Numeracy
Advancing Education in Quantitative Literacy

Volume 8 | Issue 2 Article 4

2015

Quantitative Reasoning in Environmental Science: Rasch Measurement to Support QR Assessment

Robert L. Mayes
Georgia Southern University, rmayes@georgiasouthern.edu
Kent Rittschof
Georgia Southern University, kent_r@georgiasouthern.edu
Jennifer H. Forrester
University of Wyoming, jforres5@uwyo.edu
Jennifer D. Schuttlefield Christus
University of Wisconsin Oskhosh, schuttlj@uwosh.edu
Lisa Watson
Georgia Southern University, lw02732@georgiasouthern.edu

See next page for additional authors

Recommended Citation

Mayes, Robert L., Kent Rittschof, Jennifer H. Forrester, Jennifer D. Schuttlefield Christus, Lisa Watson, and Franziska Peterson. "Quantitative Reasoning in Environmental Science: Rasch Measurement to Support QR Assessment." *Numeracy* 8, Iss. 2 (2015): Article 4. DOI: http://dx.doi.org/10.5038/1936-4660.8.2.4

Quantitative Reasoning in Environmental Science: Rasch Measurement to Support QR Assessment

Abstract

The ability of middle and high school students to reason quantitatively within the context of environmental science was investigated. A quantitative reasoning (QR) learning progression, with associated QR assessments in the content areas of biodiversity, water, and carbon, was developed based on three QR progress variables: quantification act, quantitative interpretation, and quantitative modeling. Diagnostic instruments were developed specifically for the progress variable quantitative interpretation (QI), each consisting of 96 Likert-scale items. Each content version of the instrument focused on three scale levels (macro scale, micro scale, and landscape scale) and four elements of QI identified in prior research (trend, translation, prediction, and revision). The QI assessments were completed by 362, 6th to 12th grade students in three U.S. states. Rasch (1960/1980) measurement was used to determine item and person measures for the QI instruments, both to examine validity and reliability characteristics of the instrument administration and inform the evolution of the learning progression. Rasch methods allowed identification of several QI instrument revisions, including modification of specific items, reducing number of items to avoid cognitive fatigue, reconsidering proposed item difficulty levels, and reducing Likert scale to 4 levels. Rasch diagnostics also indicated favorable levels of instrument reliability and appropriate targeting of item abilities to student abilities for the majority of participants. A revised QI instrument is available for STEM researchers and educators.

Keywords

quantitative reasoning, quantitative interpretation, modeling, learning progression, environmnetal science education

Creative Commons License

Cover Page Footnote

Robert Mayes is a Research Professor in Education at Georgia Southern University, United States, and Director of the Institute for Interdisciplinary STEM Education (i2 STEMe). His research focus is on systems reasoning, computational reasoning, and quantitative reasoning in STEM and development of reasoning learning progressions within the context of science. His work supports interdisciplinary STEM teaching that impacts underrepresented populations and rural areas.
Kent Rittschof is a Professor of Educational Psychology at Georgia Southern University, and Chair of the Department of Curriculum, Foundations, & Reading. His expertise is in learning and cognition. His recent research has emphasized contemporary measurement approaches for examining student attitudes and cognitive abilities. He was the lead on conducting Rasch analysis in this study.
Jennifer Harris Forrester is an Assistant Professor in Elementary and Early Childhood Education at the University of Wyoming. Her research focuses on the use of quantitative reasoning skills in different science contents and contexts. She is interested in documenting how professional scientists use QR skills in their research (specifically field ecology) and how they teach students QR within the context of fieldwork. Her work supports best practices in STEM teaching at the K-12 and undergraduate level.
Jennifer Schuttlefield Christus is an Assistant Professor of Chemistry at the University of Wisconsin Oshkosh. Her current research focuses on semiconductor photoelectrochemistry, solar energy conversion, atmospheric reactions on clay mineral aerosols and chemical and science education

including quantitative reasoning and the development of learning progressions in the context of science.
Lisa Watson is a graduate student at Georgia Southern University. She is working towards her doctorate in Clinical Psychology. She served as the research graduate student on the study.
Franziska Peterson is a graduate student at the University of Wyoming. She is working towards her Ph.D. in Mathematics Education. She has served as the research graduate student on the Pathways Project over the past four years.

Authors

Robert L. Mayes, Kent Rittschof, Jennifer H. Forrester, Jennifer D. Schuttlefield Christus, Lisa Watson, and Franziska Peterson

Introduction

The Next Generation Science Standards (NGSS 2013) and the Common Core State Standards for Mathematics (NGAC 2010) call for improving scientific, engineering, and mathematical practices. Among the practices called for are model-based reasoning which engages students in developing and using models, analyzing and interpreting data, and using mathematics and computational thinking. Fundamental to these processes is quantitative reasoning (QR), which for this project is defined as:

> Quantitative reasoning is mathematics and statistics applied in real-life, authentic situations that impact an individual's life as a constructive, concerned, and reflective citizen (Mayes et al. 2014a).

In the NSF project, *Culturally Relevant Ecology, Learning Progressions, and Environmental Literacy* [1] (or simply the *Pathways project*) a QR learning progression was developed to explore the trajectory of QR development across sixth to twelfth grades. A learning progression is a set of empirically grounded and testable hypotheses about how students' understanding of, and ability to use, core scientific concepts, explanations, and related scientific practices grow and become more sophisticated over time with appropriate instruction (Corcoran et al. 2009). Learning progressions provide levels of understanding through which students develop mastery of a concept over an extended period of time. The QR learning progression is conceptualized as having four levels: the lower anchor, upper anchor and two intermediate levels of understanding. The lower anchor is grounded in data collected on sixth graders understanding of QR (Mayes et al. 2014a). The upper anchor is based on expert views of what a scientifically literate citizen who is well versed in QR should know and be able to apply by the twelfth grade. A learning progression defines progress variables which are essential categories for the overall concept across which the levels are established. The QR progress variables for the QR learning progression are:

- Quantification Act (QA): mathematical process of conceptualizing an object and an attribute of it so that the attribute has a unit measure. Included in QA is quantitative literacy (the use of fundamental mathematical concepts in sophisticated ways) which allows one to describe, compare, manipulate, and draw conclusions from the quantified variables.
- Quantitative Interpretation (QI): ability to use models to discover trends and make predictions.
- Quantitative Modeling (QM): ability to create representations to explain phenomenon and to revise them based on fit to reality.

[1] Award #0832173, Division of Research on Learning, MSP-Targeted Awards. PI: John Moore, Colorado State University

Finally, each of the progress variables were elucidated by identifying a collection of elements determined through student interviews which indicate essential abilities within the categories:

- Quantification Act Elements: Variation, Quantitative Literacy, Context, Variable.
- Quantitative Interpretation Elements: Trends, Predictions, Translation, Revision.
- Quantitative Modeling Elements: Create model, Refine model, Reason with model, Statistical analysis.

For a detailed presentation of the learning progression see Mayes et al. (2014b).

In the study reported here, Rasch (1960/1980) measurement methods were used to support development of three selected response (or rating scale) assessment instruments (hereafter referred to as "assessments") that can be used to inform the QR progression and provide an efficient and accurate diagnostic assessment of quantitative interpretation (QI). The assessments were designed to be easily implemented within classrooms and to complement other means of assessing and evaluating QI student outcomes by providing an objectively scored alternative that reflects the QR learning progression.

The QI progress variable was selected as the focus for the first QR assessment development. QI was selected due to the central role it plays in developing environmentally literate citizens who can interpret quantitative models and make informed decisions based on them. The elements identified for QI are defined as follows for the upper anchor:

- Trends: determine multiple types of trends including linear, power, and exponential trends; recognize and provide quantitative explanations of trends in model representation within context of problem.
- Translation: translates between models; challenges quantitative variation between models as estimates or due to measurement error; identifies best model representing a context.
- Predictions: makes predictions using covariation and provides a quantitative account which is applied within context of problem.
- Revision: revise models theoretically without data, evaluate competing models for possible combination.

