Validating for Use and Interpretation: A Mixed Methods Contribution Illustrated

Linda Morell and Rachael Jin Bee Tan

Journal of Mixed Methods Research 2009; 3; 242 originally published online May 8, 2009;
DOI: 10.1177/1558689809335079

Journal of Mixed Methods Research
Volume 3 Number 3
July 2009 242-264
© 2009 The Author(s)
10.1177/1558689809335079
http://jmmr.sagepub.com

Validating for Use and Interpretation

A Mixed Methods Contribution Illustrated

Linda Morell
University of California, Berkeley
Rachael Jin Bee Tan
Schroeder Measurement Technologies, Dunedin, Florida

Researchers in the areas of psychology and education strive to understand the intersections among validity, educational measurement, and cognitive theory. Guided by a mixed model conceptual framework, this study investigates how respondents' opinions inform the validation argument. Validity evidence for a science assessment was collected through traditional paper-and-pencil tests, surveys, and think-aloud and exit interviews of fifth- and sixth-grade students. Item response theory analyses supplied technical descriptions of evidence investigating the internal structure. Surveys provided information regarding perceived item difficulty and fairness. Think-aloud and exit interviews provided context and response processes information to clarify and explain issues. This research demonstrates how quantitative and qualitative data can be used in concert to inform the validation process and highlights the use of think-aloud interviews as an explanatory tool.

Keywords: *mixed methods research; validity; response processes; research methods; measurement; science assessment; think-aloud interviews; middle school science*

This research shows how a mixed methods approach can be implemented to gather evidence supporting different aspects of construct validity to inform and ultimately strengthen an overall validity argument regarding test development. The study addresses the integration of qualitative and quantitative data to form a validity argument for appropriate use and interpretation of a middle school science assessment.

The fields of human cognition and measurement are on critical paths of crossover and fusion. Cognitive theories allow for greater understanding of human thinking, and measurement models, along with technological innovation, allow for careful examination of theory-based cognitive interpretation. Validity studies highlight this intersection and provide a lens to better understand how people understand material while refining instruments designed to measure thinking, attitudes, knowledge, aptitudes, and/or other individual differences.

Furthermore, this study provides a practical example of how validity issues can be addressed in the integration of disparate methods in mixed research studies, a problem that Onwuegbuzie and Johnson (2006) identified as neglected in the literature thus far. "Validity" in quantitative research is relatively well-defined, with many different types of validity having been proposed in the literature (discussion in the next section); however, the term *validity* is ambiguous in qualitative research, with no agreed-on definition of the concept (Dellinger & Leech, 2007). This poses a problem for mixed methods research, because a new conception of validity must be created to reconcile the well-defined (quantitative) and ambiguous (qualitative) viewpoints of the term. By proposing and examining a new test validation process, achieved through the combination of quantitative and qualitative methodologies, the integration of disparate data sources provides a unique and powerful way of conducting test validation as well as an example of the successful integration of validity issues into mixed methods research.

Conceptions of Validity

Because of a sizeable increase in test and study validity research during the past century, the American Psychological Association (APA), the American Educational Research Association (AERA), and the National Council on Mathematics Education (NCME) revised the *Standards for Educational and Psychological Testing* (APA, AERA, & NCME, 1999). According to the *Standards*, a sound validity argument combines

> various strands of evidence into a coherent account of the degree to which existing evidence and theory support the intended interpretation of test scores for specific uses. . . . The validity argument may indicate the need for refining the definition of the construct, may suggest revisions in the test or other aspects of the testing process, and may indicate areas needing further study. (APA, AERA, & NCME, 1999, p. 17)

This conception of validity has been influenced by many researchers, including Samuel Messick (1989, 1993).

According to the *Standards*, there are some general sources of test validity evidence that can be collected to advance a validity argument. The *Standards* specify different aspects of validity; these aspects do not represent distinct types of validity (APA, AERA, & NCME, 1999). Evidence based on internal structure, test content, response processes, relations to other variables, and test consequences comprise the current conception of sources of validity evidence.

