correlational structure to improve the reliability of subscale estimates. These reliability analyses were reported in Analyses 2 and when compared to the subscale reliabilities from the consecutive approach (all < 0.7), the multidimensional subscale reliabilities were all higher (> 0.7). The results of the correlational structure of the five dimensions support the conclusion made for Analyses 2c; that is, the internal structure of the NOS construct is best represented by the multidimensional model and that the convergent scores produced by the model are more reliable and suitable for use in science education research. The next analyses – discrepant case analyses - provide additional support for this conclusion.
Table 4.14

*Rasch-based Correlations between NOS dimensions¹*

<table>
<thead>
<tr>
<th></th>
<th>Empirical</th>
<th>Inventive</th>
<th>Theory-laden</th>
<th>Certainty</th>
<th>Socially &amp; Culturally Embedded</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Empirical</strong></td>
<td>1.0</td>
<td>0.39</td>
<td>0.48</td>
<td>0.48</td>
<td>0.41</td>
</tr>
<tr>
<td><strong>Inventive</strong></td>
<td>0.67</td>
<td>1.0</td>
<td>0.29</td>
<td>0.34</td>
<td>0.42</td>
</tr>
<tr>
<td><strong>Theory-laden</strong></td>
<td>0.82</td>
<td>0.51</td>
<td>1.0</td>
<td>0.46</td>
<td>0.39</td>
</tr>
<tr>
<td><strong>Certainty</strong></td>
<td>0.84</td>
<td>0.58</td>
<td>0.81</td>
<td>1.0</td>
<td>0.32</td>
</tr>
<tr>
<td><strong>Socially &amp; Culturally Embedded</strong></td>
<td>0.67</td>
<td>0.69</td>
<td>0.66</td>
<td>0.63</td>
<td>1.0</td>
</tr>
</tbody>
</table>

¹ Multidimensional correlations are shown below the diagonal; consecutive correlations are above the diagonal (based on standardized measures).

**Discrepant Case Analyses.** These analyses investigate whether student ability estimates from the multidimensional model differ substantially across the five dimensions when each dimension is treated as a “general” measure of NOS. In addition, the variance of the five dimension scores from the composite unidimensional estimate is estimated and the percentage of discrepant cases are provided using cutoffs reported by Briggs and Wilson (2003) and Allen and Wilson (2006). If student ability estimates differ substantially across the five dimensions, this would indicate that using one dimension as a general measure of NOS would likely underestimate or overestimate student ability estimates. Similarly, if scores from the subscales, as a whole, differ substantially from the
unidimensional composite score, this would also suggest that the composite score is not reflective of a student’s ability in each dimension and the opportunity to provide differential information on a student is lost. Figure 4.7 and 4.8 show the relationship between standardized ability estimates of the two dimensions with the highest correlations (Empirical and Certainty) and the lowest correlation (Inventive and Theory-laden). For reference, the lines around the fit lines in Figure 4.7 and Figure 4.8 represent a one standard deviation confidence interval band.

The percent of discrepant cases (differ by more than one standard deviation) were 30.1% when the ability estimates were compared across the Empirical and Certainty dimensions (R = 0.48); the comparable figure was 36.8% when the ability estimates were compared across the Inventive and Theory-laden dimensions (R = 0.29). These data show that these discrepant students’ abilities would be either under-reported or over-reported providing a false picture of their “true” ability.

When all dimensions were considered simultaneously, the results from computing the sums of squares discrepancy indicator, $DI_p$, indicated that the number of discrepant cases was even larger. Using the 0.5 cutoff (Briggs & Wilson, 2003), 91.7% of the cases were discrepant; using the 1.0 cutoff (Allen & Wilson, 2006), 75.9% of the cases were discrepant. These data provide supporting evidence that each dimension offers differential information on students’ NOS ability and could not, individually, be used to provide an estimate of a student’s “general” NOS ability. Similarly, if the composite (unidimensional) score was only reported, this differential information would be lost along with important diagnostic information on students’ strengths and weaknesses.
Figure 4.7. Relationship between Certainty and Empirical Person Estimates

Figure 4.8. Relationship between Theory-laden and Inventive Person Estimates
The structural validity evidence from the discrepant case analyses helps justify the claim that the multidimensional model best represents the internal structure of the NOS construct.

In assessing the totality of the evidence from the structural validity analyses, the evidence supports the conclusion that the NOS construct is multidimensional and, in the Rasch framework, best measured using the multidimensional Rasch model. Both the Rasch-based methodology and CFA provide congruent evidence that the multidimensional model is the best fitting model for the NOSI-E data. The disattenuated correlations of the multidimensional subscale scores are consistent with the theoretical framework for the NOS construct and with those reported in the literature (Lin & Tsai, 2008). In addition, the results from the discrepant case analyses suggest that if a unidimensional approach was used to represent the NOS construct, researchers and practitioners would not receive an accurate picture of students’ abilities and have no means to assess students’ strengths and weaknesses across the five conceptually distinct dimensions.

Results from Analyses 4: External Validity Evidence

Thus far, the structural validity analyses have provided evidence to support the conclusion that the NOS construct is multidimensional and composed of five inter-related dimensions. The results from the content, substantive and generalizability aspects of construct validity have shown that the NOSI-E dimension scores from the multidimensional approach are theoretically grounded, psychometrically sound and reliable. The last inference of the validity argument outlined in Figure 3.4 is concerned
with whether scores from the multidimensional construct can be utilized to determine if theory-based, predicted changes in person measures can be extrapolated or realized from the measures. In the Rasch framework, this is partly assessed by determining if the instrument is responsive and able to detect these changes. If the instrument is responsive to changes in students’ NOS understanding using the multidimensional approach, then, it is justifiable to use the dimension scores from the multidimensional model in the remaining external validity analyses that seek to assess their expected relationship with other measures.

Messick (1995, p. 746) proposes that of “special importance” are the external relationships that provide evidence that “attests to the utility of the scores for the applied purpose” of the assessment. In this study, these other measures are scores from the CIER content knowledge assessment related to the ER intervention and classroom perception scores from the ESSCES instrument, both of which were administered concurrently with the NOSI-E. The analyses to support the external validity aspect of NOS construct validity are summarized in Figure 4.9.

**Analyses 4: External Validity Evidence**

- a. Responsiveness – Variable Maps & Number of Person Strata.
- d. Literature Resources.

*Figure 4.9. Summary of external validity analyses*
These analyses included: (a) an assessment of the responsiveness of the person measures across the three Rasch models using the Person Strata Index and variable maps; (b) five separate multi-level regression models to determine if NOSI-E dimension scores can predict student performance on the CIER content assessment; and (c) five separate multi-level models to determine if NOSI-E dimension scores can predict students’ perceptions of the constructivist nature of their science classroom learning environment. The results are reported in the following sections.

**Responsiveness of Person Measures.** The ability of a scale to measure change in individuals is central to the utility of measurement instruments. If an instrument in particular is not responsive to individual change, then the purpose and utility for which it was built is called into question (Wolfe & Smith, 2007b). In a Rasch framework, item calibrations (and their related item thresholds) should be reasonably distributed on the scale metric axis and near juxtaposed to the person distribution. In this manner, measures are reliable and any change in student understanding of the latent construct can be precisely estimated. If, for example, items are too easy for students prior to an intervention, then measuring student improvement in the target construct due to the intervention will not be feasible as ceiling effects will prevent an accurate assessment of the change in student understanding (Wolfe & Smith, 2007b). In this situation, the instrument is not considered responsive.

An examination of the variable maps and the calculation of the person strata index was undertaken to assess the responsiveness of the NOSI-E instrument across the three Rasch models. The unidimensional model item threshold map is shown in Figure 4.10.
The Consecutive Empirical, Inventive, Theory-laden, Certainty and Socially and Culturally Embedded threshold maps are shown in Figures 4.11, 4.12, 4.13, 4.14, and 4.15, respectively. The variable threshold map for the Multidimensional model is shown in Figure 4.16. For the four response categories of the NOSI-E, there are three thresholds. In each variable map, the threshold number is separated from the item number by a period. The portion of the person distribution not covered by the item threshold calibrations is shaded in gray in each map. The percent of persons that are covered by the item calibrations is reported in Table 4.15 along with the Person Strata Index.

Table 4:15

<table>
<thead>
<tr>
<th>Dimension</th>
<th>Consecutive [% Coverage]</th>
<th>Multidimensional [% Coverage]</th>
</tr>
</thead>
<tbody>
<tr>
<td>Empirical</td>
<td>2.2 [78.0%]</td>
<td>2.8 [76.4%]</td>
</tr>
<tr>
<td>Inventive</td>
<td>2.2 [77.1%]</td>
<td>2.5 [86.6%]</td>
</tr>
<tr>
<td>Theory-laden</td>
<td>1.8 [69.1%]</td>
<td>2.5 [78.7%]</td>
</tr>
<tr>
<td>Certainty</td>
<td>1.9 [79.9%]</td>
<td>2.5 [84.3%]</td>
</tr>
<tr>
<td>Soc. &amp; Cult. Emb.</td>
<td>2.0 [91.6%]</td>
<td>2.6 [97.5%]</td>
</tr>
</tbody>
</table>


**Unidimensional Approach.**

The responsiveness of the unidimensional model is good with a Person Strata Index (H) of 3.4 (Table 4.15); over 90% of the student distribution is covered by item calibrations (Figures 4.10) indicating that the instrument is reasonably well targeted for
this sample. The mean and standard deviation of the unidimensional model is 0.87 and 0.69, respectively. The positive mean indicates that, on average, the sample found the items too easy suggesting that more difficult items are needed to better match up the item calibrations and person distribution. If the purpose of the study was to differentiate students into three statistically distinct groups (e.g., into those with naïve, adequate and informed understanding); the instrument would serve this purpose well. A Person Strata Index of 3.4 would enable students to be reliably classified as having naïve, adequate or informed understanding of NOS. However, if the purpose of the instrument is to measure change in student NOS understanding (e.g., resulting from an intervention), for some students, there would be ceiling effects and their true understanding of NOS would not be measured accurately. There are 33 students (shaded in gray) who have a greater than 50% probability of fully endorsing their understanding of the most difficult item on the instrument (SCE13I). Given that some NOS concepts were not measured by the instrument (e.g., student understanding of the difference between observation and inference), more difficult items could be added to the test in order for change to be measured in these students and to obtain a more reliable estimate of their true abilities.

In addition, the preponderance of the student distribution in this sample is above the first threshold for each item in the scale. This suggests that the number of response categories could be reduced without impacting the efficiency of the instrument. However, given that this sample was a convenience sample and relatively small for Rasch analyses, the targeting of the item calibration needs further investigation before a decision should be made on this aspect of the instrument’s external validity.
**Consecutive Approach.**

The coverage (Figures 4.11 – 4.15) and Person Strata Index (Table 4.15) of the consecutive dimensions is, in general, very poor and is associated with the poor reliability of the measures. With the exception of the Socially & Culturally Embedded dimension, the coverage of students by the item calibrations is below 80% raising the distinct possibility that for 20% of students ceiling effects would occur. This poor coverage results in relatively poor reliability for each dimension (all below 0.7). The poor coverage of item calibrations is also evident in the modal region of the person distribution. For example, the Inventive threshold map (Figure 4.12) reveals a lack of item calibrations between 0.25 logits and 1.30 logits; the estimation of students’ abilities situated between these two locations is problematic and contributes to the overall poor reliability of the subscale. These gaps are also evident for all the other consecutive subscales as evident in Figure 4.11 (Empirical); Figure 4.13 (Theory-laden); Figure 4.14 (Certainty) and Figure 4.15 (Socially & Culturally Embedded). If change in student NOS dimensional understanding was the purpose of a research study, this lack of coverage and responsiveness would exacerbate efforts to gain an accurate picture of students’ improvement resulting from an intervention for example. If the measures are not reliably estimated, the inferences made about the abilities of the examinees are questionable.

The item means for the Empirical ($\bar{x}$: 1.18; $s.~d.$ 0.94), Theory-laden ($\bar{x}$: 1.26; $s.~d.$ 0.85) and Certainty ($\bar{x}$: 1.15; $s.~d.$ 0.82) subscales are all above 1.0 logit indicating that the students, on average, found the items too easy. The consequence of this for the Theory-laden dimension for example, is that 133 students (~31% of sample)...
are poorly estimated as there are no item calibrations to cover the person distribution at the upper portion of the scale metric axis (shaded area). The Inventive mean ($\bar{x}: 0.78; s. d. 1.02$) and Socially & Culturally Embedded mean ($\bar{x}: 0.43; s. d. 0.74$) are between 0 and 1 logits; this indicates that for these dimensions, the items are reasonably well targeted for the sample. Based on the Person Strata Index (Table 4.15), students in the Empirical (H = 2.2), Inventive (H = 2.2) and Socially & Culturally Embedded (H = 2.0) dimensions can reliably be classified into those with a naïve understanding and those with an informed understanding. In contrast, for the Theory-laden (H = 1.8) and Certainty (H= 1.9) dimensions where coverage is relatively poor, this level of differentiation is not feasible. As mentioned, Khisfe (2008) postulates that the transition from naïve NOS understanding to an informed NOS perspective is one that shifts student epistemology along a developmental continuum. It involves a transition phase that can include fluctuating student views that can be unstable and context dependent and one that can include multiple co-occurring and, at times, conflicting levels of understanding. Given this reality, without the addition of more items, none of the dimensional scales using the consecutive approach could reliably assess students whose views are transitional in nature.