Assessment Development and Implementation

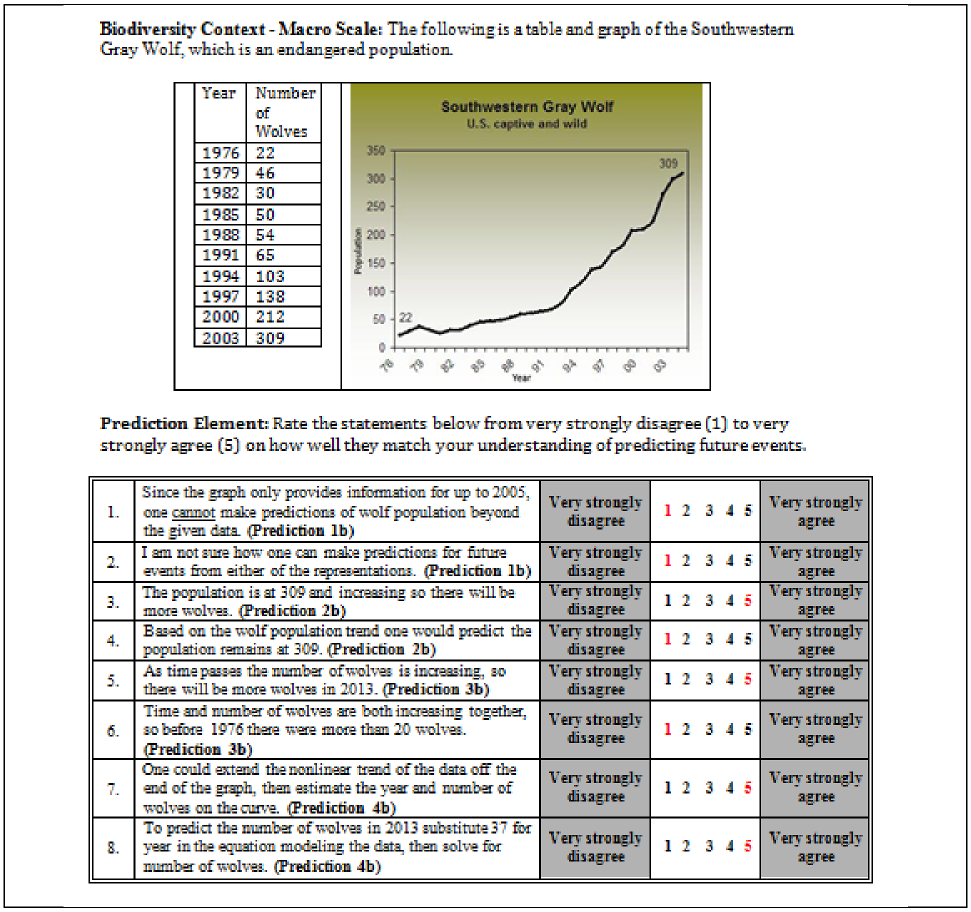
Three parallel assessments were developed to efficiently and accurately assess quantitative interpretation (QI) within the three respective environmental contexts of biodiversity, carbon cycle, and water cycle. Each assessment can be considered a context-specific version of an assessment which enables testing the hypotheses described below. Each of these three assessment versions includes the set of items within one of the three environmental contexts and the administration process used to implement the assessment. The NSF Pathways project identified these three environmental contexts as the essential progress variables for the development of an environmentally literate citizen. The central focus of each assessment was on the QI elements identified above (Mayes et al. 2014a). Each assessment included three scale levels: macro scale (what one can see with their eyes), micro scale (hidden mechanisms that underlie what one sees that require a microscope to view), and landscape scale (larger than what one can see, requiring a telescope or other aid to view). The QR research team viewed scale as a central quantitative issue in science, as the NSF Pathways project identified the concept of scale as a key potential barrier in students developing a deeper understanding of environmental science. The assessments included items developed based on the four learning progression levels: Level 1 (lower anchor - novice), Level 2 (lower intermediate), Level 3 (upper intermediate), and Level 4 (upper anchor - expert). Learning progression theory calls for a limited number of levels, with four to five being common (Corcoran et al. 2009). Two items were written for each of the elements at each of the learning progression levels for each of the scales, giving 32 items per scale and 96 items per assessment (Table 1).

Table 1: Quantitative Interpretation Assessment Structure

Environ Topic	Scale	QI Element	Level (4 per element)	Questions
Biodiversity Carbon Cycle Water Cycle	Macro	Trend	novice, lower intermediate, upper intermediate, expert	1,2
		Translation	novice, lower intermediate, upper intermediate, expert	1,2
		Prediction	novice, lower intermediate, upper intermediate, expert	1,2
		Revision	novice, lower intermediate, upper intermediate, expert	1,2
	Micro	Trend	novice, lower intermediate, upper intermediate, expert	1,2
		Translation	novice, lower intermediate, upper intermediate, expert	1,2
		Prediction	novice, lower intermediate, upper intermediate, expert	1,2
		Revision	novice, lower intermediate, upper intermediate, expert	1,2
	Landscape	Trend	novice, lower intermediate, upper intermediate, expert	1,2
		Translation	novice, lower intermediate, upper intermediate, expert	1,2
		Prediction	novice, lower intermediate, upper intermediate, expert	1,2
		Revision	novice, lower intermediate, upper intermediate, expert	1,2

An example of QI assessment items from the Biodiversity version, macro scale level, prediction element is provided in Figure 1. Each assessment consists of blocks of eight questions per QI element ranging from Level 1 through Level 4 with two questions per level. The five-category Likert scale provided students an opportunity to express their confidence in agreeing with a statement concerning QI.

The assessments were conducted across sixth to twelfth grades, with the levels providing an entry point for students from different grades. The students were provided only one version of the assessment, with one of the three versions being assigned by the teachers to an equal number of students in each participating class. The assessments were administered in Qualtrics so students could take them online. Students were not offered an enticement by the research team for taking the assessment and could choose not to participate. However, students were encouraged by their teachers to take the assessments.



Biodiversity Context - Macro Scale: The following is a table and graph of the Southwestern Gray Wolf, which is an endangered population.

Year	Number of Wolves
1976	22
1979	46
1982	30
1985	50
1988	54
1991	65
1994	103
1997	138
2000	212
2003	309

Southwestern Gray Wolf
U.S. captive and wild
Population
0
50
100
150
200
250
300
350
22
309
76
79
82
85
88
91
94
97
00
03
Year

Prediction Element: Rate the statements below from very strongly disagree (1) to very strongly agree (5) on how well they match your understanding of predicting future events.

1.	Since the graph only provides information for up to 2005, one cannot make predictions of wolf population beyond the given data. **(Prediction 1b)**	Very strongly disagree	1 2 3 4 5	Very strongly agree
2.	I am not sure how one can make predictions for future events from either of the representations. **(Prediction 1b)**	Very strongly disagree	1 2 3 4 5	Very strongly agree
3.	The population is at 309 and increasing so there will be more wolves. **(Prediction 2b)**	Very strongly disagree	1 2 3 4 5	Very strongly agree
4.	Based on the wolf population trend one would predict the population remains at 309. **(Prediction 2b)**	Very strongly disagree	1 2 3 4 5	Very strongly agree
5.	As time passes the number of wolves is increasing, so there will be more wolves in 2013. **(Prediction 3b)**	Very strongly disagree	1 2 3 4 5	Very strongly agree
6.	Time and number of wolves are both increasing together, so before 1976 there were more than 20 wolves. **(Prediction 3b)**	Very strongly disagree	1 2 3 4 5	Very strongly agree
7.	One could extend the nonlinear trend of the data off the end of the graph, then estimate the year and number of wolves on the curve. **(Prediction 4b)**	Very strongly disagree	1 2 3 4 5	Very strongly agree
8.	To predict the number of wolves in 2013 substitute 37 for year in the equation modeling the data, then solve for number of wolves. **(Prediction 4b)**	Very strongly disagree	1 2 3 4 5	Very strongly agree

Figure 1. QI assessment example. Questions are from the QI biodiversity assessment and are at the macro scale for the prediction element. Example includes eight items using the five rating categories. Red coded category labels indicate best responses.

The a-priori hypotheses concerning performance on the QI assessments were:

1. The difficulty level of QI items would vary by item level with the rank from easiest to most difficult being: Lower Level 1 (novice lower anchor), Level 2, Level 3, and Level 4 (expert upper anchor).
2. The difficulty level of QI items would vary by scale with the rank from easiest to most difficult being: macro scale, landscape scale, and micro scale.
3. The difficulty level of items would vary by QR elements with the rank from easiest to most difficult being: trend, translation, prediction, and revision.
4. The three assessment versions measure QI, so across the contexts of water cycle, carbon cycle and biodiversity the student QI outcomes would be similar. More formally each assessment version would reflect a primary QI construct dimension.

The first phases of developing, diagnosing, and refining the assessments are discussed in this paper to provide examples of the process and a description of the application.

Literature Review

Taking Science to School (Duschl et al. 2007) calls for science education to incorporate modeling practices and model-based reasoning. The call is echoed in the *Framework for K-12 Science Education* (National Research Council 2011), the *Next Generation Science Standards* (NGSS 2013), and the *Common Core State Standards* (NGAC 2010). Science as model-building is a fundamental practice of science which includes building models using evidence, checking them for internal consistency and coherence, testing them empirically, and the metaknowledge that guides and motivates the practice (Duschl et al. 2007; Schwarz et al. 2009). Inherent in model building is interpretation of the resulting model.

QI is the ability to analyze a model of a scientific phenomenon (either one provided to or created by the student) to determine trends, to translate between models to compare and contrast them, to revise models to fit new situations, and to make predictions. It is imperative for scientifically literate citizens to be able to interpret and use data provided to them to make decisions -- data that are often represented in a model (table, graph, equation, or science diagram) (Madison and Steen 2003; Steen 2004). “Representations are necessary to students’ understanding of mathematical concepts and relationships” (AERO 2011, p. 13).

Zahner and Corter (2010) propose in their model of probability problem solving that students pass through four stages when problem solving: Stage 1, Text Comprehension; Stage 2, Mathematical Problem Representation; Stage 3, Strategy Formulation and Selection;, and Stage 4, Execution of the Strategy. According to their model, to reach stages 3 and 4, students must pass through stage 2 first. Therefore, the inability to represent a problem and interpret it could be a barrier to student execution of their strategy. QI focuses on interpreting an existing model, such as one found in a newspaper article, but students must still interpret the representation if they are going to apply it to solve the problem. Thus QI could serve as a barrier to problem solving since students would not be able to make an informed decision about the environmental problem being modeled.

A complete review establishing the inclusion of the progress variables in the quantitative reasoning learning progression can be found in Mayes et al. (2012). Here a brief overview of that review is provided that supports the inclusion of the three progress variables in the QR learning progression. First, the inclusion of quantitative act is supported by the work of Thompson (2011). His research presents the quantitative act as an essential first step in moving from the science context to a mathematical representation. He defines quantification as the process of conceptualizing an object and an attribute of it so that the attribute has a unit measure, and the attribute's measure entails a proportional relationship (linear, bi-linear, or multi-linear) with its unit. In addition, covariational reasoning, defined as coordinating two varying quantities while attending to the ways in which they change in relation to each other (Carlson et al 2002), is an important aspect of quantification.

Quantitative literacy was included under the quantitative act progress variable since it is the ability to use fundamental mathematic concepts to manipulate the variables quantified. Quantitative literacy provides the tools to compare, combine, and manipulate the quantities. The work of Steen (2004) and Madison and Steen (2003) establishes that quantitative literacy is essential for all citizens if they are to make data-informed decisions, yet it is often neglected in curriculums due to its interdisciplinary nature.

Second, the inclusion of QI as a progress variable is supported by the work of Schwartz and Martin (2004), who found that early understanding of multiple representations within a context is important for students to progress mathematically. It is essential to their ability to apply models to make informed decisions.