Internal Structure

According to the *Standards*, "analyses of the internal structure of a test can indicate the degree to which the relationships among test items and test components conform to the construct on which the proposed test score interpretations are based" (APA, AERA, & NCME, 1999, p. 13). Studies of the internal structure of tests could be designed to explore whether certain test items function differently for particular subgroups of students. An investigation of the internal structure of a given examination depends on how the exam would be used. For example, if the exam positioned test items or series of test components in order of increasing difficulty, evidence of the extent to which response patterns matched the conceived expectation would need to be gathered.

Test Content

Evidence based on test content could include empirical analyses of the adequacy with which the test content represents the proposed content domain and of the bearing that the content domain has on the proposed interpretation of test scores. This type of evidence can also come from expert judgments of the relationship between parts of the test and the construct. It is often discussed in terms of “content validity” or “logical validity.” Often in evaluation, a table of specifications is used to explicate the content under investigation. It is important to note that the *Standards* specify that when

> student mastery of a subject domain is tested for the purpose of summative decision-making or grade assignment (or matriculation) the framework for testing the student should be limited to information or procedures that a student has had an opportunity to learn through a specific curriculum. (APA, AERA, & NCME, 1999, p. 12)

Response Processes

Evidence gathered based on response processes could provide additional support to and bridge evidence based on test content. According to the *Standards*, “theoretical and empirical analyses of the response processes of test takers can provide evidence concerning the fit between the construct and the detailed nature of performance or response actually engaged in by examinee” (APA, AERA, & NCME, 1999, p. 12). Evidence based on response processes generally comes from a close examination of the individual respondent and his or her individual responses. For example, if an assessment is designed to measure mathematical reasoning, it becomes important to verify that mathematical reasoning is taking place as the student completes the exam and he or she is not using a standard algorithm (like always guessing “c” to multiple choice type questions) or other prop to answer questions.

Relation to Other Variables

This aspect of validity addresses questions about the degree to which relationships between validity evidence gathered for an assessment and evidence gathered regarding external variables (such as other tests hypothesized to measure the same constructs) are consistent with the construct underlying the proposed test interpretations. Evidence based on external factors takes many forms, including (a) convergent and discriminant evidence, (b) test–criterion relationships, and (c) validity generalization.

Consequence

It is important to distinguish between issues of validity and issues regarding social policy when considering the consequential aspect of construct validity. For example, consider the scenario that different hiring rates for members of different groups result from the use of a particular test. If the difference is because of an unequal distribution of the skills the test purports to measure, and if the skills are indeed important to successful job performance, then the finding of group difference does not necessarily imply any lack of validity for the intended inference. However, if the test measured skills that are unrelated to the actual job performance needed, then validity would be called into question. Evidences based on consequences that can be traced to either construct irrelevant variance or construct underrepresentation are legitimate issues to consider within the validation process. However, if evidence cannot be traced to construct irrelevant variance or construct underrepresentation it might be an issue of policy rather than validity (APA, AERA, & NCME, 1999).

Importance of Response Processes Validity Evidence

According to *Knowing What Students Know* (National Research Council, 2001), there is an increasing recognition with the assessment community that traditional forms of validation emphasizing consistency with other measures, as well as the search for indirect indicators that can show this consistency statistically, should be supplemented with evidence of the cognitive or substantive aspect of validity.

Using a mixed methods framework, this can be examined.

Mixed Methods Evaluation

Whether to use quantitative or qualitative methods has been a consistent controversy within the evaluation community. Much of the debate stems from evaluator subscription to a particular paradigm (Greene & Caracelli, 1997). The ‘‘incompatibility thesis’’ (Howe, 1988) is a direct result of this debate, which states that qualitative and quantitative research paradigms cannot and should not be mixed (Johnson & Onwuegbuzie, 2004). Those who adhere to the incompatibility thesis are considered purists, believing that the attributes of a paradigm are inseparable, and different paradigms embody incompatible assumptions about the nature of the world and what is important to know (Greene, Caracelli, & Graham, 1989).