**Multidimensional Approach.**

The improved reliability of the five NOS dimensions using the multidimensional approach has a direct impact on the responsiveness of the dimensions across the instrument. For each dimension, the Person Strata Index, H, is larger when compared to the same dimensions measured with the consecutive approach (Table 4.15). For example,
for the Theory-laden dimension, H increases from 1.8 to 2.5, a substantial improvement. Using the multidimensional approach, students on this dimension could reliably be differentiated into 2 statistically distinct groups providing important information on student understanding within this dimension. However, H, for each dimension is still below 3 (Table 4.15) indicating that students could not reliably be divided into three distinct ability groupings (naïve, adequate and informed) making measuring change for the higher ability students difficult. Schumacker and Smith (2007, p.400, Table 1) report that to differentiate three statistically distinct groups, the person reliability associated with an H of 3.0 is 0.8. Using the Spearman Brown Prophecy formula, to obtain this level of differentiation, the total number of items using the multidimensional approach in the Empirical, Inventive, Theory-laden, Certainty, and Socially & Culturally Embedded dimensions would have to increase from 6, 5, 6, 6, 5 items to 7, 8, 9, 9 and 7 items, respectively. This would increase the instrument length to 40 items. Although this is a substantial increase in overall test length (43%), the corresponding instrument using the consecutive approach would have to include 73 items (161% increase in test length) to provide the same level of responsiveness, a much greater and impractical burden on examinees.

Figure 4.16 presents the threshold map for the multidimensional approach. The multidimensional approach uses the information on students’ abilities from the other dimensions to provide a more reliable estimate of their abilities on the target dimension. It was reported that for the Inventive domain using the consecutive approach (Figure 4.12), there were gaps in the items calibrations between 0.25 and 1.3 contributing to the
relatively poor reliability of the subscale. Evident in the variable map for the
multidimensional approach (Figure 4.16), items from other dimensions fill this gap and
this information (students’ responses to these items) is used to provide more reliable
estimates of students’ abilities on the Inventive dimension.

The table insert of Figure 4.16 indicates that the means of the dimensions range
from 0.44 for Socially & Culturally Embedded (the most difficult on average) to 1.28 for
Theory-laden (the easiest on average). Similarly, the standard deviations indicate that the
variability of responses increases from a low of 0.74 for Socially & Culturally Embedded
to a high of 0.98 for the Empirical and Inventive dimensions. These means all have
positive logit values indicating that the students, on average, found the items of each
subscale too easy. If the goal was to measure change in the subscale variable or
differentiate students into three performance levels, the multidimensional subscales, as
they stand currently, are not sufficiently responsive. However, the previous evidence
(Rasch model fit comparisons and CFA) has shown that the NOS construct is
multidimensional and to ignore this determination would also have repercussions.

Although the consecutive approach provides similar information about students’
abilities on each dimension, the reliability of the student estimates has been shown
(Analyses 2; Table 4.9) to be too low for practical use. Using the unidimensional
approach, this differentiated information is lost completely as the model provides only a
composite score for each student. To illustrate the consequence further of not using the
multidimensional approach, two profiles of individual students are provided in Table
4.16; these students both received the same raw and modeled score using the unidimensional model.

Both students scored 60 out of a possible 84 raw score points and as a result received the same Rasch ability estimate (0.81) using the unidimensional approach. A comparison of their multidimensional scores reveals a markedly different score profile. Both students had similar understanding of the Theory-laden concepts tested (both scored 1.25 logits) but Student A had a better understanding of the Empirical and Inventive dimension content when compared to Student B. Student A scored approximately one standard deviation higher on the Inventive dimension than Student B. In contrast, Student B had a much better understanding of the Certainty and Socially & Culturally Embedded content than Student A. Student B’s score on the Certainty dimension was greater than Student A’s score by over one standard deviation.

If a teacher is presented with the two unidimensional scores for these students, the teacher could make the incorrect inference that these students are of equal ability across the different facets of the NOS construct. The ability estimates from the multidimensional approach show that this is not the case with each student having different strengths and weaknesses. This differentiated information could be used by the teacher to individualize instruction and help students overcome their particular misconceptions. For relatively low stakes formative assessment decisions, the reliability of the five dimensions (> 0.7 for all dimensions) using the multidimensional approach is adequate. The benefit of the multidimensional approach is that it provides more diagnostic information to the teacher upon which she/he can tailor their instruction.
Table 4.16

*Comparison of Two Students’ Ability Estimates*

<table>
<thead>
<tr>
<th>Student ID</th>
<th>Raw Score</th>
<th>Rasch Score</th>
<th>Unidimensional Estimates</th>
<th>Multidimensional Estimates (in logits)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>Empirical</td>
<td>Inventive</td>
</tr>
<tr>
<td>A</td>
<td>60</td>
<td>0.81</td>
<td>0.62</td>
<td>1.38</td>
</tr>
<tr>
<td>B</td>
<td>60</td>
<td>0.81</td>
<td>0.14</td>
<td>0.35</td>
</tr>
</tbody>
</table>

*Legend:* Soc. & Cult. Embedded: Socially & Culturally Embedded

The importance of ensuring an instrument is responsive to change is highlighted in Songer and Gotwals (2012) study. These researchers used Rasch-based methodology to construct a hierarchical and scaffolded content knowledge assessment of elementary (G4, G5 and G6) students’ content knowledge of biodiversity and ecology, and of their ability to construct scientific explanations. Students were exposed to an intervention that scaffolded their learning with activities to help them build scientific explanations during their curricular units. The scaffolding was not only present during the intervention but was also embedded within the assessment. Scaffolding within the assessment was implemented by having some items provide content and explanation construction hints with others having only explanation construction hints and others having no hints. Students’ performance was compared pre and post intervention using variable maps that model this hierarchical assessment. Songer and Gotwals (2012) concluded that the greater learning gains evident in the variable maps (pre versus post-test maps) for G5 and G6 students when compared to G4 students were likely due to the assessment items being too
hard for G4 students and being “off target” both pre and post intervention. As a result, the
assessment was not responsive to changes in student abilities due to the intervention and
was not as useful for this particular age group.

The results provided in this subsection and from all the prior analyses point to the
conclusion that the NOS construct is multidimensional and best represented by the
multidimensional Rasch model. With the caveat on the lower than ideal responsiveness of
the multidimensional subscales, the five dimension scores were used in assessing their
expected relationships with other measures. The remaining external validity analyses
follow.
Figure 4.10. Unidimensional NOSI-E Variable Threshold Map
Figure 4.11. Consecutive Empirical Variable Threshold Map

Legend: S.D. Standard Deviation; S.E. Standard Error; EAP: Expected a Posteriori reliability estimate; H. Person Strata Index.
Figure 4.12. Consecutive Inventive Variable Threshold Map

Legend: S.D. Standard Deviation; S.E. Standard Error; EAP: Expected a Posteriori reliability estimate; H. Person Strata Index.
**Legend:** S.D. Standard Deviation; S.E. Standard Error; EAP: Expected a Posteriori reliability estimate; H. Person Strata Index.

**Figure 4.13.** Consecutive Theory-laden Variable Threshold Map
Figure 4.14. Consecutive Certainty Variable Threshold Map

Legend: S.D. Standard Deviation; S.E. Standard Error; EAP: Expected a Posteriori reliability estimate; H. Person Strata Index.

Each “x” represents 3.0 cases

Mean 1.15
S.D. 0.82
S.E. 0.04
EAP 0.58
H. 1.9
Figure 4.15. Consecutive Socially & Culturally Embedded Variable Threshold Map.

Legend: S.D. Standard Deviation; S.E. Standard Error; EAP: Expected Posteriori reliability estimate; H. Person Strata Index.
<table>
<thead>
<tr>
<th>Dimension</th>
<th>Thurstone-Item Thresholds</th>
</tr>
</thead>
</table>

**DIMENSION**

<table>
<thead>
<tr>
<th>EMP</th>
<th>INV</th>
<th>THL</th>
<th>CER</th>
<th>SCE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean</td>
<td>1.20</td>
<td>0.76</td>
<td>1.28</td>
<td>1.14</td>
</tr>
<tr>
<td>S.D.</td>
<td>0.98</td>
<td>0.98</td>
<td>0.83</td>
<td>0.82</td>
</tr>
<tr>
<td>S.E.</td>
<td>0.05</td>
<td>0.05</td>
<td>0.04</td>
<td>0.04</td>
</tr>
<tr>
<td>EAP</td>
<td>0.78</td>
<td>0.73</td>
<td>0.73</td>
<td>0.73</td>
</tr>
<tr>
<td>H</td>
<td>2.8</td>
<td>2.5</td>
<td>2.5</td>
<td>2.5</td>
</tr>
</tbody>
</table>

Legend: S.D. Standard Deviation; S.E. Standard Error; EAP: Expected a Posteriori reliability estimate; H: Person Strata Index.

*Figure 4.16. Multidimensional NOSI-E Variable Threshold Map*
Multilevel Regression Analyses. The goal of these analyses was to provide further evidence for the external validity aspect of the NOS construct and to assess the inference that the NOS instrument provides scores that are interpretable and suitable for use in science education research and teaching (the applied purpose of the NOSI-E assessment). Given the conclusion from the prior analyses that the NOS construct is multidimensional, scores used in these analyses are from the multidimensional Rasch model. Therefore, five sets of scores are used in these analyses with each set of scores representing one of the dimensions (Empirical, Inventive, Theory-laden, Certainty and Socially and Culturally Embedded).

Concurrent with administrating the NOSI-E, students were asked to complete a content knowledge assessment (Content Inventory for Evolution Readiness; CIER) related to the ER intervention and provide a measure of their perceptions of the constructivist nature of their science classroom learning environment (Elementary School Science Classroom Environment Scale; ESSCES). These are the outcome variables for the two sets of analyses in the following two subsections. However, it is important to first understand how the data was structured in the ER cohort research design and how the structure relates to multilevel modeling; this is depicted in Figure 4.17.

Students are at (Level 1) in the model. These students were nested within cohorts at Level 2 (L2) which were synonymous with the three years of the ER project. These cohorts were taught by the same teachers over the three years of the project and, ideally, the teacher would represent Level 3 (L3) in the MLM. Only students of teachers who participated in all three years of the ER project were included in these analyses. Overall,
387 of the 432 students (90%) were included in the final regression models. As can be seen from Figure 4.17, these 387 students are nested in one of three cohorts (years) and further nested within one of 10 teachers’ classrooms. The variance available to be explained at each level is determined using the intraclass correlations.

The intraclass correlation (ICC) is used to partition the variance and conveys the proportion of variance in CIER scores that can be explained by the three levels when no predictors are added to the model (the unconditional model). The percentage of variance at each level can be calculated by using Equation 4.1 (Level 1); Equation 4.2 (Level 2) and Equation 4.3 (Level 3), respectively. With level-1 variance, $\sigma^2$, equal to 0.227; level-2 variance, $\tau_\pi$, equal to 0.791; and level-3 variance, $\tau_\beta$, equal to 0.001, the total variance
available to model for the standardized CIER scores is 1.019. The percentage of variance among students (Level 1) is 77.6%; the percentage of variance among cohorts within teachers (Level 2) is reasonably substantial at 22.3%. The percentage of variance among teachers is only 0.1% suggesting that a two-level model would suffice for the data. The data are hierarchical and the substantial variance at level-2 (among cohorts) indicates that the use of MLM is appropriate. Predictors at Level 1 were added to the MLM to try to explain the variance in CIER scores at this level; there were no predictors available to add at Level 2 or at Level 3.

\[
\text{Level 1: } \frac{\sigma^2}{\sigma^2 + \tau_\pi + \tau_\beta} \quad \text{Eq. 4.1}
\]

\[
\text{Level 2: } \frac{\tau_\pi}{\sigma^2 + \tau_\pi + \tau_\beta} \quad \text{Eq. 4.2}
\]

\[
\text{Level 3: } \frac{\tau_\beta}{\sigma^2 + \tau_\pi + \tau_\beta} \quad \text{Eq. 4.3}
\]

Where:

\( \sigma^2 = \text{Variance among students.} \)

\( \tau_\pi = \text{Variance among cohorts.} \)

\( \tau_\beta = \text{Variance among teachers.} \)

\( \sigma^2 + \tau_\pi + \tau_\beta = \text{Total Variance in CIER scores.} \)
The MLM results for the regression using CIER as the outcome and NOSI-E dimension scores as independent variables are provided in the next subsection with the MLM for the ESSCES and NOSI-E variables provided in the following subsection.

**Predicting CIER Achievement.**

This section begins with a discussion of the problems associated with performing these analyses; these problems relate to power, reliability and collinearity of the predictors. By corollary, the solutions that were implemented to address these problems are discussed. The section proceeds with (1) the treatment of the variables used in the final MLM; (2) the statistical model used for the MLM; (3) the findings and discussion of the regression coefficients effect sizes and the variance components of the final MLM developed, and (4) providing supporting evidence from the literature for the results determined herein.