Finally, quantitative modeling was included as a progress variable based on the work of Duschl et al. (2007). They propose a move from science as inquiry to science as model-building and model-refining. Science as model-building is defined as learning science as a process of building theories and models using evidence, checking them for internal consistency and coherence, and testing them empirically (Duschl et al. 2007). The seminal work done by Schwarz et al. (2009) in the Modeling Designs for Learning Science project created a learning progression for scientific modeling which has two dimensions: (1) scientific models as tools for predicting and explaining and (2) models change as understanding improves.

The iterative research process that underpins the development of learning progressions is pivotal to the theoretical framework for our study. *Taking Science to School* (Duschl et al. 2007) recommends that learning and curriculum designs be organized around learning progressions as a means of supporting learners' development. The Consortium for Policy Research in Education report *Learning Progressions in Science: An evidence-based approach to reform* (Corcoran et al. 2009) identified learning progressions as a promising model that can advance effective adaptive instruction teaching techniques and thereby change the norms of practice in schools. A number of learning progressions in science have incorporated components of QR (Louca et al. 2011; Pluta et al. 2011; Schwarz, et al., 2009; Stefani and Tsaparlis 2009; Taylor and Jones 2009; Lehrer and

Schauble 2002, Smith et al. 2006), but the one proposed here is the first progression specifically addressing the development of QR in the sciences.

Purpose and Rationale

The purpose of the current study was to: (1) determine and improve the validity and reliability of a QI diagnostic assessment process using a Rasch (1960/1980) measurement model; and (2) inform the evolution of the current QR learning progression (Mayes et. al. 2014a, b). The Rasch approach was utilized in order to construct additive measures from the data and examine both item statistics and individual student statistics as the QI assessment was revised in support of improvements to the existing QR learning progression (Wilson 2009). The resulting assessments are intended to be used in conjunction with science curricula as a means of efficiently estimating QR development for grades six to twelve.

Methods

Development of the QR learning progression was guided by the iterative research process. First, an intense review of the literature was conducted to establish a hypothetical framework for the progression (Mayes et al. 2013). Second, student interviews were conducted to inform the development of a hypothetical QR learning progression (Mayes et al. 2014a). As stated above, the lower anchor is grounded in QR abilities demonstrated by sixth grade students; the intermediate levels of understanding are the levels through which the students pass on their way to attainment of the upper anchor; and the upper anchor is based on expert views of what QR a scientifically literate citizen should know and be able to apply by the twelfth grade. Here the findings from the third step of the iterative research cycle are reported. QR interviews served as a basis for development of items for a diagnostic assessment which could be implemented online to a large sample of students from grades six to twelve. The diagnostic assessment focused on one component of QR, quantitative interpretation (QI) of scientific models. The diagnostic assessment provided quantitative data informing revision of the QR learning progression and provided baseline data on the current status of QI among middle and high school students. Rasch measurement methods were used to model and analyze both the student outcomes and assessment items simultaneously (Bond and Fox 2007; de Ayala 2011; Engelhard 2013; Linacre 2014).

Rasch Measurement

A contemporary measurement approach, the Rasch (1960/1980) model, was chosen to apply a rigorous scientific framework to the examination and interpretation of the assessment data. The Rasch measurement model was named for the Danish mathematician, Georg Rasch, who originally developed a model for use with dichotomous item data (e.g., correct or incorrect). A Rasch approach assumes a fundamental measurement model, which implies that data should be examined to determine the degree to which an ideal measurement model has been realized (Bond and Fox 2007; Engelhard 2013). A Rasch measurement model was selected for use to study QR because it allows researchers to construct interval measures from ordinal assessment data to allow improved accuracy and use of crucial diagnostics. In addition, a Rasch model is well suited to the investigation because it permits analysis of both the student outcomes and assessment items placed on the same measurement scale frame of reference. These Rasch measures are based on a probabilistic relation between an item's endorsement difficulty and a person's ability (or willingness) to endorse item statements correctly with respect to the construct of interest. This probabilistic relation stems from the common observation that people tend to have a higher probability of correctly responding to easier items and incorrectly answering more difficult items. Resources such as Bond and Fox (2007), de Ayala (2011), Engelhard (2013), Linacre (2014), and Wright and Mok (2004) provide excellent descriptions of the historical and technical developments supporting the growing applications of Rasch methods, including those used for this investigation.

The Rasch model is more accurately a family of modern latent trait models, including one of the members known as the rating scale model developed by Andrich (1978). The rating scale model is a polytomous extension of Georg Rasch's dichotomous model, but modified for data that result from rating scales including Likert instruments with specific numbers of rating categories. Mathematically, the rating scale model describes that the probability of a person correctly responding to an item (P_{nij}) is a logistic function of the relative distance on a linear scale between the respondent measure location (θ_n), the item measure location (b_i), and the 0.5 probability point threshold (τ_j) for choosing between adjacent rating categories of the item

$$\ln\left(\frac{P_{nij}}{1-P_{nij-1}}\right)= \theta_n - b_i - \tau_j$$

where the subscripts refer to the person (*n*), the item (*i*), and the category (*j*). The τ_j threshold is the point at which the probability of opting for one Likert category is equal to that for the prior adjacent category. The formula represents the log of the odds of the correct responding probability (Wright and Mok, 2004).

The resulting transformed values of the ordinal raw scores are considered log-odds units and are referred to as logits. These logits can be seen as units of a Rasch ruler (e.g., Figures 1 through 3 of the Appendix) depicting both item measures and person measures. Graphic depictions of Rasch rulers are commonly referred to as item-person maps, or variable maps. The Rasch rating scale (Andrich 1978) model was used for this investigation in order to construct such linear measures from five ordinal Likert rating categories within the QI assessment. Winsteps (Linacre 2012) and SPSS computer programs were used for Rasch measurement calibrations and the corresponding diagnostic analyses. The following discussions of the use of the Likert scale and an overview of the calibration tools will highlight the specific application of the Rasch model within this investigation.

The Likert-scale assessment items allowed students to choose from a five-category scale. On approximately 60% of the items, category 1 represented very strongly disagree, reflecting the most accurate reasoning, and category 5 represented very strongly agree. The other 40% of items were reversed, meaning category 1 (very strongly disagree) was considered the response reflecting the most accurate reasoning. Prior to Rasch calibration of data, the reversed items were recoded so that for all items category 5 was registered as the best response, with categories 4, 3, 2, and 1 representing respectively lower levels of accurate responding. Thus the minimum raw score on the 96 item assessments was 96 and the maximum raw score was 96 × 5 = 480. The items were written so that they required little to no calculation, with a focus on assessing students' prior knowledge and understanding about using QI in context.

Rasch calibration analyses were conducted to identify needed measurement adjustments. A primary Rasch calibration was run on each of the assessments to identify potential problematic items and students with inconsistent patterns. Calibrations included the use of selected statistical and graphic diagnostic tools discussed by Linacre (2014) and Bond and Fox (2007) to effectively interpret the data in support of decisions regarding strengths, weaknesses, and valid, reliable measurement. These diagnostic categories included (a) *item polarity* (positively or negatively correlated assessment items) to determine whether all items were aligned in the same direction on the latent variable of QI; (b) *category function* (Likert item five-category rating; see item example in Figure 1) to determine whether all categorizations functioned as intended such that the average measures for the categories advanced; (c) *dimensionality* to determine whether all items within the instrument function in unison to represent the same dominant dimension of QI; (d) *item fit* (underfit items are unpredictable, overfit items are too predictable) to determine whether items functioned together to measure in correspondence with the model; (e) *person fit* to determine whether participants responses functioned together to measure in correspondence with the model; (f) *separation* as standard errors of spread existing among persons taking the assessments; (g) *reliability,* which was examined both to determine whether the persons consistently discriminated different levels of ratings (Likert categories of one to five) and whether items discriminate different levels of endorsement difficulty (four levels of learning progression from novice to expert); and (h) *sample targeting* to determine whether the range of item difficulties match well with the participants' responses. *Fit, separation, and reliability* categories are described in further detail below. Each diagnostic category will also be discussed more directly with respect to the findings.

Fit analyses were conducted using the information-weighted fit statistic, or *infit,* and the outlier-sensitive fit statistic, or *outfit,* procedures as part of an examination of the measurement model (Bond and Fox 2007). That is, the items and participants shown to fit the Rasch model can be considered supportive of valid measurement. To diagnose item and person fit the *infit* and *outfit* were used both as mean square statistics (MnSq) and as a standardized conversion (Zstd) of the statistic which provides symmetry and a *t*-test of significance. We reported Zstd statistics for item fit to facilitate comparisons of fit. Criteria of Zstd values above 2.0 and below − 2.0 for identifying potentially misfitting items or participants are commonly used and provided a standard for this investigation as well.

Test reliability was indicated by a Rasch *person reliability index* and the associated *person separation* index. These two indices can be used, as opposed to only one index, in order to enhance our interpretation of our reliability analyses relative to the two respective units represented by these indices. Specifically, person separation reflects the number of standard errors of spread that exist among the persons; it has the advantage over other indices of not being restricted in range between 0 and 1. The higher the person separation, the greater the confidence one can have in person measure order. Person separation index levels greater than 2 represent a typical desired range indicating two distinct groupings of items (e.g., difficult and easy to endorse). Person reliability indices, on the other hand, use the familiar range between 0 and 1, similar to the Cronbach alpha test reliability index, which is calculated using ordinal data. Rasch person reliability is calculated using linear measures and supports the determination of whether items are sufficient for classification of people into groups with respect to their ratings. Person reliability levels of 0.8 or above represent a typical desired level. The two Rasch person indices, separation and reliability, reflect the reproducibility (i.e., likelihood that this result would be repeated) of person ordering one could expect if these same participants were given another similar set of items measuring the diversity attitudes. These indices help us determine whether there are enough items along the measurement continuum at different levels to classify people.

Rasch item indices of separation and reliability were also examined, but in contrast to person separation and reliability these item indices do not reflect test reliability. The item reliability and separation indices reflect the reproducibility of item placements along the continuum if these same items were given to another similar group of participants of the same size that had the same attitudes. These item indices allow determination of whether there are enough participants along the measurement continuum at different levels. Item separation index levels greater than 3 represent a typical desired range and correspond with three distinct groupings of persons (e.g., low, middle, and high). Similarly, item reliability supports determination of whether there is a sufficient sample to classify the items into difficulty groups with levels 0.9 or above representing typical desired levels.

