Measuring Students’ Scientific Content and Inquiry Reasoning

Amelia Wenk Gotwals & Nancy Butler Songer, University of Michigan, 610 East University St. Ann Arbor, MI 48109
Email: agotwals@umich.edu, songer@umich.edu

Abstract: Recently, several important documents have promoted inquiry-based science as the main way for science to be taught and learned. In addition, there have been advancements made in the measurement sciences that allow for sophisticated and complex ways to score and interpret student responses on assessment tasks. However, while many studies have shown the benefits of scientific inquiry in the classroom and others have described new types of psychometric models available for scoring analysis, few have combined the two to develop a better understanding of how students “know” science. This paper briefly describes an assessment system used to create items that systematically measure both students’ content knowledge as well as two complex inquiry-reasoning skills. Then, using student responses to an assessment made using this system we employ multidimensional psychometric models to allow us to explain the nature of the types of knowledge students draw on when encountering scientific scenarios.

Introduction
Recent research and science education reform documents (e.g. National Research Council, 1996) have given strong endorsements for inquiry-based science teaching and learning. However, science education researchers have not fully evaluated how students develop content knowledge in relation to the development of inquiry-reasoning abilities. It is unclear the extent to which students’ content knowledge and their inquiry reasoning skills are intertwined. Discovering the nature of this relationship is important both in informing how content knowledge and inquiry-reasoning skills should be presented to students as well as how these types of knowledge and skills should be assessed. Systematically measuring complex reasoning in science, however, is difficult because many current assessments are not good measures of student understanding being that they are based on outdated theories of how students learn, outdated measurement systems, and often only test fact-based knowledge (Black, 2003; Pellegrino, Chudowsky, & Glaser, 2001). In science education, we need measures that can gather information not just about students’ science content knowledge, but also about their inquiry-reasoning skills and how science content and inquiry-reasoning skills interact in students’ abilities to work with complex scientific ideas. This study examines the extent to which we can disentangle students’ content knowledge and inquiry-reasoning abilities using a systematic assessment of content knowledge and two inquiry-reasoning abilities: formulating scientific explanations and interpreting data.

Literature Review

Inquiry Science
The process of inquiry is modeled on the scientist's method of discovery. It focuses on asking questions, exploring these questions, considering alternative explanations, and weighing evidence. Part of why inquiry is important is because it can provide students with “real” science experiences, e.g. experiences with many important features of science as practiced by professional scientists (Brown, Collins, & Duguid, 1989). According to the National Science Education Standards (1996),

Inquiry is a multifaceted activity that involves making observations; posing questions; examining books and other sources of information to see what is already known; planning and conducting investigations; reviewing what is already known in light of experimental evidence; using tools to gather, analyze, and interpret data; proposing answers, explanations, and predictions; and communicating the results. (p. 23)

In inquiry-based science classrooms, students are expected to reason with scientific knowledge through activities such as conducting investigations, formulating explanations and interpreting data. Various inquiry methods have been shown to encourage the inclusion of all students in science classrooms and also promote greater student achievement gains in both scientific content and inquiry knowledge (Krajcik et al., 1998; Mistler-Jackson & Songer,
2000; Songer, Lee, & McDonald, 2003; White & Frederiksen, 1998). For these reasons and others, the National Science Education Standards state that, “Inquiry into authentic questions generated from student experiences is the central strategy for teaching science” (National Research Council, p. 31).

Assessment

One of the main reasons for assessment is to determine what students know and to try to understand “how” they know it. Because we cannot see into students’ heads, and because knowledge, itself, is unobservable, one of the only ways to determine what students know is through observations of their behavior or performance on an assessment (Mislevy, 2003). Despite much research on the effectiveness of inquiry methods in the science classroom, there is little research on systematic methods to measure what students know and how they know it in the realm of science inquiry-reasoning skills (Haertel & Mislevy, 2001). In order to gather information about the relationship between students’ content knowledge and inquiry reasoning skills, we must take advantage of recent advances in both theories of learning, which promote inquiry methods of teaching and learning science and theories of measurement science that have influenced the way assessments are created and scored (Pellegrino et al., 2001).

One of these advances in assessment theory is Evidence Centered Design – ECD (Mislevy, 2003), which views assessment development as not just task creation, but more as the conceptualization of an assessment argument from which we can reason about students’ knowledge and abilities. In order to do this, assessment developers must first specify the complex of knowledge, skills and abilities that are to be focused on. Next they must answer the question, “What behaviors or performances would enable us to determine if students possess the knowledge, skills, and abilities [thus] defined?” (Mislevy, Steinberg, Almond, Haertel, & Penuel, 2003). And finally they must describe and create the types of situations and tasks that would allow students to demonstrate these performances or behaviors. This is generally an iterative process through which tasks are created based on the principles of ECD, given to students, and then further refined to ensure that they meet the criteria put forth by criteria described above. Through this articulation of the process of assessment development, we are better able to make valid claims about students’ knowledge and abilities.