**Problems with CIER MLM.** There were several problems in conducting the MLM analyses and this resulted in changes to the analyses strategy. “Optimal Design Plus Software” (Raudenbush et. al., 2011) was used to determine the power underpinning the analyses. After evaluating the results of this power analysis combined with the ICC computations above, it was determined that it was better to perform a two-level MLM and not a three-level one. Using a three-level model (in which 387 students are nested within 3 cohorts and further nested among 10 teachers) and assuming a two-tailed test; an $\alpha$ equal to 0.05; a level-2 ICC of 0.223; and power of 0.8, the analyses only had the power to detect an effect size of 0.8 (a large effect). An additional problem was that the reliability of the level-2 (cohort) intercept was very low (< 0.05). Raudenbush and Bryk
(2002, p. 125) posit that when the reliability of a coefficient “drops below 0.05, that coefficient is a candidate for treatment as a fixed or nonrandomly varying”. These results justified changing the MLM to two levels with students at level-1 and teachers at level-2.

Unfortunately, there were additional problems related to the level-2 model. There was still a lack of power to detect the fixed effects of the models. Using a two-tailed test; an $\alpha$ equal to 0.05; a level-2 ICC of 0.086; and power of 0.8; the two level-model only had the potential to detect an effect of 0.65 (a moderate to large effect). In addition, and not unsurprisingly, there was a significant problem with collinearity of the five NOSI-E predictors. The collinearity among NOS variables was a major constraint on using the level-2 model and prevented the NOS predictors from being added simultaneously into one MLM. To address these issues, the predictor slopes were fixed (intercepts only model) and the CIER outcome variable was regressed on each NOS predictor separately. Specifically, five intercept-only models were run, one for each NOS dimension. Of note, gender was not a significant predictor of student CIER achievement and this predictor was not included in the models presented here. Given these realities, the data structure for the MLM can be better represented in Figure 4.18.
In Figure 4.18, students at level-1 are now nested within teachers’ classrooms (level-2). For this data structure (Figure 4.18), the ICC was computed. The percentage of variance at each level can be calculated by using Equation 4.4 (Level 1); and Equation 4.5 (Level 2), respectively. With level-1 variance, $\sigma^2$, equal to 0.932 and level-2 variance, $\tau_\beta$, equal to 0.087; the total variance available to model for the CIER scores is 1.019. The percentage of variance among students (Level 1) is 91.5%; the percentage of variance among teachers (Level 2) is 8.5%.

\[
\text{Level 1: } \frac{\sigma^2}{\sigma^2 + \tau_\beta} \quad \text{Eq. 4.4}
\]

\[
\text{Level 2: } \frac{\tau_\pi}{\sigma^2 + \tau_\beta} \quad \text{Eq. 4.5}
\]

Where:

$\sigma^2 = \text{Variance among students.}$

$\tau_\beta = \text{Variance among teachers.}$

$\sigma^2 + \tau_\beta = \text{Total Variance in CIER scores.}$
All NOS subscale scores from the multidimensional approach and the outcome CIER variable were standardized to put each set of scores on the same scale. By standardizing the variables, the regression coefficients can be considered to represent an effect size for each NOSI-E predictor entered; given the low power of these analyses, the discussion of the results will focus on the effect sizes and less on the significance of the predictors.

To account for the nesting of students within cohorts (that the research design portrayed in Figure 4.18 ignores), a cohort dummy variable was developed. Two cohort dummy variables were used to represent Cohort 2 (Year 2) and Cohort 3 (Year 3) in the model with Cohort 1 (baseline) representing the comparison variable. These dummy variables were entered at level-1. Normally, the first model (Model 1) entered into the MLM is the null model with no predictors added and this is used as the reference model to determine the impact of entering predictors at each level (Raudenbush & Bryk, 2002). However, to provide a more accurate representation of the ER data portrayed in Figure 4.17, the reference model used in this study is the one with the cohort dummy predictors entered into the model (Model 2). As a result, the conditional variance components of Model 2 (Table 4.17) are used to determine the percentage of variance explained by the five NOSI-E predictors added.

The results of the five NOSI-E models are presented in Table 4.17. In Model 2, the cohort dummy variables were entered; the reference group for the cohort dummy variables (Cohort 2 and Cohort 3) was Cohort 1 in the model. In Model 3 through Model 7, the respective NOS subscale scores were added separately. There were no available
teacher measures (i.e., level-2 predictors) to add to the model to explain the variation in student achievement among teachers. Each of the five CIER models estimates two parameters. To account for the multiple tests undertaken in these analyses and in the ESSCES MLMs to follow, a significance level of 0.0025 (\(\alpha/20\)) was used throughout the analyses to determine if the parameter estimates were statistically significantly different from zero. The statistical model used for the analyses is presented next and this is followed by a discussion of the results from the models.

**CIER Statistical Model.** The multidimensional NOSI-E model provided five dimensional scores of reasonable reliability (person reliabilities, \(R\), were > 0.7); these NOSI-E scores were entered into five separate fixed-effects MLMs to determine if the NOSI-E scores could predict student achievement on the content knowledge test (CIER). The conditional, intercepts only, statistical model for these analyses is shown in Equation 4.6. Equation 4.6 portrays the mixed model for the analyses presented.

\[
\text{CIERSTD}_{ij} = \beta_{00j} + \beta_{10j} \text{(Cohort 2)} + \beta_{20j} \text{(Cohort 3)} + B_{kj} \text{(NOSI-E subscale scores)} + r_{ij} + \mu_{0j}
\]

Eq. 4.6

Where fixed components are:

- \(\text{CIERSTD}_{ij}\) is the adjusted CIER score for student \(i\) in teacher \(j\)’s cluster.
- \(\beta_{00}\) is the level-2 mean achievement across all students.
- \(\beta_{10j}\) is the pooled within regression coefficient of \(\text{CIERSTD}_{ij}\) on cohort 2.
- \(\beta_{20j}\) is the pooled within regression coefficient of \(\text{CIERSTD}_{ij}\) on cohort 3.
\( \beta_{kj} \) is the pooled within regression coefficient of CIERSTD\(_j\) on each respective NOSI-E subscale, \( k \).

Where random components are:

\( r_{ij} \) is the error (adjusted deviation) in CIER achievement associated with student \( i \) in teacher cluster \( j \).

\( \mu_{0j} \) is the adjusted deviation in CIER achievement associated with teacher cluster \( j \).

The cohort dummy variables were entered first to partial out the variance due to students being in different cohorts. Then, one of the five NOSI-E predictors was entered into the model. As mentioned, all continuous variables were standardized to have a mean of zero and variance of one. Therefore, because the variables are all standardized, the regression coefficient in each model can be interpreted as: for every one standard deviation increase in the NOSI-E predictor, CIER achievement increases by “\( \beta \)” standard deviations. The output from the five models is shown in Table 4.17; in Table 4.17, significant predictors and variances are bolded.

In the final models (Model 3 through Model 7), the reliability of the level-2 intercept was above 0.75 for each of the five models. Further, although the parameter estimates are reported with robust standard errors, i.e., the largest reported standard errors, the HLM program reported that the number of level-2 units (teachers) was too low to make their use appropriate. Robust standard errors adjust for the dependency in the errors due to the hierarchical nature of the data; given the hierarchical nature of the data used in this study, using the robust standard errors was still justified. A comparison of the
robust standard errors with the ordinary least square errors revealed that in all comparisons, the robust standard errors were larger. Therefore, the use of the robust standard errors was deemed appropriate as it was a more conservative approach to assessing the significance of the regression coefficients.

Regression Results. Table 4.17 shows that students’ CIER scores are significantly and positively related to their Theory-laden NOS dimension scores (Model 5: $p < 0.0025$). For every one standard deviation unit increase in NOS Theory-laden (THL) scores, CIER scores are predicted to increase by just over a quarter of a standard deviation (0.27). Students who have a better understanding of the subjective nature of science and the conjectural theory building process are predicted to perform higher on the CIER science content knowledge test. No other NOSI-E dimensions were a significant predictor of students’ scores on the CIER assessment. However, it is has to be remembered the power analysis showed that the level-2 model only has the power to see moderate to large effects so these results are somewhat expected. When comparing the effect sizes of the regression coefficients, the Theory-laden dimension had the largest positive effect (0.27) followed by Certainty (0.14) and Empirical (0.06). The effect sizes for both the Inventive (-0.09) and Socially and Culturally Embedded (-0.07) were both very small and negative.

It is hard to make any substantive conclusion on these results due to the analyses being underpowered. It does seem, however, that despite this lack of power (and the caveat that it was the only predictor entered, once the cohort effect was partialled out), one conclusion from the analyses is that students’ theory-laden NOSI-E scores were
predictive of students’ scores on the CIER achievement test. Theory-laden subscale scores explained 8.0% of the student-to-student variance in student CIER scores (Table 4.17); in addition, the THL subscale scores explained a small amount of the variance (1.8%) in student achievement among teachers (level-2). The total variance explained (7.4%) by the THL subscales scores combines the two sources of variability (within and between teachers’ classrooms) in the model.

It could be postulated that the lack of responsiveness of the multidimensional subscales, highlighted in the previous section, accounts for the low effect sizes of the four other subscales (EMP; INV; CER and SCE). A restriction in range of the four other dimensions could, in theory, impact the ability to detect the relationship between them and CIER. Items from the Theory-laden were however, on average, the easiest for students to understand. This would suggest that this subscale would be the most susceptible to restriction of range as ceiling effects may occur. Further support to refute this hypothesis was gleamed from the person strata index and percentage of person coverage by the item calibrations. The THL subscale had a person strata index of 2.5 (lower than both the EMP and SCE dimensions and of the same magnitude as INV and CER) with 78.7% of the student distribution covered by item calibrations (the second lowest coverage of all the dimensions). These data would argue against this hypothesis as the THL effect was still evident even when it was one of the most unresponsive subscales in comparison to the other subscales.

*Literature Resources.* The literature review found only a small number of quantitative studies that examined students’ relationship between their NOS
understanding and their science content knowledge or achievement. Two studies examined student science achievement and its relationship with students’ understanding of the theory-laden nature of science (Chuy et. al., 2010; Songer & Gotwals, 2012). Two other research studies examined students’ content knowledge and its relationship with a composite measure of NOS (Tsai, 1998b; Peters & Kitsantas, 2010).

Chuy et al., (2010) in their comparative study of fourth grade students determined that students who were part of a class using a theory-building approach to learning science exhibited higher scientific literacy than students in a comparison class that followed a more traditional curriculum. Chuy et al., results indicate that fourth grade students are capable of understanding the theory-laden nature of science and it can be hypothesized that students with a greater understanding of this dimension would perform better on science achievement tests. It can further be postulated that these higher performing students likely have more problematic epistemologies and the critical thinking skills associated with this epistemology may aid them in content knowledge tests especially those that contain open-response items such as those used in the CIER.

Support for this hypothesis is provided in Songer and Gotwals (2012) research study. In this study, the authors developed an intervention that integrated learning progressions and scaffolds to teach elementary students the core ideas related to biodiversity and ecology and to the scientific practice of developing explanations (theory-laden). Using multiple regression analyses (but not MLMs), Songer and Gotwals found that fourth, fifth and sixth grade students made significant gains in their assessment of
### Table 4.17

**Multilevel Models: Predicting students’ science achievement (CIER)**

| Predictor                          | CIER Achievement (Standard Deviation Units) | | | | |
|-----------------------------------|--------------------------------------------|--|--|--|--| |
| | Conditional Reference\(^2\) Model 2 | Empirical Model 3 | Inventive Model 4 | Theory-laden Model 5 | Certainty Model 6 | Soc. & Cult. Embedded Model 7 |
| Adjusted Intercept (Standard Error) | -0.330 (0.138) | -0.334 (0.141) | -0.326 (0.135) | -0.303 (0.143) | -0.343 (0.136) | -0.331 (0.134) |
| Cohort 2                          | 0.460 (0.158) | 0.465 (0.161) | 0.456 (0.158) | 0.400 (0.144) | 0.477 (0.156) | 0.459 (0.158) |
| Cohort 3                          | 0.374 (0.254) | 0.373 (0.251) | 0.371 (0.244) | 0.364 (0.235) | 0.391 (0.249) | 0.385 (0.248) |

**NOS Subscale Scores**

<table>
<thead>
<tr>
<th></th>
<th>Empirical (EMPSTD)</th>
<th>Inventive (INVSTD)</th>
<th>Theory-laden (THLSTD)</th>
<th>Certainty (CERSTD)</th>
<th>Soc. &amp; Cult. Embedded (SCESTD)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Empirical (EMPSTD)</td>
<td>0.065 (0.049)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Inventive (INVSTD)</td>
<td></td>
<td>-0.092 (0.079)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Theory-laden (THLSTD)</td>
<td></td>
<td></td>
<td>0.273 (0.035)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Certainty (CERSTD)</td>
<td></td>
<td></td>
<td></td>
<td>0.138 (0.063)</td>
<td></td>
</tr>
<tr>
<td>Soc. &amp; Cult. Embedded (SCESTD)</td>
<td></td>
<td></td>
<td></td>
<td>-0.069 (0.071)</td>
<td></td>
</tr>
</tbody>
</table>

**Variance Components**

<table>
<thead>
<tr>
<th></th>
<th>Available Variance</th>
<th>Residual Variance</th>
</tr>
</thead>
<tbody>
<tr>
<td>Level 1 (among students)</td>
<td>0.900</td>
<td>0.898</td>
</tr>
<tr>
<td>Level 2 (among teachers)</td>
<td><strong>0.084</strong></td>
<td>0.084</td>
</tr>
<tr>
<td>Level 1 (among students)</td>
<td>0.893</td>
<td>0.828</td>
</tr>
<tr>
<td>Level 2 (among teachers)</td>
<td>0.0860</td>
<td>0.082</td>
</tr>
<tr>
<td>Level 1 (among students)</td>
<td>0.884</td>
<td>0.079</td>
</tr>
<tr>
<td>Level 2 (among teachers)</td>
<td><strong>0.083</strong></td>
<td></td>
</tr>
<tr>
<td>Total Variance Explained</td>
<td>0.2%</td>
<td>0.6%</td>
</tr>
<tr>
<td>Percent Variance Explained</td>
<td>0.2%</td>
<td>0.6%</td>
</tr>
</tbody>
</table>

**Legend:**


1. Values that are **bolded** were significant at the 0.0025 level.
2. Model 1 (not shown) is the null model with no predictors added (see text for details of this model). Model 2 is the more appropriate reference model as it takes account of the cohort effect.
content knowledge and students’ ability to construct explanations between the pre-intervention assessment and post-intervention assessment. Besides pre-intervention test scores as a significant predictor of post-intervention achievement, students that had more exposure to the intervention (the number of worksheets that scaffolded how to construct explanations in science was used as a proxy for the level of student exposure to the intervention) scored significantly higher on the content assessment (effect sizes: G4: 0.128; G5: 0.093; G6: 0.184). It seems that students who developed more problematic epistemologies through their exposure to the intervention performed better on the content achievement test.