Study Sample and Assessment Administration

The QI assessments were administered to 342 sixth to twelfth grade students in three U.S. states. The sample was 45% male and 55% female; 56% White, 21% African American, 13% Asian, and 6% Native American/Pacific Islander (some participants chose not to disclose their race). The distribution of gender and race across the three assessments and state sites were relatively equivalent. The schools constituted a sample made up of districts that had participated in previous projects with members of the research team.

Teachers in these schools volunteered to administer the assessments in science classes. While students could opt out of taking the assessment, this was a rare occurrence due to the teacher requesting them to complete the assessment. Teachers were instructed to have each student take one version of the assessment and to randomly assign one third of their class to the three assessments. Rasch person fit analysis indicated that 16% of students in the sample had outcomes that were either highly predictable or highly unpredictable (e.g., the student provided contradictory responses to similar items). In addition, some students did not complete the assessments. Removal of these students from the sample can improve interpretation, maximizing measurement accuracy, because the meaning of their mis-fitting data is uncertain with respect to the model. Subsequent Rasch analysis was performed on the remaining student sample of 286 students. The students were distributed by grade as follows: 19 sixth graders, 23 seventh graders, 44 eighth graders, 40 ninth graders, 48 tenth graders, 85 eleventh graders, and 27 twelfth graders.

Results

Item Summary

Reliability Indices for the Assessments. Person reliability levels were 0.86 for the biodiversity assessment, 0.76 for the water assessment, and 0.87 for the carbon assessment. Person separation indicated 2.52 for the biodiversity assessment, 1.77 for the water assessment, and 2.55 for the carbon assessment. Person reliability at or above 0.80 and separation at or above 2.00 standard errors of spread were reached on the biodiversity and carbon assessments, while the levels for the water assessment were just short of those expectations. In general, these reliability and separation levels from the biodiversity and carbon assessment provide support for just over two distinct levels of difficulty (e.g., easier and more difficult) and indicate that items are sufficient for classification of people into groups, which is crucial for estimating QI learning progression levels.

Item reliability levels were 0.85 for the biodiversity assessment, 0.78 for the water assessment, and 0.79 for the carbon assessment. Item separation levels were 2.40 for the biodiversity assessment, 1.87 for the water assessment, and 1.95 for the carbon assessment. With desired levels at 0.90 for item reliability and 3.00 standard errors of spread for item separation, these levels were all below expectations. This finding suggests a need for a larger, more diverse sample of participants to improve measurement.

Fit and Misfit. Fit refers to how well item or person measures correspond to a pattern expected by the Rasch model. Item infit and outfit are summarized in Table 2 for each of the three QI assessment versions. Infit is an information-weighted index so it is most sensitive to the middle of a distribution of measures while outfit is not weighted, allowing it to be more sensitive to outlier measures. Any statistic with an infit or outfit value outside the standardized, or Zstd, interval $(-2, 2)$ was flagged as a concern. High values (underfit) indicate a lack of predictability, or noise, with respect to the model. Underfit can therefore be used as a possible indication of items that are not part of the primary or dominant dimension under investigation (Smith 2004). On the other hand, low values (overfit) indicate very high predictability, which in this context can result from redundancy among items with respect to how students responded. Both maximum and minimum infit and outfit values were identified as areas of concern. Misfit (either underfit or overfit) findings impact our assessment revision concerning how to reduce the number of items without increasing variability in response. While Rasch fit analysis indicated that there were concerns with some items, they will not be automatically removed from future versions of the assessment. The learning progression iterative research process emphasizes improving the items rather than simply removing them from assessments as there are construct-specific reasons for including each item within the assessment.

Table 2: Item Infit and Outfit Summary

	Biodiversity Assessment		**Carbon Assessment**		**Water Assessment**	
	Infit	Outfit	Infit	Outfit	Infit	Outfit
Mean	0.0	0.0	0.0	0.0	0.0	0.0
S.D.	1.5	1.4	1.5	1.5	1.4	1.4
Max	4.5	3.9	5.5	5.6	3.5	3.4
Min	-3.8	-3.9	-2.6	-2.7	-3.7	-3.7

Note. Standardized units (Zstd) were reported for infit and outfit.

Misfit order for items provides infit and outfit parameters for each individual item. For example, on the biodiversity assessment, the item with the highest infit (most noise) is MATD2Q2.[2] For this item, the total score for all persons was 414 of a possible maximum of 495, yielding a measurement of −0.98 logits, placing it as the lowest item on the Rasch ruler (midpoint 0, range −1 to 1). Since the infit Zstd score for this item is 4.5, which is considerably greater than 2, the item does not coincide well with the measurement model since it is not following the pattern of most other items on the assessment.

The numbers of infit and outfit items across all three assessments were similar, except for the water assessment which had only five overfit (highly predictable) item measures compared to ten for the other two assessments. The underfit (most unpredictable) item measures were predominantly macro, trend, and level 1 across all three assessments. This result is surprising, since the macro, trend, and level 1 items were considered by the research team to be the easier items, leading to the assumption that responses would be more predictable. Why were items that the research team considered to be at the more basic scale, element, and question level eliciting the most unpredictable responses? In contrast, the overfit item measures were predominantly micro, translation, and level 4 on the biodiversity assessment, but were more evenly distributed across scale and element for the other two assessments. The carbon assessment had most overfit on level 4 items, while on the water assessment the overfit items were more evenly distributed across the levels. The most surprising outcome for overfit was that the most predictable item responses were most often level 4 items. This could be due to redundancy among items that had similar levels of difficulty with respect to participants' willingness to endorse the item statements.

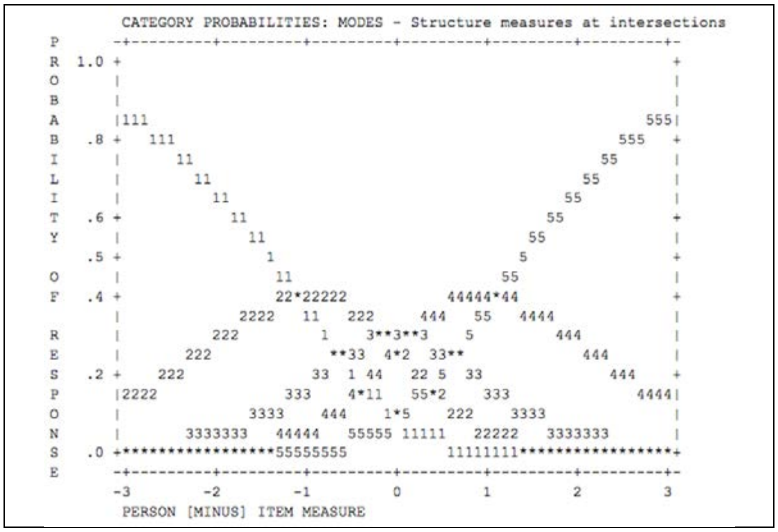
Item Polarity. Rasch analysis provides a point measure correlation for all item measures that reflect item polarity (positively or negatively correlated). Items with a negative polarity, or point measure correlations, indicated that student responses did not trend similarly to most other items. For instance, if an item that required a Likert category 1 rather than a category 5 as the most correct response was not reverse-coded prior to analysis, a negative correlation could be the result for that item. Item LAPR1Q1[3] had the lowest point measure correlation of −0.19, and a measurement of 0.03, which is very near the average item measure. Negative polarity does not always correspond with misfit, and this particular item had an infit Zstd value of 2.0 which we consider just within the model fit range. All such items were reviewed to determine if they should be re-

[2] Figure 1: Macro scale, Trend, Level 2, Question 2 - Given the Grey Wolf data in Figure 1, do you think the population is increasing?

[3] Landscape scale, Prediction, Level 1, Question 1 - One cannot predict future events from a box model of energy flow within a food chain

coded, revised, or potentially removed from the assessment as they were not functioning in concert with the other items. This can occur when items better represent different construct dimensions than the primary dimension under consideration, which is discussed in a subsequent section. The biodiversity assessment had nine negatively correlated items, the carbon assessment had eight, and the water assessment had nine. Thus there were consistent numbers of items functioning differently than most other items. These polarity findings were used with fit indices toward revisions of both items and instruments, and they will be used for reference to subsequent analyses of revised items.

Item Category Function. Rasch analysis provides item category frequency counts, average measures, outfit mean squares, between-category (Andrich) thresholds, and probabilities with corresponding graphics curves illustrating the structure. These statistics help address the issue of whether the five Likert category ratings are functioning as expected. Figure 2 illustrates the biodiversity assessment’s five category probability curves that ideally should each peak above the remaining four curves, in sequence. The category probabilities curves for biodiversity indicate that Likert categories 1 and 5 are clearly distinguished and functioning as intended, with low person measure associated with category1 and high person measure associated with category 5. Also there is some overlap (confusion) of category 1 with category 2, and category 5 with category 4. However, the middle three categories are more clearly confounded, with level 3 failing to peak above the other categories, suggesting an excess of categories. Probability curves for the water and carbon assessments were similar in shape and appearance supporting a student usage pattern of the five categories across the three assessment versions. Thus for all three assessments reducing Likert categories from five to four may improve measurement.



```
        CATEGORY PROBABILITIES: MODES - Structure measures at intersections
P      -+---------+---------+---------+---------+---------+---------+-
R  1.0 +                                                             +
O      |                                                             |
B      |                                                             |
A      |111                                                       555|
B   .8 +   111                                                 555   +
I      |      11                                             55      |
L      |        11                                         55        |
I      |          11                                     55          |
T   .6 +            11                                 55            +
Y      |              11                             55              |
    .5 +                1                           5                +
O      |                 11                       55                 |
F   .4 +                 22*22222           44444*44                 +
       |             2222   11   222     444   55   4444             |
R      |          222         1    3**3**3    5         444          |
E      |       222             **33  4*2  33**             444       |
S   .2 +    222              33  1 44   22 5  33              444    +
P      |2222              333    4*11   55*2    333              4444|
O      |              3333    444    1*5    222    3333              |
N      |       3333333   44444   55555 11111   22222   3333333       |
S   .0 +*****************55555555           11111111*****************+
E      -+---------+---------+---------+---------+---------+---------+-
       -3        -2        -1         0         1         2         3
        PERSON [MINUS] ITEM MEASURE
```

Figure 2. Category probability curves for the biodiversity assessment five-category scale (1 = very strongly disagree through 5 = very strongly agree). Each numbered category curve should have a distinctive peak in ordered sequence across the scale for optimal functioning. These curves are overlapping excessively.