Methods

Task Design

Using the principles of ECD, we developed an assessment to coordinate with an inquiry-based biodiversity curriculum. This curriculum covers content topics associated with biodiversity knowledge, such as animal interactions, animal classification, and the measurement of animal biodiversity. In addition, it focuses on helping students develop two inquiry-reasoning abilities: formulating scientific explanations and interpreting data. As part of our philosophy of how people learn, we believe that inquiry in the classroom can take various forms and can occur at many different levels (Songer et al., 2003). Therefore, it is important to develop tasks specifically oriented to different levels of complexity to accurately evaluate students’ developing abilities over time. The kinds of tasks that we need to employ in order to gather information about students’ content knowledge and inquiry reasoning skills can be structured in many ways to reflect the different ability levels that students may have. In this study, we conceptualize the difficulty of science inquiry assessment tasks as having two main dimensions: the difficulty of the science content and the difficulty of the science inquiry. To address both of these aspects of task difficulty, we created a matrix that lays out three possible levels for each dimension. Table 1, below, illustrates the matrix for the inquiry skill, formulating scientific explanations. Science content knowledge is divided into three categories: simple, moderate and complex; whereas, inquiry-reasoning skills are separated into three levels: step1, step 2, and step 3. For the formulating scientific explanations inquiry skill, we created degrees of inquiry tasks based on the amount of support or scaffolding the task provides for explanation formation (Huber, Songer, & Lee, 2003). Using this matrix as well as a similar matrix for the inquiry-reasoning skill, interpreting data, we developed a suite of assessment tasks to gather information about students’ biodiversity content knowledge as well as both of the specified inquiry-reasoning abilities.
Table 1. Levels of Content and Inquiry Knowledge Needed for Assessment Items For Inquiry-Skill: “Formulating Scientific Explanations Using Evidence”

<table>
<thead>
<tr>
<th>Complexity of Reasoning</th>
<th>Amount of Content Required</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Simple – minimal or no extra content knowledge is required</td>
</tr>
<tr>
<td>Step 1- Students match a given claim to relevant evidence</td>
<td>Students are given all of the evidence or data and the claim. Minimal or no extra content knowledge is required</td>
</tr>
<tr>
<td>Step 2- Students are prompted (or asked) to give a claim and are specifically asked for pieces of evidence</td>
<td>Students are given evidence, to choose the claim and construct the explanation, minimal or no additional knowledge or interpretation of evidence is required</td>
</tr>
<tr>
<td>Step 3-Students must construct a claim and explanation using relevant data and evidence without any scaffolding</td>
<td>Students must construct a claim and explanation however, they need to bring minimal or no additional content knowledge to the task</td>
</tr>
</tbody>
</table>

Data Collection
This study takes place within the implementation of an inquiry-based biodiversity curricular program in three inner-city Midwestern schools’ sixth grade classes. After completing this curriculum that specifically scaffolded students reasoning abilities, students were given a test containing nineteen items representative of all cells of the matrix for both inquiry-reasoning skills: formulating scientific explanations and interpreting data. The items were both multiple-choice items as well as constructed response items. The data for this study come from over three hundred students’ responses to this assessment.

Coding/Analyzing Data
Using both a validity analysis of the items as well as the cells of the matrix as a guide for what types of knowledge should be demonstrated in each question, we coded student responses for demonstrations of the knowledge and skills that the tasks were build around, in particular the inquiry skills (interpreting data and explanations) and amount of content infused into the task. Depending on the particular cell of the matrix (or matrices) that the item falls into, the question may be coded solely for the content knowledge that the student draws on when answering the question, the types of reasoning that they draw on, or a both. The rubric was specified for each item, but the general scheme is shown below in table 2. Some questions will have only one code, while others may involve up to seven codes (one for each piece of the rubric below).
Table 2. Generalized Coding Rubric for Assessment