Two other studies that have reported the relationship between students’ NOS understanding using composite scores and student content knowledge. Tsai (1998b) found no correlation between G8 students’ epistemological views and achievement in a small-scale study of cognitive structures and their relationship to students’ understanding of atomic theory and students’ epistemological views. In Tsai’s study (1998b), students were not taught about the nature of science explicitly and the study was very small (48 students) which suggests they may not have had enough power to see an effect. Peters and Kitsantas (2010) randomly assigned classes of students to an intervention that used metacognitive prompts to teach students about the nature of science during a science unit on electricity and magnetism; comparison classes received the same instructional unit but did not receive the metacognitive intervention. Peters and Kitsantas developed a scoring rubric for the VNOS-B instrument to measure student understanding of NOS and produce a composite score. Based on all participants in the study, Peter and Kitsantas (2010)
report a correlation of 0.52 between students’ VNOS-B scores and students’ unit content knowledge scores. Unfortunately, Peter and Kitsantas (2010) did not report separate correlations for each group of students to tease out the impact of the intervention on the relationship between NOS understanding and science content knowledge.

Overall, the positive correlation between VNOS-B and content knowledge scores in Peters and Kitsantas’s study (2010) and the positive results of the theory-laden intervention in Songer and Gotwals study (2012) do suggest that the positive effect between NOSI-E theory-laden scores and CIER content knowledge scores witnessed in the MLM is reasonable especially given the metacognitive nature of the ER study’s intervention. There may be other NOS dimensions that are related to students’ achievement in this study but there is a lack of power to detect these effects. Therefore, the evidence presented here to support the external validity aspect of NOS construct validity is not sufficiently robust and studies using more rigorous methodology such as used by Peters and Kitsantas (2010) are needed to fully understand the relationships between NOS understanding and student science achievement.

*Predicting Students’ Perceptions of their Classroom Learning Environment.*

The analysis strategy was the same as the one developed for the CIER MLM. The problems identified in the CIER MLM are relevant to these analyses too. Due to the lack of power for these analyses and the multicollinearity of the predictors; a two-level model was performed for each NOSI-E predictor using the unidimensional ESSCES score as the outcome variable. Specifically, scores from the five dimensions of the multidimensional NOSI-E model were entered into five separate intercept-only, multilevel regression
models (MLM) to determine if the NOSI-E scores could predict students’ perceptions of their science classroom learning (ESSCES). As a reminder, students were asked to endorse the extent that various constructivist practices were used within their science classroom learning environment when completing the ESCCES instrument.

The prior analyses reported that a three-level model was not feasible given the structure (387 students nested within 3 cohorts and further nested within 10 teachers’ classrooms) of the ER data. Therefore, similar to the CIER analyses a two-level model was performed for the ESSCES outcome. Using Equation 4.4 and Equation 4.5, the first null model (Model 1) performed was to calculate the ICC for the power analyses. With level-1 variance, $\sigma^2$, equal to 0.920 and level-2 variance, $\tau_{\rho}$, equal to 0.150; the total variance available to model for the ESSCES outcome scores is 1.070. The percentage of variance among students (Level 1) is 86.0%; the percentage of variance among teachers (Level 2) is 14%. The ICC of 0.14 was used in the power analyses for this two-level MLM. Using a two-tailed test; an $\alpha$ equal to 0.05; a level-2 ICC of 0.14; and power of 0.8; the two level-model only had the potential to detect an effect of 0.81 (a large effect). Therefore, as in the case of the CIER MLM, these analyses only have the power to detect a large effect.

As stated, normally the first model that would be performed is to obtain the ICC with no predictors added and it is this model that is used to assess the impact of other predictors in the model (Raudenbush & Bryk, 2002). However, as in the CIER MLM, the appropriate reference model is the one in which the cohort dummy predictors are added. The two-level model ignores the nesting of students within cohorts; as a result, the cohort
dummy variables were added at level-1 to control for the cohort effect. Therefore, the variance components of this model (Model 2; Table 4.18) are more appropriate and are used to assess the impact of adding students’ NOSI-E dimension scores as predictors. As with the NOSI-E predictors, the ESSCES outcome variable was standardized to have a mean of zero and a standard deviation of one. This enables effect sizes of the predictors to be compared across the five MLMs as all variables are on the same scale.

The results of the five NOSI-E models are presented in Table 4.18. In Model 2, the cohort dummy variables were entered (again using Cohort 1 as the reference variable). In Model 3 through Model 7, the respective NOS subscale scores were added separately. There were no teacher measures (i.e., level-2 predictors) to add to the model to explain the variation in student perceptions among teachers. Similar to the CIER MLMs, a significance level of 0.0025 ($\alpha/20$) was used throughout the analyses to determine if the parameter estimates were statistically significantly different from zero. The statistical model used for the analyses is presented next and this is followed by a discussion of the results from the models.

**ESSCES Statistical Model.** The mixed statistical model for these analyses is shown in Equation 4.7. As with the previous analyses, the cohort dummy variables were entered first to partial out the variance due to the cohort effect. Then, one of the five NOSI-E predictors was entered into the model. The output from the five models is shown in Table 4.18. Significant predictors are bolded.
\[
\text{ESSCES}_{ij} = \beta_{00j} + \beta_{10j} (\text{Cohort 2}) + \beta_{20j} (\text{Cohort 3}) + B_{kj} (\text{NOSI-E subscale scores}) + r_{ij} + \mu_{0j}
\]

Eq. 4.7

Where:

Fixed Components are:

- \text{ESSCES}_{ij} is the adjusted ESSCES score for student \(i\) in teacher \(j\)’s cluster.
- \(\beta_{00j}\) is the level-2 mean achievement across all students.
- \(\beta_{10j}\) is the pooled within regression coefficient of ESSCES\(_{ij}\) on cohort 2.
- \(\beta_{20j}\) is the pooled within regression coefficient of ESSCES\(_{ij}\) on cohort 3.
- \(B_{kj}\) is the pooled within regression coefficient of ESSCES\(_{ij}\) on each respective NOSI-E subscale, \(k\).

Random Components are:

- \(r_{ij}\) is the error (adjusted deviation) in ESSCES endorsement associated with student \(i\) in teacher cluster \(j\).
- \(\mu_{0j}\) is the adjusted deviation in ESSCES endorsement associated with teacher cluster \(j\).

It should be noted that in the final models (Model 3 through Model 7), the reliability of the level-2 intercept was above 0.80 for each of the five models. Robust standard errors are reported with the knowledge that the appropriateness of using them is doubtful given the low number of level-2 units.
Regression Results. Students’ ESSCES scores are significantly and positively related to all five NOSI-E dimension scores (Table 4.18). The regression coefficients for the Empirical (Model3; 0.271); Inventive (Model 4; 0.233); Theory-laden (Model 5; 0.275); Certainty (Model 6, 0.217) and Socially & Culturally Embedded (Model 7; 0.213) dimensions were all significant at the 0.0025 level of confidence. For every one standard deviation unit increase in NOS Empirical (EMP) and Theory-laden (THL) scores, ESSCES scores are predicted to increase by just over a quarter of a standard deviation. Similarly, for every one standard deviation increase in Inventive (INV), Certainty and Socially & Culturally Embedded (SCE) scores, ESSCES scores are predicted to increase by just over a fifth of a standard deviation. Because the model variables are all standardized, these coefficients can be considered to be effect sizes; the effect sizes are all relatively small.

Empirical (8.1%) and Theory-laden (8.5%) subscale scores explained the highest amount of level-1 variance in student ESSCES scores of all NOS predictors; in addition, the EMP and THL subscale scores explained a relatively substantial amount of the variance (15.5% and 17.7%, respectively) in students’ among teachers (level-2). The Inventive, Certainty and Socially & Culturally Embedded predictors explained 5.8%, 4.9% and 4.7% of the level-1 variance in ESSCES scores, respectively; the corresponding percentages at level-2 were 17.1% (INV); 14.5% (EMP); and 19.4% (SCE). Each of the fixed-effects model (Model 3 –Model 7) estimates two parameters as does the reference model (Model 2). Therefore, it is not possible to use the “deviance” statistic (-2LL) to compare model fit. In reality, if these NOS predictors were entered together, they would
likely share variance. The effect size of each NOS predictor on its own is small but indicative of a positive relationship between students’ nature of science understanding and the extent that they perceive their classroom learning environment as constructivist.

To make any definitive conclusions based on this data is difficult due to the lack of power for the analyses. Overall, however, the results indicate that students who perceived their science classroom learning environment to be more constructivists in nature appear to have more informed epistemologies on each dimension of the NOS construct i.e., students’ scores on the NOSI-E dimensions were predictive of students’ ESSCES scores. Evidence from the literature can help support this interpretation of the results.

*Literature Resources.* The pivotal role that teachers and the classroom environment can play in students’ ability to develop more informed epistemologies is evident when one contrasts the results from the qualitative studies that investigated students’ uptake of more sophisticated epistemologies using an implicit approach designed to inculcate NOS (Kang, Scharmann, & Noh, 2005; Akerson & Abd-El-Khalick, 2005; Sandoval & Morrison, 2003, Bell, Blair, Crawford, & Lederman, 2003; Wu & Wu, 2011) and the results from studies that used a more explicit approach to teach NOS (Smith, Maclin, Houghton, & Hennessey, 2000; Khisfe & Abd-El-Khalick, 2002, Kawasaki, Herrenkohl, & Yeary, 2004; Akerson & Volrich, 2006; Khisfe, 2008; Akerson & Donnelly, 2010; Akerson, Buck, & Quigley, 2011; Quigley, Pongsanon, & Akerson, 2011). Students, in general, did not develop more informed views of NOS unless they
were taught NOS through the use of an explicit approach that attempted to model how scientists in real life would undertake their work.

Essentially, what distinguished the two approaches to teaching NOS were the teaching practices used to instruct students within the classrooms. The classroom environments of the “explicit studies” cited were constructivist in nature and this learning environment helped students develop a more informed understanding of NOS. In these “explicit” studies, the science activities were for the most part hands-on, often student-led and the teacher expressively and frequently used reflective discussions and metacognitive prompts to elicit students’ views on NOS and their discernment of its relationship with science knowledge construction. In contrast, the pedagogy used in the “implicit” studies cited above relied more heavily on teacher-led science activities where “how to” directions were provided to students and the activities were expected to lead to a given result. It seems that this difference in pedagogy may explain the difference in uptake of more informed epistemologies among students.

In a qualitative study that examined the relationship between NOS understanding and learning environments more directly (i.e., it was the focus of the study), Tsai (1998a) found that students who preferred a more constructivist learning environment had more informed scientific epistemological views. In a follow-up study, Tsai (2000) quantified this relationship and reported statistically significant positive correlations (between 0.17 and 0.22) between students’ epistemological views and their perceptions of three aspects (the classroom encouraged a high degree of student on student negotiation; provided ample time to use prior knowledge to construct new knowledge; and encouraged student-
led activities) of the classroom learning environment that was used to measure its orientation toward constructivism. In contrast to Tsai’s (2000) finding, Peters and Kitsantas’s study (2010) found no relationship between students’ VNOS-B scores and the Metacognitive Orientation Scale – Science (MOLES-S). The MOLES-S was designed to measure the social constructivist nature of the science classroom learning environment. There was a substantial difference in the number of students involved in the two studies with a much larger sample (1,176) in Tsai’s study than in Peter and Kitsantas’s study (83); this may help explain the different results.

Overall, the consensus from the qualitative studies suggests that a higher degree of constructivist pedagogy (found predominantly within the “explicit studies) within the science classroom is positively related to more informed NOS understanding. Given that NOS was not “explicitly” taught in the ER project or was a part of the students’ regular curriculum, finding positive relationships (albeit of small effect sizes) between NOS understanding and students’ views on the degree to which they view their science classroom learning environment reformed provides support for the external validity aspect of NOS construct validity.