In contrast to the purist stance, the dialectical and pragmatic positions posit that although competing paradigms give rise to contradicting ideas, these differences are not necessarily irreconcilable. The dialectical perspective seeks to honor and respect the different ways of making knowledge claims by maintaining the integrity of different paradigms when they are combined in one research study (Greene & Caracelli, 1997; Hanson, Creswell, Plano Clark, Petska, & Creswell, 2005). Pragmatists believe that methods can be mixed and matched to best answer the research questions (Hanson et al., 2005; Johnson & Onwuegbuzie, 2004). This study uses a pragmatic lens by allowing the research question to dictate which methods are most appropriate.

This study was developed within a mixed methods framework, as was requisite to fully examine each research question. Cognitive theories suggest that response processes can differ depending on the nature of the cognitive demands placed on the subject, so multiple data collection modes should be used to gain a more complete understanding of the thinking processes and performance (Taylor & Dionne, 2000). From conception through design and implementation, a mixed method design was necessary because of the complexity of the study, the variety of data sources, and types of data required.

The research design has a complementarity purpose, as described by Greene et al. (1989). Both methods measure overlapping, but yet different aspects of the same phenomenon—test validity. The methods were chosen with the intent that they would have complementary strengths and nonoverlapping weaknesses. The independent use of each method to assess the phenomenon is necessary to ensure a fair interpretation of the findings (Campbell & Fiske, 1959; Greene & McClintock, 1985). The intent was to use distinct methods to elaborate understanding of test validity, while implementing compatible methods that can corroborate findings and also balance biases present in any one method.

Method

A parallel mixed methods design (Taskakkori & Teddlie, 2003) was used to collect validity data. Each type of data was collected independently and integration, explanation, and interpretation occurred after each source was analyzed separately. In the literature, this has also been referred to as a concurrent triangulation design (Hanson et al., 2005), where integration usually occurs only at the data interpretation stage. Such designs are particularly useful for exploring whether different sources of evidence corroborate study findings. If disparate methods, which were not integrated until the end of the study, provide similar conclusions about the data, then this affords stronger evidence for those conclusions.

This research provides examples of how validity evidence based on response processes and evidence based on internal structure work in concert to inform the validation argument. At levels above classroom instruction, it is often the case that students contribute to the validity process by responding to knowledge (and/or procedural) questions only. This research investigates how students contribute to the validation process beyond that of test taker. Specifically, this research was guided by the following research question: How does evidence based on response processes inform the validity argument?

Participants

Fifth- and sixth-grade students in two public school districts took a paper-and-pencil test followed by a survey that asked questions about the test that was just taken. There were four classes per school district for a total of eight classes. For the think-aloud and exit interviews, teachers selected the students who participated. Thirty-four students participated in the think-aloud and exit interviews. Students from each of the eight classes participated in the think-aloud and the exit interviews. Out of a total of 230 study participants, 114 were in the fifth grade and 116 were in the sixth grade. There were 107 female and 123 male students.

Data Sources and Procedure

Validity evidence was collected from a variety of sources. Data were gathered through the Assessing Science Knowledge (ASK) research project at the Lawrence Hall of Science (LHS). The ASK project was funded through the National Science Foundation to develop an assessment system for the LHS Full Option Science System (FOSS) science curriculum for Grades 3 to 4 and 5 to 6. The goal of the ASK project is to provide valid and reliable information regarding student understanding of science (LHS, 2003). The FOSS curriculum was developed at LHS and uses a ‘‘hands-on’’ approach to teaching science, where students are not only the recipients but also creators of knowledge. The FOSS curriculum engages students in laboratory investigations, multimedia interactions, data processing, discourse, reading, writing, and assessment (LHS, 2003).

In collaboration with LHS project staff, the earth science module titled "Environments" was chosen to supply information for this research. The Environments module is designed to teach students that all living things have dependent relationships with their environments. Students learn about the relationship between one organism and its environment and what limitations the environment places on the living organism. With this knowledge comes an understanding that all living things have limitations, and changes to their environment can be difficult on organisms. The expectation of students is that they will gain an understanding of the basic concepts in environmental biology (FOSS, 2000).