<table>
<thead>
<tr>
<th>Content Codes</th>
<th>Explanation Codes</th>
<th>Interpreting Data Codes</th>
</tr>
</thead>
<tbody>
<tr>
<td>One point for each piece of relevant content used in answer</td>
<td>Claim (1 total point) Correct (1): Formulates a claim statement (regardless of correct content) Incomplete (0): other response or no claim statement Evidence: (2 total points) *must be consistent with claim Full (2): Gives two pieces of relevant evidence - related to scenario provided regardless of correct content Partial (1): Gives one piece of relevant evidence Incomplete (0): irrelevant or no evidence</td>
<td>Draw conclusion (1 total point) (1): Correctly identifies correct answer to question (0): Other or no conclusion</td>
</tr>
<tr>
<td></td>
<td>Reasoning: (2 total points) Full (2): Ties claim to evidence explicitly Partial (1): Ties claim to evidence implicitly Incomplete: (0)=No reasoning or incorrect / irrelevant reasoning given</td>
<td>Identify pattern (2 total pts) Full (2): Ranked/compared each item across each category Partial (1): Ranked/ compared some items across each category OR all items across some categories Incomplete (0): provides no evidence</td>
</tr>
<tr>
<td></td>
<td>Disconfirming Data (2 total points) (2): Acknowledged disconfirming data and reasoned with it (1): Acknowledged disconfirming evidence but does not reconcile it (0): Does not address disconfirming evidence</td>
<td></td>
</tr>
</tbody>
</table>

Dimensionality Analysis

In the types of items given on this test, it is required that students possess different kinds of knowledge: content knowledge, the ability to formulate a scientific explanation, and the ability to interpret data. One of the main benefits of Item Response Theory (IRT) is that it allows us use performances on assessment in order to estimate a person’s “trait level,” using maximum likelihood methods. A trait level, also called a latent trait (and denoted Θ), is essentially a person’s capacity or ability on a given construct (such as biodiversity content knowledge or ability to formulate explanations) (Embretson & Reise, 2000; Mislevy, Wilson, Ercikan, & Chudowsky, 2002). In Item Response Theory (IRT), a test that only addresses one kind of knowledge or ability is called a unidimensional model. “The unidimensional IRT models define a single trait level for each person. These models are appropriate for items that involve a single underlying trait or a combination of traits that is contingent across items” (Embretson & Reise, 2000, p. 92). On the other hand, a certain class of IRT models, called multidimensional models, allow for “multiple intelligences” to play a role in completing certain items. “Multidimensional IRT models contain two or more parameters to represent each person…. If success on an item depends on more than one latent trait, then the person’s response provides information about two or more traits simultaneously” (Embretson & Reise, 2000, p. 82).

In order to model the data from the assessments, we used the ConQuest software program (Wu, Adams, & Wilson, 1998), which uses an extension of Rasch’s Logistic Model (Rasch, 1960), the multidimensional random coefficients multinomial logit model, MRCMLM, as its basis (Adams, Wilson, & Wang, 1997). The assessment that we created has a combination of dichotomous items (coded 0, 1) and polytomous items (coded 0, 1, 2, 3), therefore we used a partial credit model (Masters, 1982) which is part of the MRCML-family of models that allows for analyzing items that require multiple steps and where assigning partial credit is important in understanding students’ trait levels. In order to determine whether students use an overall science ability or whether their content knowledge is a separate entity from their inquiry reasoning, we fit three different models to the data: a unidimensional model, a two-dimensional model, and a three-dimensional model. For the unidimensional model, we considered content combined with both inquiry skills (explanations and interpreting data) as one latent trait. If this model fits the data best, it would indicate that content and inquiry reasoning abilities are intricately intertwined and that when students encounter scientific situations they would be using an overall science ability that included both content knowledge and inquiry reasoning to solve these problems.
For the two-dimensional model, we consider both types of inquiry reasoning as a single latent trait (an overall inquiry reasoning ability) and content knowledge as a separate latent trait. If this model fits the data best, it would indicate that students have developed their content knowledge separately from their inquiry reasoning abilities. In this situation, students could fully grasp the content knowledge, but not necessarily have developed inquiry-reasoning abilities or vice versa. However, this model assumes that inquiry reasoning is a uniform trait and if a student is able to formulate explanations, they would also be able to interpret data. Finally, for the three-dimensional model, we consider content knowledge, ability to formulate scientific explanations, and ability to interpret data as three separate latent traits. In this model, students could possess any combination of traits, such as being able to formulate an explanation, but not have content knowledge or be able to interpret data. This model would indicate that students have separately developed each of these skills and that possessing one of these skills is not necessarily required to possess any of the others.