**Externality Validity Conclusion.** Of concern in making the inference that the NOSI-E scores support the external validity aspect of the multidimensional NOS construct is the responsiveness of the multidimensional subscales to changes in student understanding on each dimension. The responsiveness of the subscales require improvement which can be accomplished by added a few items to each subscale that are of greater average difficulty than those currently used. Improving the responsiveness of
the subscales will help ensure that the researchers can measure change in the variables and use the subscale scores to examine relationships with other measures with more confidence. In addition, the lack of power in the MLMs is of concern and likely impacted the ability to see meaningful relationships among variables.

These concerns are not ameliorated by being able to provide support from the literature for the MLM results reported in this study. There are a limited number of quantitative research studies that are concerned with the relationship between students’ NOS understanding at the dimensional level and students’ content knowledge achievement and between students’ NOS understandings and their relationship with the classroom learning environment (Deng, Chen, Tsai, & Chai, 2011). This makes it hard to assess if the relationships identified here are meaningful and supportive of the external validity aspect of the NOS construct. However, despite these caveats, the limited quantitative studies supporting the findings of the relationship between NOS dimensional scores and CIER achievement scores and the qualitative and quantitative studies that support the findings of a relationship between NOS dimensional scores and students’ views on their classroom learning environment suggest that the results reported here are reasonably plausible. It also has to be remembered that these results are based on a convenient sample so any inferences made have to be viewed with caution and are not generalizable. On balance, however, it is concluded that the results of the two MLMs provide reasonable support for the external validity aspect of the multidimensional NOS construct. The decision as to whether the NOSI-E instrument is ready to “attest to the utility of the scores for its applied purpose” (Messick, 1995, p. 746) should be left until
the NOSI-E is used in further research that uses more powerful and rigorous methodologies.

**Conclusion**

This chapter presented and discussed the results of several analyses that were used to assess the different validity aspects needed to support the validity argument that the NOS construct, as measured by the NOSI-E, is multidimensional and best represented by a multidimensional Rasch model. In Chapter 5, the chain of inferences that links the building blocks of the validity argument from its premise to its conclusion is reflected upon along with a discussion of the suitability of using the scores from the NOSI-E in science education research and teaching. In addition, Chapter 5 will discuss the limitations of the study and recommendations for future research in this field of study.
Table 4.18:

Multilevel Models: Predicting Student’s Views of their Classroom Learning Environment. (ESSCES)

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Adjusted Intercept (Standard Error)</td>
<td>-0.169 (0.174)</td>
<td>-0.198 (0.166)</td>
<td>-0.189 (0.164)</td>
<td>-0.142 (0.156)</td>
<td>-0.197 (0.164)</td>
<td>-0.172 (0.157)</td>
</tr>
<tr>
<td>Cohort 2</td>
<td>0.554 (0.164)</td>
<td>0.595 (0.158)</td>
<td>0.567 (0.168)</td>
<td>0.490 (0.159)</td>
<td>0.589 (0.158)</td>
<td>0.559 (0.160)</td>
</tr>
<tr>
<td>Cohort 3</td>
<td>-0.134 (0.139)</td>
<td>-0.124 (0.131)</td>
<td>-0.117 (0.132)</td>
<td>-0.144 (0.132)</td>
<td>-0.100 (0.139)</td>
<td>-0.161 (0.124)</td>
</tr>
</tbody>
</table>

NOS Subscale Scores

<table>
<thead>
<tr>
<th>Subscale</th>
<th>Available Variance</th>
<th>Residual Variance</th>
</tr>
</thead>
<tbody>
<tr>
<td>Empirical (EMPSTD)</td>
<td>0.271 (0.055)</td>
<td></td>
</tr>
<tr>
<td>Inventive (INVSTD)</td>
<td>0.233 (0.049)</td>
<td></td>
</tr>
<tr>
<td>Theory-laden (THLSTD)</td>
<td>0.275 (0.041)</td>
<td></td>
</tr>
<tr>
<td>Certainty (CERSTD)</td>
<td></td>
<td>0.217 (0.067)</td>
</tr>
<tr>
<td>Soc. &amp; Cult. Embedded (SCESSTD)</td>
<td></td>
<td>0.213 (0.059)</td>
</tr>
</tbody>
</table>

Variance Components

<table>
<thead>
<tr>
<th>Level</th>
<th>Available Variance</th>
<th>Residual Variance</th>
</tr>
</thead>
<tbody>
<tr>
<td>Level 1 (among students)</td>
<td>0.828</td>
<td>0.760</td>
</tr>
<tr>
<td>Level 2 (among teachers)</td>
<td>0.141</td>
<td>0.119</td>
</tr>
</tbody>
</table>

Percent Variance Explained

<table>
<thead>
<tr>
<th>Level</th>
<th>Explained Variance</th>
</tr>
</thead>
<tbody>
<tr>
<td>Level 1 (among students)</td>
<td>8.1%</td>
</tr>
<tr>
<td>Level 2 (among teachers)</td>
<td>15.5%</td>
</tr>
<tr>
<td>Total Variance Explained</td>
<td>9.2%</td>
</tr>
</tbody>
</table>


1. Values that are **bolded** were significant at the 0.0025 level.
2. Model 1 (not shown) is the null model with no predictors added (see text for details of this model). Model 2 is the more appropriate reference model as it takes account of the cohort effect.
Chapter Five: Conclusion

The results from each set of analyses in Chapter 4, along with the associated literature cited, were each designed to address an aspect of NOS construct validity. The results supply the evidence needed to justify the claim that the NOS construct is multidimensional and the measures derived from the NOSI-E are reliable, interpretable and suitable for use in science education research and teaching. These empirical analyses were framed by a specific validity argument that was constructed to outline what evidence is needed to determine if the NOS construct is, as hypothesized, multidimensional and to justify the use of scores from the NOSI-E instrument in science education research. Messick (1980, 1995) stresses that it is not just the empirical evidence that justifies the meaningfulness and interpretability of scores but the rationales used to support the trustworthiness of score interpretations and the relationship of scores to other measures.

In this Chapter, the summary evidence and rationales for each step of the validity argument, outlined in Figure 3.4 (p. 125) is discussed with the goal of providing a coherent explanatory framework for the major conclusion resulting from this study. The chapter will begin with an overview of the study. It will then proceed with a discussion of the evidence marshaled from the analyses to support the validity argument and, by corollary, provides a unifying explanation of the evidence to support the meaningfulness, interpretability and utility of the multidimensional NOSI-E scores. The chapter will conclude with the limitations of the current study and recommendations for future
research related to the NOSI-E instrument’s development and to its use in science education research and teaching.

**Overview of Study**

In the NRC’s (2011) new framework, the NRC explicitly states that research is needed into “how engagement in specific practices supports the development of both specific (core) ideas in science and understanding of the nature of science” (p. 13-4). The NRC suggests that researchers need to develop instruments that can be efficiently used in “large-scale testing contexts” (NRC, 2011, p. 13-6) in order to examine the relationship between scientific practices, content knowledge and NOS understanding. Given the importance that the NRC (2011) and other researchers (Ryan & Aikenhead, 1992; Carey & Smith, 1993; McComas, Clough, & Almazroa, 1998; Smith, Maclin, Houghton, & Hennessey, 2000; Tsai, 2000; Conley, Pintrich, Vekiri, & Harrison, 2004; Lederman, 2007; Khisfe, 2008; Akerson & Donnelly, 2010) place on having students understand the nature of scientific knowledge construction, a need existed for a measure that can reliably and appropriately assess students’ NOS views. The Nature of Science Instrument-Elementary (NOSI-E) was developed to address this need (Peoples, O’Dwyer, Shields, & Wang, in review). The Rasch-based NOSI-E measures elementary students’ views on the five domains (Empirical, Inventive, Theory-laden, Certainty and Socially & Culturally Embedded) that comprise the internal structure of the NOS construct; these dimensions are premised on Lederman’s theoretical framework for the NOS construct (Lederman, 1992; Lederman, Abd-El-Khalick, Bell, & Schwartz, 2002; Lederman, 2007).
Prior research work, related to the Evolution Readiness project, had developed items that were designed to measure students’ understanding of NOS (Peoples, O’Dwyer, Shields, & Wang, in review); 28 items were retained from this instrument (NOSI-E) development process and were used in this study. This study described and reported on the validation activities performed to assess the internal structure of the NOS construct and to evaluate the suitability of using the scores derived from the administration of the NOSI-E in science education research and teaching. This dissertation addressed one encompassing research question related to these validation activities. The research question of this study was concerned with determining which of three Rasch-based models (unidimensional, consecutive or multidimensional) best represents the internal structure of the NOS construct. Inherent in this question is the underlying issue of whether the NOS construct, as measured by the NOSI-E, is multidimensional or composed of one or five unidimensional constructs.

The rich quantitative and qualitative data produced by Conquest (Wu, Adams, Wilson, & Haldane, 2007) enabled the three Rasch models to be compared on several aspects of construct validity (content, substantive, generalizability, structural and external). The results from these comparative Rasch analyses (and from other methodologies) showed that the NOS construct is multidimensional, composed of five inter-related but separate dimensions and, in the Rasch framework, is best measured using the multidimensional Rasch model. The perspective taken in this dissertation is that construct validity is the unifying force for all aspects of validity highlighted above. Wolfe (2007b, p. 220) posits that the external aspect of validity is “arguably the most important
aspect”; it is concerned with the meaning of the construct as it relates to other measures. Messick (1995) suggests that it is these relationships which are critical to establishing the utility of the construct under “applied conditions” (p. 1017). In this dissertation scores from the NOSI-E instrument were examined to determine if they were meaningful as they relate to other measures, namely students’ science achievement (as measured by CIER) and students’ views on the constructivist nature of their science classroom learning environment (as measured by ESSCES).

The responsiveness of the NOSI-E instrument across the three Rasch models was compared using the Person Strata Index and variable maps. In addition, two sets of multilevel models (MLMs) were run to determine if the NOSE-I dimension scores were predictive of student science achievement (CIER) and of students’ perceptions of their classroom learning environment (ESSCES). Although in need of improvement, the subscales of the multidimensional Rasch model were found to be reasonably responsive indicating that they would be able to measure change in the construct. The results from the MLMs also provided reasonable support for the claim that the scores from the five NOSI-E subscales behaved as expected in terms of their predicted relationship with both science achievement and students views of their classroom environment. However, this claim is accompanied by the caveat that the MLMs were problematic (multicollinearity was an issue and resulted in the need for simpler models to be formulated) and lacked the power to accurately portray the relationships between the variables.

When building evidence for the validity argument for this study, an important component of this was to provide corroborating or conflicting evidence from the
literature for the findings reported in Chapter 4. As a result, the evidence from the literature became an integral part of the results section. Therefore, the discussion of the findings is primarily found in Chapter 4. However, given the centrality of the validity argument to this dissertation, it is incumbent here to revisit the validity argument and provide readers with a coherent explanatory synthesis of the evidence, inferences, interpretations and claims made at each step of the validity argument to reach the conclusion that the NOS construct is multidimensional, best represented by the multidimensional Rasch model and the subscale scores derived from the model are reliable, interpretable and suitable for use in science education and teaching.

Validity Argument Revisited

The proposition was made that the internal structure of the NOS construct is best represented and measured by the multidimensional Rasch model. Before one can begin to address this proposition, one needs to provide evidence for the first inference of the validity argument i.e., the NOS construct can be appropriately described and defined. Evidence from a literature review was used to support the theoretical model used in this study for the NOS construct. The theoretical framework was heavily influenced by Lederman’s work (Lederman, 1992; Lederman, Abd-El-Khalick, Bell, & Schwartz, 2002; Lederman, 2007) on the nature of science. Lederman has spent his life’s work on studying students’ and teachers’ understanding of the nature of science and as a result his expertise in this area is profound. Based on Lederman’s conceptualization of the NOS construct, items were developed to measure five (Empirical, Inventive, Theory-laden, Certainty and Socially & Culturally Embedded) of the eight possible NOS dimensions;
these five dimensions were considered age-appropriate for elementary students. 

Lederman’s conceptualization of the NOS construct theorizes that the five dimensions are inter-related but separate dimensions and further infers that the items developed for each dimension are not related to the other dimensions of the construct. The assumption being made related to this domain description inference is that the content of the items developed for the NOSI-E is representative and relevant to the target latent construct(s). This multidimensional proposition competes with other plausible internal structures for the NOS construct; the NOS construct could be unidimensional or composed of five independent consecutive dimensions. These other internal models for the NOS construct can similarly be assessed using Rasch models. The hypothesis related to the internal structure of the NOS construct is testable by using the observed student responses to the NOSI-E.

The observed responses are used to evaluate whether the items’ quality and scoring structures provide scores that are measurable and reflective of the students’ understanding of the latent NOS construct. The technical quality, rating scale functioning and the item difficulty hierarchy of the 28 items were compared across the three Rasch models. The results of these analyses (Analyses 1) showed that, regardless of the Rasch model used (unidimensional, consecutive or multidimensional) the 28 items exhibited excellent model fit, good rating scale functioning and item hierarchies (between and within subscales) that conformed to theoretical expectations and evidence derived from the literature review. Inferred from these results, is the claim that the scoring structure used for items performed as expected and students’ observed performance on the “test”
provides the evidence needed to assess their level of NOS understanding. This claim is plausible across all three Rasch models. The findings and claim from the evaluation inference provide the foundation for the next building block of the validity argument. The validity argument can progress up the chain of inferences.