The frequency and percent of use for each of the five categories, the outfit mean squares values, the average measures, and the Andrich thresholds are specified in Table 3 (e.g., for the biodiversity assessment, 644 responses were at category 1, 7% of overall responses of all students on all 96 items on the assessment). We expect the measures and Andrich thresholds to be ordered in correspondence, or in step, with the rating category. The category distribution of frequency counts and percentages were similar across the three instruments with the lowest use at categories one (7%) and five (13%) and the highest use categories at three (28% to 34%) and four (26% to 30%). Average observed measures were ordered, except for categories one and two for both the carbon and water assessments that were very similar average measures. All three assessments had ordered Andrich thresholds (step measurements), supporting expected step functioning of the categories. Step measurements should advance by approximately 1.0 logit when using five categories to show distinctions but not more than approximately 5.0 logits, as this gap would represent an excessive range (Bond and Fox, 2007; Linacre, 1999). The differences between pairs of Andrich thresholds on Table 3 indicate step advances of 0.97, 0.28, and 0.97 for the biodiversity assessment, 0.22, 0.98, and 0.46 for the water assessment, and 0.33, 0.94, and 0.72 for the carbon assessment. These findings further support a potential benefit of utilizing fewer categories on future versions of the assessment.

Table 3: Category Structure for Learning Progressions Instruments with Five Rating Categories

Category	1	2	3	4	5
Biodiversity					
Count	644	2028	2628	2851	1339
% of Total	7	21	28	30	14
Outfit MnSQ.	1.17	1.03	0.64	1.00	0.95
Ave. Meas.	.11	.09	.10	.28	.49
Andrich Threshold		-1.11	-.14	.14	1.11
Water					
Count	954	2118	3980	2995	1534
% of Total	8	18	34	26	13
Outfit MnSQ.	1.09	1.04	0.69	0.98	0.97
Ave. Meas.	.06	.07	.07	.19	.30
Andrich Threshold		-.77	-.55	.43	.89
Carbon					
Count	757	1894	3669	3039	1375
% of Total	7	18	34	28	13
Outfit MnSQ.	1.19	1.00	0.72	0.96	0.97
Ave. Meas.	.08	.06	.09	.24	.43
Andrich Threshold		-.90	-.57	.37	1.09

Dimensionality. We examine the unidimensionality of the measurements to determine whether they reflect one dominant construct, or dimension. Unidimensionality does not mean that only one psychological process is influencing responses, but rather that the multiple psychological processes that make up a construct, such as QI, affect the items such that they function similarly

(Smith 2004). For each of the three assessments whether calibrated measures share a primary QI construct dimension was estimated through examination of the variance explained using principal components analyses (PCA) of residuals, or what is left over after predicted variability is accounted for (Linacre 2014; Smith 2004). PCA of residuals differs from typical PCA or other factor analysis studies of scores in that one is not looking for a factor structure, but instead determining whether there is evidence of one primary dimension through examining variance explained by measures and checking for residual contrasts following removal of the variance explained. Eigenvalues, which represent approximately the number of items in PCA, also reflect the variance explained that can be calculated as a percentage. For the biodiversity assessment, PCA of residuals indicated that only 11.7% of variance was explained (eigenvalue of 12.8) by the Rasch measures. Similarly for the carbon assessment only 10.0% of variance was explained (eigenvalue of 10.6) by measures, and for the water assessment only 7.5 % of variance was explained (eigenvalue of 7.8). Variance that is not dominant is not supportive of unidimensionality, so this dimensional variance is low and unsupportive of unidimensionality. That is, the Rasch modeled dimension did not account for a dominant proportion of variance which may indicate additional dimensions. Principal components of residuals decomposed the unexplained variance to determine the relative strength of any secondary dimensions. If only one dimension is dominant, the contrasts should yield relatively small eigenvalues (ideally values less than 2). However, the first contrast in the residuals explained 8.1% of the variance (eigenvalue of 8.8) for the biodiversity assessment, 12.3% of the variance (eigenvalue of 13.1) for the carbon assessment, and 9.1% of the variance (eigenvalue of 9.4) for the water assessment. For the carbon and water assessments, these eigenvalues and corresponding percentages were large and exceeded that of the variance explained by the measures, further indicating multidimensionality within the data. These dimensionality findings will be considered with respect to the other diagnostics and measurement results.

Rasch Ruler. The measures for student and item are jointly considered in Rasch measurement. One of the primary ways of viewing the relationship between student and items is the Rasch ruler, or variable map (Wilson 2009; Wright and Stone 1979), which places the students and item measures on the same scale graphically. The Rasch rulers are provided in the Appendix: for the biodiversity assessment as Figure A.1; for the carbon assessment analysis as Figure A.2; and for the water assessment analysis as Figure A.3.

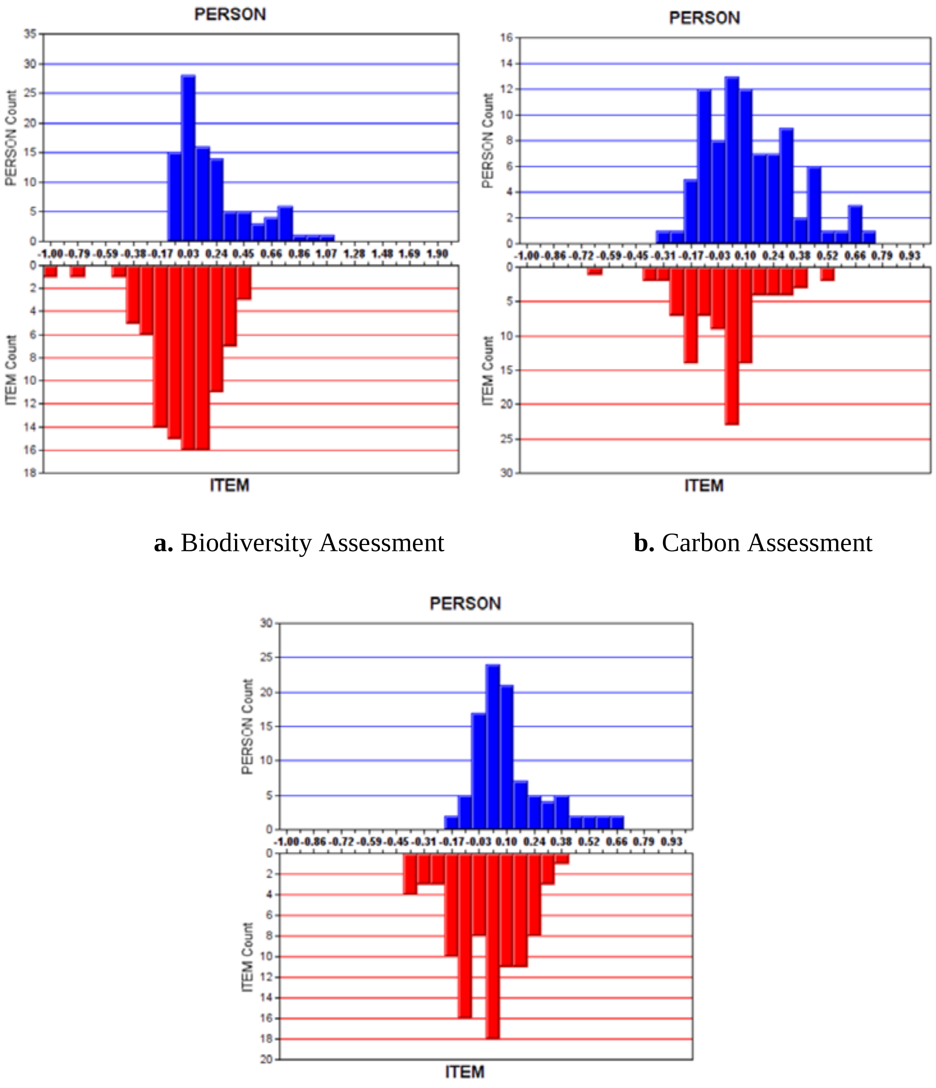
On the Rasch rulers, student measures are plotted on the left and item measures on the right of the vertical line, where the mean (M), standard deviation (S), and two standard deviations (T) are shown. The total measure mean for items was calculated using measures by students on each individual item. In Rasch measurement, item difficulty measures are based on the probability that a student will respond to an item correctly, or with the most accurate endorsement (agreement category) response in the case of the five-category Likert-scale assessment. A person has a 50% chance of "correctly" responding (indicating the most ideal response) to items that have the same item measure value as their student measure value. For example, on the biodiversity assessment, those students at the mean score measure of 0.21 (raw score of 309.9) have a 50% chance of responding to item MAPR4Q1[4] with the best answer. The items higher on the difficulty scale than the student measure are less likely to elicit agreement by that individual. The higher the items are on the scale the more difficult they are for the student to answer correctly through their level of agreement. Similarly, the lower the item is on the scale the easier it is for the student to agree appropriately. When unexpected responses are flagged by Rasch fit statistics these occurrences may represent a student correctly responding to questions that are especially difficult for them (above their student measure) or incorrectly responding to items that are predicted to be especially easy for them (below their student measure).