In order to compare the relative fit of the hierarchical unidimensional and multidimensional models, we can use the deviance statistics for the three models as a measure of how well the data fit the intended model. Since the one-dimensional model is nested within the two-dimensional model and the two-dimensional model is nested within the three-dimensional model, we can compare the deviance of model pairs with a chi-square distribution, with the degrees of freedom equal to the number of additional parameters in the more complex model.

Results

The comparison of the unidimensional, two-dimensional and three-dimensional models show that the three-dimensional model fits the data best (see Table 3 below).

Table 3. Comparison of One, Two, and Three-Dimensional Models (N=312)

<table>
<thead>
<tr>
<th></th>
<th>Deviance (-2* loglikelihood)</th>
<th>Number of parameters</th>
</tr>
</thead>
<tbody>
<tr>
<td>One-Dimension</td>
<td>16576.479</td>
<td>76</td>
</tr>
<tr>
<td>Two-Dimension</td>
<td>16573.578</td>
<td>78</td>
</tr>
<tr>
<td>Three-Dimension</td>
<td>16562.602</td>
<td>81</td>
</tr>
</tbody>
</table>

This result shows that the two-dimensional model fits the data slightly better than the one-dimensional model (p<0.10; critical value: χ²(2, 0.10)=2.706). The difference between the two-dimensional model and three-dimensional model, though, is more compelling at 10.976, showing that the three-dimensional model fits the data significantly better than the two-dimensional model (p<0.025; critical value: χ²(3, 0.025)=9.348). The difference in deviance between the three-dimensional model and the unidimensional model is 13.877, and shows that the three-dimensional model fits the data better than the one-dimensional model (p<0.025; critical value: χ²(5, 0.025)=12.833). These results illustrate that the three-dimensional model is the best fit to the data, meaning that when students interact with the items on this test, they are using three distinct abilities (their content knowledge, their ability to formulate explanations, and their ability to interpret data) to answer questions.

Discussion

The data from the dimensionality analysis show that students use three unique skills when answering this set of assessment tasks. This assessment, that was systematically designed to measure biodiversity content knowledge and two complex reasoning skills, formulating explanations and interpreting data, was able to disentangle students’ content knowledge and inquiry reasoning abilities. These results show both that the design of these assessment tasks allowed access to students’ knowledge in each of these areas and that they did measure the three skills that they were designed to measure, as well as showing that it is possible to disentangle students’ content knowledge from their complex reasoning abilities. The second of these two findings is of most importance for this paper.

It is very interesting that students do possess separate abilities or skills within the domain of science and that when they encounter scientific scenarios similar to those seen in this assessment, they draw on these unique skills in solving scientific problems. Perhaps this is not entirely surprising, considering that these skills require students to do very different things. For instance, being able to correctly describe a producer or consumer is not the
same as being able to create a scientific explanation consisting of a claim, evidence, and reasoning, and this is not the same as being able to read a table or graph and draw conclusions from it. However, in many assessments, including many standardized assessments, students are only tested on a unidimensional or at most a two-dimensional type of science knowledge. Many large-scale tests focus only on content knowledge. Some assessments, however, may bring in a component of reasoning, but rarely, if ever, do they separate out reasoning components. By doing this, other assessments may not be tapping into all of a students’ science knowledge. In addition, by measuring science in this way, assessments will tend to privilege content knowledge over any type of reasoning ability. While we would not argue with the importance of content knowledge, giving it priority over reasoning abilities downplays what standards documents and research programs have been stressing, which is the integral nature of having students’ reason with their content knowledge rather than just possessing the content knowledge as a separate entity of knowledge.

If we are interested in promoting complex reasoning skills in the science classroom, developing items that tap into the important aspects of students’ scientific abilities and measuring students’ responses on these separate dimensions is very important. While it is likely that some students will be at similar levels in each latent ability, there will be some students who may have understood the content knowledge, and yet did not do well on the reasoning abilities. Having tests that can tease this information out is essential for both the teacher and the student so that each can more fully understand their strengths and weaknesses. Teachers may find that students are grasping the content knowledge, but are not able to perform complex inquiry-reasoning. This type of information can be used to inform teachers where they can focus or alter their teaching to ensure that students not only possess content knowledge, but can use that content knowledge in complex reasoning situations.

In summary, emerging research into the ways in which students interact with content and complex inquiry reasoning skills shows that students possess and tap into distinct abilities when encountering scientific scenarios. Developing systems for measuring these skills is essential in order to fully appreciate students’ understandings of science and in order to provide adequate feedback to both teachers and students about their performance on an assessment.

**Endnotes**

1. There is not enough space in this paper to fully describe the system of task design utilized to create this assessment. For more information, please see (Songer & Gotwals, 2004)

**Works Cited**