The generalization inference is predicated on the claim that the observed scores are reliable and accurate estimates of expected scores if the NOSI-E was used across different contexts and subgroups. Analyses were performed to provide evidence that the scores from the NOSI-E are generalizable and not impacted by the Rasch model used. The results of the analyses performed indicated that items were invariant across the three Rasch models and across gender groups. The findings of the DIF gender analyses were consistent regardless of the Rasch model used. The reliability analyses, however, produced evidence that the scores from the consecutive model would not provide reliable ability estimates as all subscale person separation reliabilities were below 0.7. This finding suggests that the scores from the NOSI-E as modeled by the consecutive approach would produce scores that are not generalizable across contexts and calls into question if the consecutive scores could be interpreted appropriately and used for their intended purpose.

Similarly, the comparison of the standard errors in relation to the person ability estimates indicated that the unidimensional scores, although reliable (person separation reliability of 0.84) may not accurately reflect students’ abilities on the NOS construct, especially given the premise that the NOS construct is multidimensional. For both the consecutive and multidimensional approaches, there were a substantial number of
students who scored below -0.5 logits on each dimension; this information was lost using the unidimensional approach where the ability distribution predominantly situated above -0.5 logits. This suggests that for the unidimensional approach, scores may be misinterpreted and students’ abilities may be misrepresented. It is at this stage of the validity argument and associated analyses that the evidence collected to support the different validity aspects of the NOS construct diverged across the three Rasch models. The multidimensional model produced scores that did not exhibit DIF; were reasonably reliable (all subscales had person separation reliabilities above 0.7) and appeared to provide a more accurate picture of students’ ability estimates on each of the five dimensions. However, this is only one body of evidence that suggests that the NOS construct is multidimensional and best represented by the multidimensional Rasch model. The evidence is not sufficiently compelling to disconfirm the plausibility of using the rival unidimensional or consecutive approaches to represent the internal structure of the NOS construct. For example, items could be added to the consecutive dimensions to improve their reliability and the scores from each of the dimensions would be interpretable, meaningful and generalizable across contexts.

One of the assumptions for the explanation inference in the chain of inferences is that the internal structure of NOSI-E scores is consistent with the theoretical framework used to represent the construct. The expected scores can be “attributed to” (Chapelle & Jamieson, 2010, p. 8) a construct of NOS proficiency. The theoretical framework for the NOS construct used in this study postulates that it is composed of five inter-related but separate dimensions (Empirical, Inventive, Theory-laden, Certainty and Socially &
Culturally Embedded). However, to support this postulate requires the disconfirmation of plausible rival internal structures for the NOS construct; namely that the construct is unidimensional or is composed of five independent consecutive dimensions. The results from the analyses predicated on Rasch methodology showed that the multidimensional model (when compared to the unidimensional model) was a significantly better fit or explanation of the item response data and that the subscales were inter-related. These results provide further disconfirming evidence that the unidimensional model is an accurate representation of the internal structure of the NOS construct. Through these analyses, evidence was also provided that undermines the use of the consecutive approach to represent the NOS construct. Disattenuated correlations of the multidimensional approach ranged from 0.51 between the Inventive and Theory-laden subscales to 0.84 between the Empirical and Certainty dimensions. These moderate to large correlations indicate that the five dimensions are inter-related and are not separate independent constructs.

Other methodology was employed to substantiate the finding that the multidimensional Rasch model best represents the internal structure of the NOS construct i.e., the NOS construct is multidimensional. Confirmatory Factor Analyses is a confirmatory technique that can be used to compare model fit between different plausible representations of a construct. The results mirrored those found using Rasch methodology. The five-factor model was a significantly better fit for the response data than the one-factor model. The reported correlations between the factors were of the same magnitude as those reported by the multidimensional Rasch model indicating that the
NOS construct is composed of five inter-related dimensions. In addition, the literature review produced one CFA study whose results buttressed the proposition that the NOS construct is multidimensional and composed of inter-related dimensions.

With both methodologies providing comparable evidence in support of a NOS construct composed of five inter-related dimensions, the scores from the multidimensional Rasch model were investigated further to determine if they were interpretable and suitable for their intended purpose, namely to be used in science education research and teaching. Discrepant case analyses modeled on Briggs and Wilson (2003) and Allen and Wilson’s (2006) research revealed the multidimensional subscale scores differed substantially from the unidimensional composite score which indicates that students’ abilities varied across dimensions and any one of these estimates could not reliably used as a measure of a students’ NOS ability. This claim was reinforced by the portrayal of two students’ score profiles across the five dimensions. These students had the same scores on the unidimensional model but their scores differed considerably across the five dimensions. Therefore, to obtain a more accurate picture of students’ abilities on their understanding of NOS, scores from each of the five dimensions are needed. These scores are interpretable and “intrinsic to the meaning and outcomes of the testing” (Messick, 1995, p. 748).

Whether scores are interpretable, meaningful and suitable for use in science education research and teaching is central to the extrapolation inference in the chain of inferences. This inference is related to the claim that students’ NOS ability, as measured by the NOSI-E, can be used to account for group-level and/or individual changes in NOS
understanding in the elementary school context. It can similarly be used to explore relationships with other variables that are “useful to identify under applied conditions” (Messick, 1995, 1017). The NRC (2011) in its new framework for K-12 science education explicitly states that it is now possible to “characterize student achievement in terms of multiple aspects of proficiency rather than a single score” (p. 13-6) and to “chart students’ progress over time instead of simply measuring performance at a particular time point” (p. 13-6). If the NOSI-E’s subscales are capable of detecting change in student abilities (i.e., they are responsive and can “chart students’ progress over time”), then support is provided for the external validity aspect of the NOS construct and the use of multiple scores of proficiency rather than one score. The results indicated that the multidimensional subscales were, as measured by the person strata index, reasonably responsive to changes in students’ NOS understanding (the person strata index were all equal to or above 2.5). In contrast, the consecutive subscales were not responsive to change (the person strata index were all equal to or below 2.2) and, as a result, could not be counted on to reliably measure changes in students’ NOS understanding. However, the unidimensional model was responsive to changes in students’ NOS and was the only model that had the capability of categorizing students into three performance levels (naïve, adequate and informed). Given that the preponderance of prior evidence indicates that the NOS construct is multidimensional, future research should target improving the responsiveness and utility of the subscales.

The NRC (2011) also stresses the need for research that examines the complex relationships between scientific classroom practices, learning environments and students’
understanding of core ideas and concepts. To provide evidence in support of this external validity aspect of the NOS construct; two sets of multilevel models (MLMs) were run using the five multidimensional subscale scores as independent variables. The outcome variables for the separate analyses were a measure of student science achievement (CIER) and a measure of students’ perceptions of the constructivist nature of their science classroom learning environment (ESSCES).

The results of the MLMs using scores on the ESSCES as an outcome indicated that students’ NOS understanding across the five NOS dimensions were predictive of students’ perceptions of their learning environment. Evidence from the literature review was used to corroborate this finding; this form of evidence helps ameliorate the concern that these analyses were problematic (due to multicollinearity of NOS predictors) and under-powered. The results of the regression of the NOSI-E subscale scores with the CIER outcome were less conclusive. Student achievement appears related to students’ understanding of the Theory-laden NOS dimension but unrelated to the other four NOS dimensions. However, the analyses were under-powered and evidence from the literature review to uphold this finding was sparse and not directly comparable. Extrapolating score-based interpretations beyond the instrument helps to assess if the scores are useful as Messick (1995) states in their “applied conditions”. The validity argument and associated analyses was designed to provide and build validity evidence that supports its inferences and conclusion; namely the scores produced by the multidimensional model are reliable, interpretable and suitable for use in science education research and teaching.
In summary, it is important to first review the totality of evidence as it relates to the rival approaches to modeling the NOS construct and, inherent in this, the competing structures (unidimensional and consecutive) for the internal model of the NOS construct. The results of the unidimensional Rasch model of the NOS construct indicated that the items were of good technical quality and did not exhibit DIF. In addition, the scores were reliable and the scale would be responsive to changes in students’ NOS understanding. However, the dimensionality analyses revealed that the unidimensional model did not fit the data as well as the multidimensional model; this finding was supported across two different but complementary methodologies (Rasch and CFA). In addition, the interpretability of the scores is called into question as the discrepant case analyses revealed that students’ abilities differ across dimensions and cannot be accurately reflected by one composite score. Therefore, the critical structural validity evidence did not support the unidimensional approach to represent the NOS construct; scores from this model could lead to teachers or researchers misinterpreting students’ NOS abilities.

The other competing model for the internal structure of the NOS construct was the consecutive approach to represent the internal structure. Results from this approach showed that the items were of good technical quality, exhibited no DIF, but the subscale scores were of poor reliability and responsiveness. A lack of reliability and responsiveness indicates that the reproducibility of scores across different test context may not be consistent and the five subscale instruments will not be able to reliably detect changes in students’ understanding of the NOS construct. Most importantly, perhaps, the consecutive model of the internal structure of the NOS construct does not conform to the
theoretical conceptualization of the NOS construct as proposed in the literature (Lederman, 1992; Lederman, Abd-El-Khalick, Bell, & Schwartz, 2002; Lederman, 2007). In comparing the consecutive subscale correlations to those of the multidimensional ones, the consecutive subscale correlations of the five dimensions were attenuated possibly providing an inaccurate picture of the “true” relationship among the dimensions. Therefore, the consecutive representation is not theoretically grounded as it ignores the conceptualization and empirical evidence that the five dimensions are inter-related resulting in a multidimensional construct.

Lastly it is important to summarize the body of evidence that supports the assertion above that the NOS construct is multidimensional and best represented by the multidimensional Rasch model. The results of the multidimensional analyses indicated that the items of the five subscales were of excellent technical quality, exhibited no DIF, had an item hierarchy that conformed to theoretical expectations; and were of reasonable reliability (> 0.7 on each subscale). In the dimensionality analyses, the multidimensional Rasch model was a significantly better fit than the unidimensional model and this was supported by the CFA which indicated that the five-factor model was a significantly better model than the one-factor model. The discrepant case analyses provided evidence that students’ abilities differ across NOS dimensions. Of the three models compared, the multidimensional model is the only model that could enable a teacher, for example, to reliably estimate students’ abilities across the dimensions and use the scores and differential information to tailor their classroom instruction according to their students’ strengths and weaknesses.
Although the reliabilities of the multidimensional subscales were suitable for this type of formative assessment, the reliabilities are too low for high-stakes decisions. Similarly, as they stand, the multidimensional subscales are not sufficiently responsive to measuring change in students’ abilities (all person strata indices were below 3). To address this issue, each subscale requires more difficult items to be developed to increase the reliability, variance and responsiveness of the scale. The burden on the respondents of adding items to the instrument could possibly be ameliorated by simultaneously removing some of the easier items on the subscales. The effects on the subscales of removing items would have to be tested empirically to ensure the integrity and interpretability of the subscales. Given that research into the relationship between classroom learning environments, science achievement and NOS understanding is in its infancy, the positive relationships reported in this study between CIER variable scores and theory-laden scores and between ESSCES variable scores and all five NOS dimension scores suggest that the measures from the multidimensional model are applicable and can be used in science education research. Overall, the preponderance of evidence with regard to the multidimensional model supports the conclusion of the validity argument that the internal structure of the NOS construct is best represented by the multidimensional Rasch model and the scores provided by the model are reliable, interpretable and suitable for use in science education research and teaching.

Study Limitations

One of the major limitations of this research is concerned with the generalizability and research design of the Evolution Readiness (ER) project. A convenience sample was
used; therefore, any findings, inferences and conclusions cannot be generalized to all elementary or, more specifically, to all fourth grade students within the U.S. The data presented in this study are for a sample that is likely not representative of the U.S. population. Given this reality, any decision made about the generalizability and suitability of using the multidimensional Rasch-based scores in science education research is, at best, tentative. The researchers were also not always present for the administration of NOSI-E instrument; therefore, it is not certain if the administration of instrument in the ER classrooms was undertaken in a standardized and consistent manner. A lack of standardization can also affect the generalizability of results. This lack of generalizability of the study’s results is in part due to the research design used for the ER project.

The ER project was based on a cohort research design and of a small scale. This type of research design is observational which limits any causal inferences from being made. For example, in presenting the external validity evidence for the NOS construct, the empirical positive relationships reported in the MLM between the CIER achievement scores and the NOSI-E theory-laden scores cannot be interpreted to mean that improvement in students’ NOS theory-laden understanding will “cause” an increase in students’ achievement. Therefore, the results need to be treated cautiously and understood in the context of the sample used.

The small-scale and limited data collection of the ER project had particular impact on the MLMs performed to provide external validity evidence for the study. The results of the MLM were limited by the lack of power for the analyses. To properly
model the data structure of the ER project, a three-level MLM should have been performed. However, due to power considerations, two sets of two-level MLMs were used. Therefore, the complex error structure may be biased as the student-level errors were correlated by not including cohort as the second level in the models. This may lead to regression coefficients being reported as significant (Type-I error) when, in reality, they are not significant. Similarly, the lack of predictors related to students’ characteristics (e.g., race, socio-economic status) at level-1 of the models and the absence of predictors at the teacher level of the two-level model make the analyses somewhat artificial and not representative of using the NOS scores under “applied conditions” (Messick, 1995, p. 1017). Excluding important explanatory variables at any level may lead to the wrong conclusions on the statistical significance of the variables of interest (i.e., NOS subscale coefficients) as their omission may potentially bias parameter estimates.