Information was added to the Rasch ruler that is not provided by the Winsteps program in order to visualize distribution of items. The items were shade-coded to provide a visual of distribution of items by proposed difficulty level. Questions written to assess at level 1 are light grey text; level-2 questions are dark text; level 3 are light shade of grey, and level 4 are dark shade of grey. The levels one through four discussed here refer to the hypothesized levels of the learning progression to which the questions corresponded. The color coding allows for a visual analysis of distribution of proposed level of endorsement difficulty versus student measure of ability. Note for example that level 2 items are disproportionately represented in the upper fourth of the biodiversity assessment Rasch ruler (Figure A.1).

The Rasch ruler allows for a comparison of the distribution of students to the distribution of items. The student distribution and mean is higher than the item distribution and mean for all three QI assessment versions, indicating that, overall, the assessments' items were not too difficult for the students. For the biodiversity assessment, 14 student measures exceeded all items, indicating they had better than a 50% chance to respond correctly to all items. On the water assessment, only four student measures exceeded all items, and, on the carbon assessment, six student measures exceeded all items. There were no student measures on the biodiversity assessment that were more than one standard deviation below the student mean. Specifically 33 of the 96 items were below all student measures (34%). If we omit the student measures on the carbon and water assessments that were more than one standard deviation below the student mean—only four students on each assessment—then approximately one third of the items on each

[4] Figure 1: Macro scale, Prediction, Level 4, Question 1 - One could extend the nonlinear trend of the data off the end of the graph, then estimate the year and number of wolves on the curve

of the assessments (31% carbon, 36% water) were below remaining student measures. This indicates that approximately a third of the items on each assessment would have more than a 50% chance of being answered correctly.



a. Biodiversity Assessment

b. Carbon Assessment

c. Water Assessment

Figure 3. Rasch ruler histograms illustrating targeting for each assessment instrument

Figure 3 provides histograms of the Rasch Rulers, which are visual representations of the correspondence, or overlap, of students with items. The histograms indicate a positive overlap of student and item measures. Overlap of item and student measures helps maximize measurement accuracy and identifies whether the assessment is well suited, or targeted, to the ability level of the participants. It is also evident that the student measures are typically higher than the item measures, that there are student measures above all item measures, and a number of item measures do not target well with student measures on the lower end of the scale. This lack of targeting tends to increase error for those item and student measures that do not overlap. A goal for revising the assessments will be to better align the targeting between student and items using these findings.

Empirical Performance Level. To further examine the alignment of proposed item level difficulty with the Rasch rating of measure order difficulty, the Rasch ruler was divided into four empirical performance levels based on student data:

- Level 1: one standard-deviation bin or more below the mean (easiest items).
- Level 2: between one standard-deviation bin below mean and the mean.
- Level 3: the mean and up to one standard-deviation bin above the mean.
- Level 4: one standard-deviation bin above the mean or more (most difficult items).

The assessment items were written to reflect increasing levels of complication, with level-1 items representing the lower anchor of a learning progression (novice level, e.g., Figure 1 Prediction 1b items), up to level-4 items representing the upper anchor of the learning progression (expert level, e.g., Figure 1 Prediction 4b items). Table 4 provides a count of items with respect to learning progression level (expected item challenge level) by actual performance level (empirical performance difficulty level). For example, Table 4 shows that the 24 biodiversity assessment level-4 items were distributed across all four empirical performance levels (e.g., three level-4 items appeared in the lowest empirical performance level). Level-1 items on the biodiversity assessment were found more often on empirical performance levels 1 and 2 as expected. However on the carbon and water assessments the level 1 items were more evenly distributed across all four empirical performance levels. Level-2 items were prevalent in empirical performance level 3, which was higher than expected on all three assessments. Level-3 items were found more at empirical performance level 4 on the biodiversity assessment which was higher than expected, but on empirical performance level 3 on the carbon assessment, and empirical performance levels 2 and 3 on the water assessment, which meets expectations. Unexpectedly, the carbon assessment had a large number of level-3 items at empirical performance level 1, which is counter to what was expected. The most unexpected trend was that on all three assessments, level-4 items were found most frequently in empirical performance levels 2 and 3, not in empirical performance level 4 as predicted. This provides evidence that either the hypothesized complexity of the levels of the learning progression are in question or that the items did not elicit the desired level of required understanding on the part of the persons taking the assessments.

Table 4: Expected Item Challenge Level by Empirical Performance Difficulty Level

Expected Item Challenge Level	Biodiversity Assessment				Carbon Assessment				Water Assessment			
	1st	2nd	3rd	4th	1st	2nd	3rd	4th	1st	2nd	3rd	4th
Level 4	3	7	9	5	3	7	13	1	2	8	10	4
Level 3	5	3	6	10	8	4	10	2	3	9	9	3
Level 2	4	5	7	8	4	3	9	8	5	4	9	6
Level 1	11	6	4	3	6	7	5	6	7	6	6	5
TOTAL	23	21	26	26	21	21	37	17	17	27	34	18

Assessment Items by Scale. Within each of the three QI assessment versions, items were developed to assess across three scales: macro scale, micro scale, and landscape scale (see Figure 1 for example of macro scale items). Student's ability to use quantitative reasoning may vary across these scales. At the macro scale, comfort with the context may reduce cognitive load and encourage quantitative accounts; at the micro scale, the context becomes inherently more quantitative as physical science is often required which may be more difficult for students; and at the landscape scale, quantitative accounts are driven by the need to generalize from local to regional or global contexts providing a different quantitative challenge. The hypothesis was that the scales from easiest to hardest would be: macro, landscape, micro. Table 5 presents data on scale by empirical performance level, where the number of scale items at each empirical performance level is listed in the table (e.g., Table 5 indicates that 13 of the 32 micro level items were at the first empirical performance level). The hypothesis was that macro scale items would be on the lower empirical performance levels, but they were more evenly distributed than expected on all three assessments. The micro scale items were spread relatively even across empirical performance levels 2, 3 and 4 for the biodiversity assessment, across the lower three performance levels for the carbon assessment, and at the empirical performance levels 2 and 3 for the water assessment. It was expected that more of the micro scale items would occur in empirical performance level 4 due to the quantitative nature of science required, but this is not supported by the data. Landscape items were clustered more in the upper three performance levels for all three assessments. It was expected the landscape scale items would be clustered on empirical performance levels 2 and 3, which is supported by the data. Overall there is considerable spread of the scale items across empirical performance levels, which supports the development of easier and harder level items within each scale.

Table 5: Scale by Empirical Performance Difficulty Level

Scale	Biodiversity Assessment				Carbon Assessment				Water Assessment			
	1st	2nd	3rd	4th	1st	2nd	3rd	4th	1st	2nd	3rd	4th
Macro	13	4	7	8	10	7	11	4	9	9	8	6
Micro	5	8	9	10	9	8	10	5	6	8	14	4
Landscape	5	9	10	8	2	6	16	8	2	10	12	8
TOTAL	23	21	26	26	21	21	37	17	17	27	34	18

How can one rank the level of difficulty of scales? One way is to calculate a weighted score across all three science strand assessments by multiplying the number of items by the performance level and summing. The weighted score for scale on the assessment indicates that the easiest scale for students was macro (222), followed by micro (243) and landscape (269). This supports the conjecture that students would do best on QI at the macro level, but inverts the hypothesized difficulty level for landscape and micro scales. However, the length of the assessment and fatigue could have influenced this order since this is precisely the order of the scales on the assessment.

Assessment Items by QI Elements. The distribution of items by QI elements and performance level was also analyzed (Table 6). The hypothesis was that students would find trend the easiest element, followed by translation, prediction, and revision. While all three assessments had a number of trend items in empirical performance level 1 as predicted, they also had an inordinate number of trend items at empirical performance level 3 and 4. The translation element items were evenly distributed across empirical performance levels for the biodiversity assessment, but were more prevalent in performance level 2 and 3 for the other assessments. The latter of these supports the hypothesis. Prediction element items were evenly distributed across all empirical performance levels on the biodiversity assessment, across the lower three empirical performance levels on the carbon assessment, and at empirical performance level 3 on the water assessment. This is counter to the expectation that these items would be more prevalent in empirical performance level 3 and 4. The revision element was most prevalent at empirical performance level 3 for the biodiversity assessment, empirical performance level 3 and 4 for the carbon assessment, and empirical performance level 2 for the water assessment. Thus the distribution for the biodiversity and carbon assessments supported the contention that revision would be more difficult for students, but the water assessment did not support this expectation. In fact, empirical performance level 4 was relatively evenly populated by items from all four elements on the water assessment and was reversed for the biodiversity assessment with the greatest number of items at the trend level.

Table 6: Elements by Empirical Performance Difficulty Level

Element	**Biodiversity Assessment**				**Carbon Assessment**				**Water Assessment**			
	1st	2nd	3rd	4th	1st	2nd	3rd	4th	1st	2nd	3rd	4th
Trend	8	2	3	11	9	1	11	3	7	5	7	5
Translation	7	5	6	6	2	6	12	4	5	6	9	4
Prediction	6	8	6	4	6	9	7	2	2	5	12	5
Revision	2	6	11	5	4	5	7	8	3	11	6	4
TOTAL	23	21	26	26	21	21	37	17	17	27	34	18

A weighted score was calculated to determine a ranking of difficulty for QI elements. There was no discernible difference on the assessments between trend (179) and prediction (177). Translation (185) was ranked higher than prediction, while revision (193) was ranked the highest. The rankings do not support the hypothesis that prediction would be more difficult than trend and translation. However, the elements of trend, translation, and revision were in the predicted order.

Discussion

Examining the QI assessments through simultaneous review of item data and student response data allowed for improvement of the current measurement accuracy by focusing on the assessment process validity. The intent was to influence future measurement accuracy following data-informed revisions of the assessment. Use of the Rasch rating scale model approach allowed for the development of additive measures from the raw ordinal ratings provided by the students. Rasch procedures include diagnostic statistics that enabled the refinement of these measures through identification of data that did not correspond with the ideal Rasch measurement model. For example, for the biodiversity assessment macro scale prediction items in Figure 1, item 2 (Prediction 1b) and item 4 (Prediction 2b) were both underfit (unpredictable) and item 8 (Prediction 4b) was overfit (too predictable). Such items were considered for revision. The items identified as very difficult, negatively correlated, and underfitting were all reviewed for possible revision. As part of this diagnostic review it was determined that items were represented across a broad range of difficulties, which allowed the assessment to better indicate the full range of student performance measurements. Specifically, item difficulty distributions showed a large proportion of items to be targeted well to student ability distributions for each assessment, supporting measurement validity. In addition, two of three assessments yielded relatively similar high levels of internal consistency and person separation reliability. However, despite favorable student to item targeting, student ability levels for some students generally exceeded the difficulty level for all items across assessments, so targeting was not seen as ideal. This finding suggests some benefit to revising some items to be more challenging.