The instruments used to collect data include (a) a paper-and-pencil test administered to students, (b) a paper-and-pencil survey administered to students, and (c) a think-aloud interview along with an exit interview conducted with students. Both qualitative and quantitative data were collected.

Paper-and-Pencil Test Administered to Students

The ASK team at LHS in collaboration with the Berkeley Evaluation and Assessment Research Center, revised the earlier Environments module items based on anecdotal feedback from teachers and item response theory (IRT) analyses of items administered to students through the ASK project during the previous year.

A 12-item posttest was administered at the end of the instructional module. This posttest contained five open-ended items, five fill-in-the-blank items, and two multipart items (fill-in-the-blank and open-ended items scored together).

Paper-and-Pencil Student Survey

The paper-and-pencil student survey contained 33 items and was administered to students immediately after the posttest. The survey was divided into three parts. The first part asked the students to estimate each posttest item's difficulty on a 3-point scale (*easy*, *medium*, and *hard*). In the second part, students were asked to rate the fairness of the items on a scale of *very unfair*, *unfair*, *fair*, and *very fair*. The third part of the survey asked students which questions were easiest and hardest and why, what subjects were taught in class but not reflected on the posttest, what question they would write if they were developing test items, and which question(s) they anticipated seeing on the test. All items on this last part were open ended.

Think-Aloud and Exit Interviews

Based on the work done by Ayala, Yin, Shavelson, and Vanides (2002), Taylor and Dionne (2000), and Leighton (2004), a think-aloud protocol and exit interview were developed and used to collect data from students. The think-aloud interview is a technique in which a test taker verbalizes all thoughts while completing an exam, so an interviewer can gain insight into the otherwise hidden thought processes of the examinee. The interviewer does not interfere with the testing process, and the role of the interviewer is only to keep the examinee talking while answering the test questions. Minimal prompting was provided in the form of “keep talking” and “I’m listening” as the student went about answering each question. The think-aloud interview was followed by an exit interview (retrospective debriefing) in which issues raised during the think-aloud interview were addressed. An interviewer may choose to conduct retrospective debriefing if they felt that examinees did not provide enough verbalizations during the think-aloud interview or if the interviewer wanted clarification on some aspect of the verbalizations. The think-aloud protocol and exit interview were developed and pilot tested with three students at a private Bay Area elementary school and with two graduate students. The collection of information was guided by the strict protocol developed from the piloting sessions. Audiotapes and transcriptions confirmed the methodology with which the sessions were conducted. Based on the pilot studies, the original think-aloud protocol and exit interview were reduced to fall within the teacher time-limit requirements and the Committee for the Protection of Human Subject specifications.

Informing the Validity Argument

One source of validity evidence came from the use of IRT analyses, used to determine item difficulty of the posttest. Although IRT models have existed for more than half a century, they are just recently being applied in assessment outside the large-scale testing arena. IRT provides a mathematical model to organize the way in which test responses exhibit latent traits. The information provided using this methodology is “useful when developing, evaluating, and scoring items” (Harvey & Hammer, 1999, p. 353). In comparison with classical test theory (CTT), the IRT measurement model is expressed at both the instrument and the item level and not just at the instrument level, making IRT a sample-independent measure, whereas CTT results cannot be generalized beyond one specific testing situation. In addition, IRT “focuses attention on modeling the probability of the observed responses rather than modeling the responses, as true score theory” (Wilson, 2005, p. 90).