The remaining limitation of this study is to do with the NOSI-E instrument itself. There has been a plethora of criticism by science education researchers on the use of self-reported measures of NOS understanding (Lederman & O’Malley, 1990; Lederman, Wade, & Bell, 1998; Lederman, Abd-El-Khalick, Bell, & Schwartz, 2002; Lederman, 2007; Khishfe & Abd-El-Khalick, 2008). There are two main criticisms of self-report measures as they relate to NOS. One is that the instrument developers assume that their conceptual understanding of what an item prompt is measuring is interpreted in the same manner by students. This criticism is not without validity as Lederman and O’Malley’s study (1990) reported that in follow up interviews with students concerning their written
responses to open-ended questions, students had not interpreted the item prompts as expected. Given the relatively small scope of this study, in future research that uses the NOSI-E instrument, it will be important for researchers to ensure that the students are interpreting the items as intended. If the NOSI-E is used in large-scale studies, this could be accomplished by taking a random sample of respondents and interviewing them.

Allaying this criticism of the NOSI-E instrument was this study’s findings that the subscales and items within the subscales conformed to theoretical expectations.

The second reproach is that standardized instruments are only useful for large-scale assessments with the information derived from these studies limited due to the aggregated nature of the data (Lederman, Abd-El-Khalick, Bell, & Schwartz, 2002). Lederman, Abd-El-Khalick, Bell and Schwartz (2002) also criticize instrument developers for assigning a numerical value to students’ NOS understanding and for not explaining what “numerical value on such instruments constitutes an adequate view of NOS” (Lederman, Abd-El-Khalick, Bell, & Schwartz, 2002, p. 503). These criticisms are at the crux of the consequential validity aspect of construct validity. This study was limited as it was not able to investigate the consequential validity aspect of NOS construct validity and, as a result, falls prey to Lederman, Abd-El-Khalick, Bell and Schwartz’s criticism. Consequential validity is concerned with “value implications of score interpretation as a source of action” (Wolfe, 2007b, p. 224). The development of the NOSI-E instrument in its current form took three years to build. With some modifications (discussed in the next section), the instrument is now ready to be used in studies where standards of performance could be set. The scale metric axis of the variable
map with the student distribution and item calibrations on the common metric can be used by researchers to establish cutoffs which could be used to indicate what constitutes naïve, adequate and informed conceptions of NOS across the five dimensions of the construct (criterion-based standards) or it could equally be used to establish norms for elementary students (normative-based standards).

The Rasch-based results provide researchers and practitioners with a rich qualitative assessment (item content) of what these performance levels actually mean in terms of student understanding; that is, the qualitative meanings can be assigned to the quantitative measures (Wolfe, 2007b; Boone, Townsend, & Staver, 2011). Rasch-based methodologies also enable practitioners to individualize this qualitatively rich source of information by the production of “construct maps” for each student on each variable (Kennedy & Draney, 2009) enabling teachers to tailor instruction to individual students’ strengths and weaknesses. The NRC (2011) calls for high quality assessment tools that can elicit “appropriate and relevant data from students” (p. 13-6) that can be interpreted to help students learn and is of practical use in the classroom. Although the NOSI-E was designed for large scale use, it has applications within the classroom and could prove, with modification, a useful tool to help understand students’ understanding of the nature of science.

**Recommendations for Future Research**

The recommendations for future research related to the NOSI-E instrument can be divided into two areas: (1) making improvements to the actual instrument and (2) its application in research that seeks to understand the complex relationship between
students’ NOS understanding, students’ achievement and students’ science learning environments.

**Instrument Improvements.** The results indicated that many of the items of the NOSI-E were too easily endorsed for some students with the placement of these students’ abilities on the scale metric axis extending beyond the item calibrations of each NOS dimension. If the purpose of using the NOSI-E was to measure growth in NOS understanding (e.g., resulting from an intervention), this mis-match of student ability estimates and item calibrations impacts the utility of the instrument as it makes measuring change problematic due to ceiling effects. If future administrations of the instrument confirm that items and person distributions are mis-matched, research will be needed to develop new, developmentally appropriate items that have the goal of increasing the overall difficulty of each subscale. Suggestions for the concepts for this item development are to measure students’ ability to understand: (1) the distinction between observation and inference (EMP); (2) the role of cause and effect in science (EMP); (3) the role of indirect evidence in making scientific claims (THL); (4) the distinction between a hypothesis and a theory (THL); (5) that scientific knowledge can be discarded in the future (CER); (6) the role of imagination in assessing indirect evidence (INV); (7) that science is equally collaborative and argumentative (SCE); and (8) that science is socially negotiated among peers (SCE). Items based on these concepts should be theoretically more difficult than those currently developed for the NOSI-E and help improve the variance of each subscale. In addition, these types of items will help in the
development of instruments targeted at older students which is another avenue for future instrument development research.

The possibility exists of developing instruments for middle and high school students (NOSI-M and NOSI-H), respectively. In this manner, the progression in student NOS understanding can be assessed as they grow and mature in their conceptions of how scientific knowledge is constructed. At the high school level, even more difficult concepts (e.g., understanding what a scientific law is and how it is different from a theory) could be added to ensure the full breadth of the NOS latent construct is measured. This would aid in the NRC’s (2011) goal of assessing learning progressions for students’ understanding of how scientific knowledge is constructed.

Another productive avenue for research as it relates to the NOSI-E instrument is to establish standards (criterion referenced and norm referenced). In this manner, these standards can be used by practitioners to understand how their students are performing compared to either a criterion or compared to their peers. Related to this research is the development of visuals such as construct maps that can be productively used by teachers to help them and their students understand their strengths and weaknesses on the concepts related to NOS understanding. All these future instrument development activities are proposed to help make the NOSI-E more useful for science education researchers and teachers.

**Use of NOSI-E in Science Education Research.** Akerson, Buck, Donnelly, Nargund-Joshi and Weiland (2011) recommend that future research should help develop a better understanding of how different teaching practices influence student NOS
understanding and whether an effective learning progression can be identified if students are exposed to NOS instruction (integrated and non-integrated) across disciplines and throughout their education K-12. This view is supported by the NRC (2011) in their new framework for science education K-12. The results from this study support Akerson, Buck, Donnelly, Nargund-Joshi and Weiland’s (2011) assertion of the need for future research. This study’s results indicate the complexity of the relationships between student NOS understanding and student achievement; and between NOS understanding and students’ classroom learning environments but also the dearth of research that examines these types of relationships in science education. Complicating this issue is the acknowledgement of science education researchers that “current science education does not typically offer the kind of educational environments that have been shown to support children’s understanding of scientific knowledge” and result in their view that “scientific knowledge is unproblematic” (NRC, 2007, p. 182). Therefore, it is important that future research studies try to disentangle these relationships and endeavor to understand what practices and classroom learning environments promote students’ NOS understanding and whether this in turn promotes students’ ability to understand core ideas and concepts in science.

The rich, dynamic and fundamental relationships between NOS understanding, classroom environments and student achievement, however, will only truly be understood if future research uses more rigorous and powerful research methodologies. These studies should be of a larger scale than the scale used in this study so there is sufficient power to detect effects and model data structures appropriately. To measure learning progressions
in students’ understanding of NOS, studies could be designed to assess individual-level longitudinal change. As the calibrations of the NOSI-E items should remain invariant over time and across samples, the instrument has the potential to measure this type of change in learning as students’ progress K-12. In addition, if the research for example is related to the use of an intervention to improve NOS understanding, the research should use a robust experimental design (e.g., random assignment of treatment to subjects) so that the researchers can postulate and measure causal relationships. These relationships will not, however, be truly understood unless the instruments used to measure these variables are psychometrically sound and reflect the “true” internal structure of the construct they are designed to measure.

Conclusion

The purpose of this study was to determine which of three competing models will provide, reliable, interpretable, and responsive measures of elementary students’ understanding of the nature of science (NOS). The Nature of Science Instrument–Elementary (NOSI-E), a 28-item Rasch-based instrument, was used to assess students’ NOS understanding. Corollary to this purpose was to provide evidence that the scores from the NOSI-E instrument could be used in science education research and has applications within the elementary science classroom.

Specifically, this research focused on the structural validity aspect of the NOS construct as measured by the Nature of Science Instrument – Elementary (NOSI-E). The encompassing research question for this dissertation asked which of three representations of the NOS construct (unidimensional, consecutive or multidimensional) best represents
the internal structure of the NOS construct and will provide researchers with scores that are reliable, interpretable and responsive for use in their research. Grounded by Lederman’s (Lederman, 1992; Lederman, Abd-El-Khalick, Bell, & Schwartz, 2002; Lederman, 2007) theoretical conceptualization of the NOS construct, it was hypothesized that the multidimensional Rasch model would best represent the internal structure of the NOS construct. The multidimensional conceptualization posits that the NOS construct is composed of five inter-related but separate dimensions.

To address this research question and its related hypothesis Messick’s (1995) unified concept for validity was used to guide the analyses for this research. A validity argument was developed that hypothesized that the internal structure of the NOS construct is best represented and measured by the multidimensional Rasch model. Four sets of analyses were performed in which the three representations were compared. These analyses addressed five validity aspects (content, substantive, generalizability, structural and external) of construct validity. These analyses primarily used evidence from Rasch methodology to support the validity argument. However, classical test theory methodology (Confirmatory Factor Analyses; CFA) was also used to substantiate the Rasch-based findings.

The study found that the vast body of evidence supported the claim that the NOS construct is composed of five separate but inter-related dimensions that is best represented by the multidimensional Rasch model. The results of the multidimensional analyses indicated that the items of the five subscales were of excellent technical quality, exhibited no differential item functioning (based on gender), had an item hierarchy that
conformed to theoretical expectations; and together formed subscales of reasonable reliability (> 0.7 on each subscale) that were responsive to change in the construct. In contrast, the consecutive subscales were all of poor reliability; lacked responsiveness and the model ignored the interrelationship of the subscales that was determined empirically. In the dimensionality analyses, the multidimensional Rasch model was a significantly better fit than the unidimensional model and this was supported by the CFA.

Theory-laden scores from the multidimensional model predicted students’ science achievement with scores from all five dimensions significantly predicting students’ perceptions of the constructivist nature of their classroom learning environment. By providing scores on each dimension, the multidimensional model can provide teachers with reliable estimates of students’ abilities across the dimensions enabling teachers to use this differential information to tailor their classroom instruction according to their students’ strengths and weaknesses. The NOSI-E instrument is a theoretically grounded scale that can measure elementary students’ NOS understanding and appears suitable for use in wider applications within science education research.

The facility to measure progression in students’ learning is fundamental to the NRC’s (2011) goal of having children build on their knowledge and understanding of science so they can develop informed and coherent views of the scientific enterprise and how scientific knowledge is constructed. As US science education moves toward students learning science through engaging in authentic scientific practices and building learning progressions in science (NRC, 2011), it will be important to assess whether this new approach to teaching science is effective. The NOSI-E, designed to assess student
understanding of NOS, can be used as one measure of whether this reform effort is impactful.
References


Şahin, C. T. & Köksal, M. S. (2010). How are the perceptions of high school students and teachers on NOS as a knowledge type presented in schools in terms of “importance” and “interest”. *International Journal of Science Education, 5* (1), 105-126.


Scientific Literacy Research Center, the University of Wisconsin. “*Wisconsin Inventory of Science Processes.*” Madison, Wisconsin: The University of Wisconsin, 1967.


Tsai, C. (2006). Biological knowledge is more tentative than physics knowledge: Taiwan high school adolescents’ views about the nature of biology and physics. Adolescence, 41 (164), 691-703.