Participants' use of the five categories of the Likert scale suggested that fewer categories, perhaps four versus the current five categories, would have provided greater measurement accuracy. Middle levels of the five category scales were typically overlapping with respect to measurement. One implication of this overlap is to reduce the number of categories in order to eliminate possible redundancy in the scale and help encourage a more meaningful distribution of responses on the scale. Reducing the categories to four removes the neutral option for students, requiring them to commit to either agreeing or disagreeing with each item. One potential advantage to eliminating the neutral option is improved identification of each participant's agreement tendency. For participants who would use a middle category to opt out of an agreement or disagreement level decision, an option outside of the agreement scale choices could also be added to help identify students who truly have no basis for responding one way or the other. With or without an opt-out choice, elimination of the middle category for this particular Likert-scale application would provide clearer information on middle-range student tendencies, potentially further improving measurement accuracy and item targeting. This reasoning is supported by the work of Wolfe and Smith (2007) who favor an even number of rating categories, stating "…the middle category is often used as a 'dumping ground' for participants that are compelled to provide a response but would not do so otherwise (p. 231−232)." This possible advantage to a four-category rather than a five-category scale will require an empirical test to determine whether such a measurement advantage exists on future administrations of the assessments that present four-category Likert items. Furthermore, providing the response choice for each item that would allow a student to opt out of any item, rather than provide a random response or misuse middle responses can be examined. That is, if a middle-range response was used by students when their intended response actually lied outside of the Likert scale, the result was simply erroneous data. An additional option such as "don't know," for example, may support increased efficacy of the Likert scale with four categories.

The length of the assessment at 96 items was a concern, and the student responses illustrated reasons for continued concern. Response patterns of some students, such as repeated use of a single level or apparently random responding, suggested student fatigue or low motivation, followed by misuse of the assessment due to deliberate careless responding. This finding supports an advantage for shorter versions of the assessments as well as consideration of means to influence motivation levels for future administration. Considering that duplicate items were developed at each scale-element-level, the test could be reduced in half by removing all duplicate items. This would help to address the likely fatigue and motivation problem. The items identified as a concern (e.g., underfitting or overfitting items) could be the first ones removed for the revised assessment. Another option is to reduce the assessment length by having students take only one level of the scale, assigning a class randomly to the three scales or having students take the assessment in three parts over three weeks. These implications depend upon the motivational character of subsequent students who are administered the revised QI assessment, but until sources of careless responding are minimized, the interpretations of findings must be tempered accordingly.

The research context of this investigation called for a unidimensional focus to examine the primary QI dimension. Thus, the assessment characteristics required examination with regard to this primary dimension. A majority of the items within each assessment fit well together according to weighted and unweighted fit analyses, supporting the broad QI dimension. However, PCA dimensionality analysis findings lead to the question of whether the improvements to the assessments following this investigation will lend greater support for a primary dimension (i.e., unidimensionality) or will it instead be necessary to divide up each revised assessment relative to dimensions to accurately analyze and interpret subsequent assessment findings. By identifying the multidimensionality evidence during this initial development stage, baseline statistics were established to allow for theoretical considerations of the dimensions that will be examined in future analyses of administrations of the revised assessments. For example, empirical investigations have shown that positive wording versus negative wording may lead to multidimensionality (Marsh 1996; Wang, Chen, and Jin 2015; Wolfe and Smith 2007, Yamaguchi 1997). Perhaps the reverse coding of negatively worded questions may have resulted in a secondary dimension. This possibility will be examined on a revised assessment by separately calibrating and comparing positively and negatively worded items to examine unidimensionality with PCA in conjunction with fit statistics. Other possible reasons for the lack of a clearly dominant dimension exist and include the influences of QI elements and scales, so empirical tests of new item sets that examine these additional aspects of the items can help to reveal the most advantageous means to examining the QI construct. Empirical tests of shorter forms of the assessment are being developed. Findings from future administrations of assessments will be compared with those from the current investigation regarding element, scale, and assessment versions to determine whether the unexpected QR progressions findings in the present study resulted from the item, assessment, and administration issues identified.

Concluding Remarks

Model-based reasoning skills are necessary for scientifically literate citizens to engage in 21st-century problem solving. This investigation represents a step toward improved diagnosis of model-based reasoning skills for educators. The analysis conducted provided a demonstration of several current measurement tools that can be used to develop and refine learning assessments that support model-based reasoning skills development. Measurement tools used in the current phase of the research include the construction of linear Rasch rating scale measures of item difficulty and student ability, as well as indices and visual graphics that allow focus on item polarity, category functioning, dimensionality, targeting, reliability, separation, and item/person fit. Following instrument revisions, additional Rasch measurement tools will be beneficial within subsequent investigations of the instrument to help refine the assessment process for wider use. Additional tools and techniques include examinations of differential item function (DIF) that involves comparison of item measures between subgroups of students (Linacre, 2014; de Ayala 2009, Bond and Fox, 2007, Smith and Smith, 2004). With refined items making up shorter instruments, both the dimensional character and item functioning can help specify whether the instruments measure the learning progressions levels intended in a consistent manner.

Findings from this investigation supported further refinement of these assessments for use by teachers, administrators, and researchers as a part of an efficient diagnostic process to improve understanding of what QR abilities students possess. The student-level and classroom-level data generated from these improved assessments, in conjunction with other available performance outcomes can also allow K-12 curriculum developers the opportunities to integrate the explicit teaching of QR within science contexts and provide data-informed support for STEM professional-development opportunities. The revised QI assessments and subsequent research that follows from this investigation are available to educators and researchers from the first author.

Acknowledgment

This project is supported in part by a grant from the NSF: *Culturally Relevant Ecology, Learning Progressions, and Environmental Literacy* (DUE-0832173).

Appendix A. Rasch rulers for the three assessments

```
MEASURE
     PERSON - MAP – ITEM
         <more>|<rare>
 1          X  +
               |
            .  |
         .XXX T|
           XX  |
          .X S|T MITD3
          XXX  |  LATD3LATD4MAPR3MATS3MIPR2MIRV1
         XXXX  |  LARV2LATD2MATD3MIRV2MITS2
     .XXXXXXX M|S LATD3LATS1LATS3LATS4MARV1MARV3MATD2MATD4MATD4MIPR2MIPR3MITD2MITD4MITS3
  -------------------------------------------------------------------------
   XXXXXXXXXX  |  LAPR1LAPR2LARV1LARV2LARV3LARV4LATS3MAPR4MARV2MARV2MATS2MATS4MATS4MIRV2QMIRV3MITD2MITD4
  -------------------------------------------------------------------------
0XXXXXXXXXXXX  +M LAPR3LAPR3LARV3LARV4LATD2LATS2LATS2LATS4MAPR4MIPR1MIPR4MIRV4MITD3MITS1MITS4
        .XXXX S|  LAPR1LAPR2 ARV1LATD4MAPR1MAPR1MARV3MARV4MIPR1MIPR4MIRV1MIRV3MIRV4MITS2MITS4
  -------------------------------------------------------------------------
               |S LAPR4LAPR4LATD1LATD1LATS1MARV1MATD3MATS1MIPR3MITD1
              T|  MAPR2MAPR2MARV4MATS3MITD1MITS1MITS3
               |  MAPR3MATD1MATS2
               |T MATD1
               |
               |
               |  MATS1
               |
   -1          +  MATD2
         <less>|<frequent>

EACH "X" IS 2. EACH "." IS 1.
```

Figure A.1. Variable map, or Rasch ruler, for biodiversity assessment illustrating item types by quartile

```
MEASURE
       PERSON - MAP - ITEM
            <more>|<rare>
    2             +
                  |
                  |
                  |
                  |
                  |
                  |
               X  |
                  |
              XX  |
    1            T+
                  |  MATD1
            XXXX  |T LARV2
           XXXXX  |  MIPR2
        XXXXXXXX S|  LATS3MATS3MIPR4
            XXXX  |  LATD2LATD4LATS1MITD4
             XXX  |S LARV1LARV3LATD1LATS4MARV2MARV2MATD2MITS4
       ---------------------------------------------------------------------------------
        XXXXXXXX M|  MARV1MATS1MIPR3MITD2MITD4
         XXXXXXX  |  LARV3LATD3MARV3MATD4MATS2MATS3MATS4MIPR1MIRV2MIRV2MITD2MITS3
          XXXXXX  |  LAPR4MAPR3MAPR4MAPR4LAPR1LARV2LARV4LATS2MARV3MARV4MIPR3MIRV1MIRV4MIRV4MITD1MITS3
    0 XXXXXXXXXX  +M LATS4LATD3MIPR2
       ---------------------------------------------------------------------------------
        XXXXXXXX S|  LARV3LARV4MARV1MATD4MIPR1MIPR4
              XX  |  LAPR1LAPR2LAPR4LATS3MAPR1MAPR1MARV4MATD1MATS4MIRV3MIRV3MITD1MITS2MITS4
              XX  |  LAPR2LARV1LATD1LATD4LATS2MAPR2MATD2MATD3MIRV1MITD3MITS1MITS2
       ---------------------------------------------------------------------------------
                 T|S MATS1MATS2
                  |  MAPR2MAPR3
                  |
                  |  MATD3MITD3
                  |T
                  |
   -1             +
                  |  LAPR3LATD2
                  |
                  |  LATS1
                  |
                  |
                  |
                  |
                  |
                  |
   -2             +
            <less>|<frequent>
```