## Appendix 1
### Item Prompts, Anchoring Items and Development Cycle for NOSI-E

<table>
<thead>
<tr>
<th>Item</th>
<th>Item Prompt</th>
<th>Item Diff.</th>
<th>Devn. Cycle</th>
</tr>
</thead>
<tbody>
<tr>
<td>CER14A</td>
<td>In science, finding an out that a hypothesis is NOT correct is as important as finding out that a hypothesis IS correct.</td>
<td>545</td>
<td>3</td>
</tr>
<tr>
<td>CER8M</td>
<td>New theories in science should only be accepted when there is a lot of evidence to support them.</td>
<td>523</td>
<td>2</td>
</tr>
<tr>
<td>CER6H</td>
<td>A lot of data is needed to decide if a hypothesis is true.</td>
<td>471</td>
<td>P ; 1</td>
</tr>
<tr>
<td>CER8L</td>
<td>When scientists have a good idea, they continue to try to make it better.</td>
<td>464</td>
<td>2</td>
</tr>
<tr>
<td>CER7H</td>
<td>Two scientists can disagree, but both can have good ideas.</td>
<td>453</td>
<td>P ; 1</td>
</tr>
<tr>
<td>CER6J</td>
<td>Trying things out helps scientists think of new ideas.</td>
<td>446</td>
<td>P ; 1</td>
</tr>
<tr>
<td>INV6D</td>
<td>You have to be creative to work in science.</td>
<td>556</td>
<td>P ; 1</td>
</tr>
<tr>
<td>INV12N</td>
<td>A good imagination is needed to create the best experiment to test an idea.</td>
<td>534</td>
<td>2</td>
</tr>
<tr>
<td>INV8G</td>
<td>To explain their results, scientists need to be creative.</td>
<td>534</td>
<td>P ; 1</td>
</tr>
<tr>
<td>INV12L</td>
<td>Although science is based on facts, scientists do need a good imagination.</td>
<td>515</td>
<td>2</td>
</tr>
<tr>
<td>INV12M</td>
<td>A good imagination is needed to make predictions about what will happen in an experiment.</td>
<td>505</td>
<td>2</td>
</tr>
<tr>
<td>SCE13K</td>
<td>Where scientists live may affect what they are allowed to work on.</td>
<td>569</td>
<td>2</td>
</tr>
<tr>
<td>SCE13C</td>
<td>The country a scientist comes from influences how they understand the results of an experiment.</td>
<td>558</td>
<td>P ; 1</td>
</tr>
<tr>
<td>SCE13I</td>
<td>How scientists see the world is influenced by the culture they grew up in.</td>
<td>553</td>
<td>P ; 1</td>
</tr>
<tr>
<td>SCE13H</td>
<td>Where scientists come from may lead to different answers to the same question.</td>
<td>526</td>
<td>P ; 1</td>
</tr>
<tr>
<td>SCE13F</td>
<td>A scientist’s beliefs may change how they do their work.</td>
<td>521</td>
<td>P ; 1</td>
</tr>
<tr>
<td>EMP9I</td>
<td>Experiments are used to see what happens in nature.</td>
<td>516</td>
<td>P ; 1</td>
</tr>
<tr>
<td>EMP14B</td>
<td>Scientists infer what they think is happening from what they already know.</td>
<td>512</td>
<td>3</td>
</tr>
<tr>
<td>EMP8I</td>
<td>Science describes what happens in nature.</td>
<td>512</td>
<td>P ; 1</td>
</tr>
<tr>
<td>EMP8D</td>
<td>A good way to know if something is true is to do an experiment.</td>
<td>486</td>
<td>P ; 1</td>
</tr>
<tr>
<td>EMP9J</td>
<td>Science helps answer questions about how something works.</td>
<td>465</td>
<td>P ; 1</td>
</tr>
<tr>
<td>EMP9A</td>
<td>Scientists explain how something works.</td>
<td>454</td>
<td>P ; 1</td>
</tr>
<tr>
<td>THL14E</td>
<td>If we do the same experiments many times, we may get different results.</td>
<td>527</td>
<td>3</td>
</tr>
<tr>
<td>THL8F</td>
<td>Scientific questions are answered by observing things.</td>
<td>504</td>
<td>P ; 1</td>
</tr>
<tr>
<td>THL9K</td>
<td>Theories can change when new evidence is found.</td>
<td>483</td>
<td>2</td>
</tr>
<tr>
<td>THL8K</td>
<td>Scientists use what they found in the past to help explain their new findings.</td>
<td>477</td>
<td>2</td>
</tr>
<tr>
<td>THL12K</td>
<td>Scientists create different types of experiments to answer their questions.</td>
<td>447</td>
<td>2</td>
</tr>
<tr>
<td>THL6C</td>
<td>Scientists use different ways to test their hypotheses.</td>
<td>435</td>
<td>P ; 1</td>
</tr>
</tbody>
</table>

Notes: 1 Items are based on Year 2 and Year 3 data combined Source: ER Project; Items used for Anchoring Scale denoted with an “A”. CER: Certainty; INV: Inventive; SCE: Socially and Culturally Embedded; EMP: Empirical; and THL: Theory-Laden Subscale. EMP8D highlighted in grey is from the following source: Conley, Pintrich, Vekiri, & Harrison (2004); THL8F highlighted in grey is from the following source: Moore & Foy (1997).
Appendix 2
Syntax for Rasch Models (Conquest Software\(^1\))

**Unidimensional Rasch Model**
Title Partial Credit Model: YR123_NOS_IT28;
data YR123.NOS.IT28.UNI2.MISFIT.dat;
format responses 1-28 name 29-34 state 35 gender 37;
labels << YR123.NOS.IT28.lab;
codes 3,2,1,0;
score (3 2 1 0) (3 2 1 0);
model item + item*step;
set constraints=items;
import init_parameters << YR123.NOS.IT28.IP.prm;
estimate;
show !estimates=latent >> YR123.NOS.IT28.MISFIT.shw;
itanal >> YR123.NOS.IT28.MISFIT.itn;
show cases !estimates=mle >> YR123.NOS.IT28.MISFIT.mle;
plot tcc;
plot tinfo;
plot icc;

**Consecutive Rasch Model: Empirical Subscale**
Title Partial Credit Model: YR123_NOS_IT28_EMP;
data YR123.NOS.IT28.UNI2.MISFIT.EMP.dat;
format responses 1-6 name 7-12 gender 13 ELL 15 YEAR 17 state 21;
labels << YR123.NOS.IT28.EMP.lab;
codes 3,2,1,0;
score (3 2 1 0) (3 2 1 0);
model item + item*step;
set constraints=items;
estimate;
show !estimates=latent >> YR123.NOS.IT28.EMP9.12.shw;
itanal >> YR123.NOS.IT28.EMP9.12.itn;
show cases !estimates=mle >> YR123.NOS.IT28.EMP9.12.mle;
plot tcc;
plot tinfo;
plot icc;
Appendix 2 (continued)

**Multidimensional Rasch Model**
Title Partial Credit Model: YR23.NOS.IT28.MULT;
data YR123.NOS.IT28.MULT2.MISFIT.dat;
format responses 1-28 name 29-34 gender 35 ELL 37 Year 39-43 state 44;
labels << YR123.NOS.IT28.MULT.lab;
codes 0,1,2,3;
score (0,1,2,3) (0,1,2,3) ( ) ( ) ( ) ! items (1-6);
score (0,1,2,3) ( ) (0,1,2,3) ( ) ( ) ( ) ! items (7-11);
score (0,1,2,3) ( ) (0,1,2,3) ( ) ( ) ( ) ! items (12-17);
score (0,1,2,3) ( ) (0,1,2,3) ( ) ( ) ( ) ! items (18-23);
score (0,1,2,3) ( ) ( ) (0,1,2,3) ( ) ( ) ! items (24-28);
model item + item*step;
set warnings=no,update=yes;
set respmiss=both;
export parameters >> YR123.NOS.IT28.MULT.MISFIT3.prm;
export reg_coefficients >> YR123.NOS.IT28.MULT.MISFIT3.reg;
export covariance >> YR123.NOS.IT28.MULT.MISFIT3.cov;
import init_parameters << YR123.NOS.IT28.MULT001.prm;
import init_covariance << YR123.NOS.IT28.MULT001.cov;
import init_reg_coefficients << YR123.NOS.IT28.MULT001.reg;
estimate!method=montecarlo,nodes=2000,converge=.005;
show !tables=1:2:3:4,estimates=latent >> YR123.NOS.IT28.MULT9.15.shw;
itanal >> YR123.NOS.IT28.MULT9.15.itn;
show cases !estimates=mle >> YR123.NOS.IT28.MULT9.15.mle;
quit;

**DIF Analyses Gender for Unidimensional Model**
Title Partial Credit Model: YR123_NOS_IT28;
data YR123.NOS.IT28.UNI2.MISFIT.dat;
format responses 1-28 name 30-34 gender 36 state 45;
labels << YR123.NOS.IT28.lab;
codes 3,2,1,0;
score (3 2 1 0) (3 2 1 0);
model item-gender+item*gender;
set constraints=items;
import init_parameters << YR123.NOS.IT28.IP.prm;
estimate !fit=no, stderr=full;
show !tables=2 >> YR123.NOS.IT28.UNI.DIF.9.15.shw;
plot icc! overlay=yes, legend=yes;

2. Syntax that is bolded was added to consecutive and multidimensional files above to obtain DIF analyses.
Appendix 3

Syntax for Confirmatory Factor Analyses (MPLUS Software)

CFA One-Factor Unidimensional Model
TITLE: YR123_NOS_IT28_UNIDIMENSIONAL;
DATA: FILE IS Yr123.NOS.IT28.SPSS.dat;
VARIABLE: NAMES ARE EMP8D EMP8I EMP9A EMP9I EMP9J EMP14B INV6D INV8G
INV12L INV12M INV12N THL6C THL8F THL8K THL9K THL12K THL14E CER6H CER6J
CER7H CER8L CER8M CER14A SCE13C SCE13F SCE13H SCE13I SCE13K;
USEVARIABLES are EMP8D EMP8I EMP9A EMP9I EMP9J EMP14B INV6D INV8G
INV12L INV12M INV12N THL6C THL8F THL8K THL9K THL12K THL14E CER6H CER6J
CER7H CER8L CER8M CER14A SCE13C SCE13F SCE13H SCE13I SCE13K;
CATEGORICAL ARE EMP8D EMP8I EMP9A EMP9I EMP9J EMP14B INV6D INV8G
INV12L INV12M INV12N THL6C THL8F THL8K THL9K THL12K THL14E CER6H CER6J
CER7H CER8L CER8M CER14A SCE13C SCE13F SCE13H SCE13I SCE13K;
MISSING ARE ALL(9);
ANALYSIS:
  TYPE = general;
DATA IMPUTATION:
MODEL: f1 BY EMP8D - SCE13K;
Output:
sampstat standardized modindices (3.84);

CFA Consecutive Factor Model: Empirical
TITLE: YR123.NOS.IT28.EMP;
DATA: FILE IS Yr123.NOS.IT28.SPSS.dat;
VARIABLE: NAMES ARE EMP8D EMP8I EMP9A EMP9I EMP9J EMP14B INV6D INV8G
INV12L INV12M INV12N THL6C THL8F THL8K THL9K THL12K THL14E CER6H CER6J
CER7H CER8L CER8M CER14A SCE13C SCE13F SCE13H SCE13I SCE13K;
USEVARIABLES are EMP8D EMP8I EMP9A EMP9I EMP9J EMP14B;
CATEGORICAL ARE EMP8D EMP8I EMP9A EMP9I EMP9J EMP14B;
MISSING ARE ALL(9);
ANALYSIS:
  TYPE = general;
DATA IMPUTATION:
MODEL: f1 BY EMP8D-EMP14B;
Output:
sampstat standardized modindices (ALL);
Appendix 3 (continued)

CFA Five-Factor Model (multidimensional)
TITLE: Yr123_NOS_IT28 Five-Factor;
DATA: FILE IS Yr123.NOS.IT28.SPSS.dat;
VARIABLE: NAMES ARE EMP8D EMP8I EMP9A EMP9I EMP9J EMP14B INV6D INV8G
INV12L INV12M INV12N THL6C THL8F THL8K THL9K THL12K THL14E CER6H CER6J
CER7H CER8L CER8M CER14A SCE13C SCE13F SCE13H SCE13I SCE13K;
USEVARIABLES are EMP8D EMP8I EMP9A EMP9I EMP9J EMP14B INV6D INV8G
INV12L INV12M INV12N THL6C THL8F THL8K THL9K THL12K THL14E CER6H CER6J
CER7H CER8L CER8M CER14A SCE13C SCE13F SCE13H SCE13I SCE13K;
CATEGORICAL ARE EMP8D EMP8I EMP9A EMP9I EMP9J EMP14B INV6D INV8G
INV12L INV12M INV12N THL6C THL8F THL8K THL9K THL12K THL14E CER6H CER6J
CER7H CER8L CER8M CER14A SCE13C SCE13F SCE13H SCE13I SCE13K;
MISSING ARE ALL(9);
ANALYSIS:
   TYPE = general;
DATA IMPUTATION:
MODEL: F1 BY EMP8D - EMP14B;
   F2 BY THL6C - THL14E;
   F3 BY INV6D - INV12N;
   F4 BY CER6H - CER14A;
   F5 BY SCE13C - SCE13K;
   F1 F2 WITH F3 F4 F5;
Output:
sampstat standardized modindices (3.84);
Appendix 3 (continued)

DIFFTEST MODEL COMPARISON
TITLE: Yr123_NOS_IT28_DIFFTEST.H1MODEL;
DATA: FILE IS Yr123.NOS.IT28.SPSS.dat;
VARIABLE: NAMES ARE EMP8D EMP8I EMP9A EMP9I EMP14B INV6D INV8G
INV12L INV12M INV12N THL6C THL8F THL8K THL9K THL12K THL14E CER6H CER6J
CER7H CER8L CER8M CER14A SCE13C SCE13F SCE13H SCE13I SCE13K;
USEVARIABLES are EMP8D EMP8I EMP9A EMP9I EMP14B INV6D INV8G
INV12L INV12M INV12N THL6C THL8F THL8K THL9K THL12K THL14E CER6H CER6J
CER7H CER8L CER8M CER14A SCE13C SCE13F SCE13H SCE13I SCE13K;
CATEGORICAL ARE EMP8D EMP8I EMP9A EMP9I EMP14B INV6D INV8G
INV12L INV12M INV12N THL6C THL8F THL8K THL9K THL12K THL14E CER6H CER6J
CER7H CER8L CER8M CER14A SCE13C SCE13F SCE13H SCE13I SCE13K;
MISSING ARE ALL(9);
ANALYSIS:
TYPE = general;
SAVEDATA:
DIFFTEST IS H1MOD.dat;
MODEL: f1 BY EMP8D - EMP14B;
  f2 BY THL6C - THL14E;
  F3 BY INV6D - INV12N;
  F4 BY CER6H - CER14A;
  F5 BY SCE13C - SCE13K;
  F1 F2 WITH F3 F4 F5;
Output:
  sampstat standardized;