Figure A.2. Variable map, or Rasch ruler, for carbon assessment illustrating item types by quartile

```
MEASURE
PERSON - MAP - ITEM
    <more>|<rare>
2         +
          |
          |
          |
          |
       .  |
          |
          |
1      .  +
          |
       X T|
          |
    .XXX  |T MATS2
      XX S|  LAPR1MIPR4
   .XXXX  |  LATD3LATD4MAPR4MATD4MITS4
  .XXXXX  |S LAPR2LARV2LATD2LATD4LATS2LATS4MAPR1MARV1MARV2MATS3MATS4MIPR2MIPR3MIRV1MITD4MITS1MITS3MITS4
  ------------------------------------------------------------------------
XXXXXXXX M|  LAPR4LARV4LATS1MARV2MATD2MIPR3MIRV2
   XXXXX  |  MAPR1MAPR2MIRV2MITD2MITD3LARV1MITD4LARV3LATD1LATD2LATS3MAPR2MATD3MATS2MIPR1MIPR4MIRV3MIRV3
0 XXXXXX  +M MIRV1LAPR4MIRV4MITD3MITS2MATD4
  ------------------------------------------------------------------------
XXXXXXXX S| LAPR2LAPR3LAPR3LARV1LARV2LATS4MAPR3MAPR4MATD3
       X  | LAPR1LARV3LATS2LATS3MARV1MARV3MATD1MATS1MATS1MIPR1MIPR2MITS2
  -------------------------------------------------------------------------
       .  |S LARV4MAPR3MARV4MARV4MATD1MATD2MATS3MATS4MITD2MITS3
         T|  MARV3MIRV4
       .  |
          |T LATD1LATD3MITD1
          |  LATS1MITD1
          |
          |
-1        +  MITS1
    <less>|<frequent>
```

Figure A.3. Variable map, or Rasch ruler, for water assessment illustrating item types by quartile

References

AERO. 2011. American Education Reaches Out, AERO Mathematics Curriculum Framework: K-8 Standards and Performance Indicators http://www.projectaero.org/aero_standards/mathematics-framework/AERO-MathematicsCurriculumFramework.pdf

Andrich, D. 1978. A rating formulation for ordered response categories. *Psychometrika* 43: 561–573. http://dx.doi.org/10.1007/BF02293814

Bond, T., and C. M. Fox. 2007. *Applying the Rasch Model: Fundamental measurement in the human sciences*, 2nd ed. Mahwah, NJ: Lawrence Erlbaum.

Carlson, M., S. Jacobs, E. Coe, S. Larsen, and E. Hsu. 2002. Applying covariational reasoning while modeling dynamic events: A framework and a study. *Journal for Research in Mathematics Education* 33: 352–378. http://dx.doi.org/10.2307/4149958 (accessed February 6, 2012)

Corcoran, T., F. Mosher, and A. Rogat. 2009. *Learning progressions in science: An evidence-based approach to reform.* Philadelphia, PA: Consortium Policy Research in Education. http://dx.doi.org/10.12698/cpre.2009.rr63

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

Duschl, R. A., H. A. Schweingruber, and A. W. Shouse. 2007. *Taking science to school: Learning and teaching science in grades K-8* .Washington, D.C: National Academies Press.

Engelhard, G. 2013. *Invariant measurement: Using Rasch Models in the social, behavioral, and health sciences*. New York: Routlege.

Lehrer, R. and L. Schauble. 2002. *Investigating real data in the classroom: Expanding children's understanding of math and science.* New York: Teachers College Press.

Linacre, J. M. 1999. Investigating rating scale category utility. *Journal of Outcome Measurement,* 3, 2: 103–122.

Linacre, J. M. 2012. *Winsteps® (Version 3.74.0) [Computer Software]*. Accessed August 22, 2012, http://www.winsteps.com/

———. 2014. *Winsteps® Rasch measurement computer program User's Guide.* Beaverton, Oregon: Winsteps.com

Louca, L. T., Z. C. Zacharia, and C. P. Constantinou. 2011. In quest of productive modeling-based learning discourse in elementary school science. *Journal of Research in Science Teaching* 48: 919–951. http://dx.doi.org/10.1002/tea.20435

Madison, B. L., and L. A. Steen, eds. 2003. *Quantitative literacy: Why numeracy matters for schools and colleges*. Princeton, NJ: National Council on Education and the Disciplines.

Marsh, H. W. 1996. Positive and negative global self-esteem: A substantively meaningful distinction or artifactors? *Journal of Personality and Social Psychology* 70: 810–819. http://dx.doi.org/10.1037/0022-3514.70.4.810

Mayes, R. L., F. Peterson and R. Bonilla. 2012, Quantitative reasoning in context. In *WISDOMe: Quantitative Reasoning and Mathematical Modeling: A Driver for STEM Integrated Education and Teaching in Context*, eds. Author and L. L. Hatfield, 7–38. Laramie, WY: University of Wyoming.

———. 2013. Quantitative reasoning learning progressions for environmental science: Developing a framework. *Numeracy* 6(1): Article 4. http://dx.doi.org/10.5038/1936-4660.6.1.4

Mayes, R. L., J. H. Forrester, J. S. Christus, F. I Peterson, R. Bonilla & N. Yestness. 2014a. Quantitative reasoning in environmental science: A learning progression. *International Journal of Science Education* 36(4): 635−658. http://dx.doi.org/10.1080/09500693.2013.819534

Mayes, R. L., J. H. Forrester, J. S. Christus, F. Peterson and R. Walker. 2014b. Quantitative reasoning learning progression: The matrix. *Numeracy* 7(2): Article 5. http://dx.doi.org/10.5038/1936-4660.7.2.5

National Governors Association Center for Best Practices, Council of Chief State School Officers. 2010. *Common Core State Standards Mathematics*. Washington DC: NGAC, CCSSO.

National Research Council. 2012. *A Framework for K-12 Science Education: Practices, Crosscutting Concepts, and Core Ideas*. Washington, DC: The National Academies Press.

NGAC. See National Governors Association Center

NGSS Lead States. 2013. Next Generation Science Standards: For States, By States. Washington, DC: The National Academies Press.

Pluta, W. .J., C. A. Chinn, and R. G. Duncan. 2011. Learners' epistemic criteria for good scientific models.*Journal of Research in Science Teaching* 48: 486−511. http://dx.doi.org/10.1002/tea.20415

Rasch, G.. 1960/1980. *Probabilistic models for some intelligence and attainment tests.* Chicago: University of Chicago Press.

Schwarz, C. V., B. J. Reiser, E. A. Davis, L. Kenyon, A. Archer, D. Fortus, Y. Shwartz, B. Hug, and J. Krajcik. 2009. Developing a learning progression for scientific modeling: Making scientific modeling accessible and meaningful for learners. *Journal of Research in Science Teaching* 46: 632−654. http://dx.doi.org/10.1002/tea.20311

Schwartz, D. L., and T. Martin. 2004. Inventing to prepare for future learning: The hidden efficiency of encouraging original student production in statistics instruction. *Cognition and Instruction* 22: 129–184. http://dx.doi.org/10.1207/s1532690xci2202_1

Smith, C. L., C. A. Wisner, and J. Krajcik. 2006. Implications of research on children’s learning for standards and assessment: A proposed learning progression for matter and the atomic molecular theory. *Measurement: Interdisciplinary Research and Perspective* 4: 1−98. http://dx.doi.org/10.1080/15366367.2006.9678570

Smith, E V. 2004. Detecting and evaluating the impact of multidimensionality using item fit statistics and principal component analysis of residuals. In *Introduction to Rasch measurement: Theory, models, and applications*, eds. E.V. Smith & R. M. Smith, 575−600. Maple Grove, MN: Jam Press.

———, and R. Smith. 2004. *Introduction to Rasch measurement: Theory, models, and applications*. Maple Grove, MN: Jam Press.

Steen, L. A. 2004. *Achieving quantitative literacy: An urgent challenge for higher education.* Washington, DC: Mathematical Association of America.

Stefani, C., and G. Tsaparlis. 2009. Students' levels of explanations, models, and misconceptions in basic quantum chemistry: A phenomenographic study. *Journal of Research in Science Teaching* 46: 520–536. http://dx.doi.org/10.1002/tea.20279

Taylor, A. R., and G. Jones. 2009. Proportional reasoning ability and concepts of scale: Surface area to volume relationships in science. *International Journal of Science Education* 31: 1231−1247. http://dx.doi.org/10.1080/09500690802017545

Thompson, P.W. 2011. Quantitative reasoning and mathematical modeling. In *New perspectives and directions for collaborative research in mathematics education,* eds. L. L. Hatfield, S. Chamberlain, and S. Belbase, 33−57. Laramie, WY: University of Wyoming.

Wang, W., H. Chen, and K. Jin. 2015. Item response theory models for wording effects in mixed-format scales. *Educational and Psychological Measurement* 75(1): 157−178. http://dx.doi.org/10.1177/0013164414528209

Wilson, M. 2009. Measuring progressions: Assessment structures underlying a learning progression. *Journal of Research in Science Teaching* 46:716−730. http://dx.doi.org/10.1002/tea.20318

Wolfe, E. W., and E. V. Smith, Jr. 2007. Instrument development tools and activities for measure validation using Rasch Models: Part I – Instrument development tools. In *Rasch measurement: Advanced and specialized applications,* eds. E. Smith, Jr., and R. M. Smith, 202−242. Maple Grove, MN: Jam Press.

Wright, B. D., and M. C. Mok. 2004. An overview of the family of Rasch measurement models. In *Introduction to Rasch measurement: Theory, models, and applications*, eds. E. V. Smith, and R. M. Smith, 1−24. Maple Grove, MN: JAM Press.

Wright, B. D., and M. H. Stone. 1979. *Best test design*. Chicago: MESA Press.

Yamaguchi, J. 1997. Positive versus negative wording. *Rasch Measurment Transactions* 11: 567.

Zahner, D., and J. E. Corter. 2010. The process of probability problem solving: Use of external visual representations. *Mathematical Thinking and Learning* 12: 177−204. http://dx.doi.org/10.1080/10986061003654240