The Environments items were developed to elicit the hierarchical and cumulative nature of knowledge or understanding of the elements and levels within each subelement proposed by the theoretical model (the Environments Progress Map). The partial credit model (PCM) is an analysis model in the IRT family and is built from Rasch’s (1960) dichotomously scored items methodology. This model can be applied to data that have an ordinal-level structure. Therefore, this model is appropriate for analyzing the Environments module items. A unidimensional IRT analysis using the PCM was performed to analyze the test items. Wright and Masters (1982) provide the formula necessary to execute the PCM. The following expression shows the general formula that was followed for this study:

$$\pi_{\mathrm{nix}} = \frac{\exp \sum_{j=0}^{x} (\beta_n - \delta_{ij})}{\sum_{k=0}^{mi} \exp \sum_{j=0}^{k} (\beta_n - \delta_{ij})}, \quad x = 0, 1, \ldots, m_i.$$

The observation x in the above equation is the count of the completed item steps (the number of total response categories minus one). The numerator contains only the difficulties of these x completed steps, $\delta_{i1}, \delta_{i2}, \ldots, \delta_{ix}$. The denominator is the sum of all $m_i + 1$ possible numerators. The formula and description were taken directly from Wright and Masters (1982, p. 43).

Using this model allows for an investigation of the measurement properties of the test by estimating the person ability estimates, item difficulty estimates, item fit statistics, and person fit statistics. Person ability (β) is interpreted as a person's knowledge of the information presented through the Environments module within the academic context. Item difficulty (δ), is interpreted as the degree to which individuals choose the correct answer to an item. This model is useful because the analysis calibrates test-free person estimates (or person ability = β) and sample-free item estimates (or item difficulty = δ). This can be done because raw information (item scores and person estimates) can be transformed into logits, yielding item scores and person ability estimates on a logit scale. Conducting this logistic transformation on the data places them on a new scale free from the particulars of persons or items (Wright & Masters, 1982), making IRT a sample-free measurement model. The analysis for this study was completed using ConQuest software.

Another source of validity evidence was the student responses to the surveys. Surveys administered to students immediately after they completed the posttest asked students to rate the difficulty and fairness of each posttest item, and then answer nine open-ended questions about the posttest items. Student survey results were categorized and coded into typologies (or ideal types as opposed to taxonomies, which completely classify something into mutually exclusive and exhaustive categories; Patton, 2002). The qualitative analysis was conducted as needed determined by whether explanations or understanding of IRT results or quantitative results were necessary.

The think-aloud and exit interview responses provided another source of validity evidence for the posttest. The qualitative data analyzed for this research were based on Taylor and Dionne's (2000) concurrent verbal protocols and retrospective debriefing analysis method as well as on other researchers' work in this area (Ayala et al., 2002; Boren & Ramey, 2000; Ericsson & Simon, 1993; Greeno, 1980; Hamilton, Nussbaum, & Snow, 1997; Leighton, 2004; Tzuriel, 2000). These methods are less used in some areas of human studies research than others. For example, Jacob Neilson (1993) suggests that "thinking aloud may be the single most valuable" method implemented in usability testing (p. 195). However, doubts about the trustworthiness and accuracy of data collected via think-aloud and exit interviews persist (Leighton, 2004). For this reason, the methodology used for this study is described in detail here.

Concurrent verbalizations (think-aloud interviews) and retrospective debriefings (exit interviews) were conducted with fifth- and sixth-grade science students. Both of these data types were analyzed using the procedures described here. Unlike some data analysis, this qualitative data analysis began before the data were collected. Before students were interviewed, a task analysis was completed for each item and a uniform protocol was developed and pilot tested before the actual study data were collected (Ayala et al., 2002; Greeno, 1980). Generally, the task entailed a student reading a test item aloud, talking through his or her response process, and then recording his or her response on the test paper provided by the interviewer. Four features were considered while defining the task requirements for each item. First, the task demands were identified. What did the student need to do to complete the task? Second, the cognitive demands were identified. What did the student need to know to complete the task? Third, the item openness was defined. What was the degree of flexibility that the student could use to respond to the item (lower for multiple-choice items and higher for constructed-response items)? And finally, the difficulty of the item was estimated. How easy or difficult is the item with "respect to material that does not pertain to the construct (reading level, use of jargon)?" (Ayala et al., 2002, p. 6). Each item on the posttest required a constructed response from the student. According to Ericsson and Simon (1993), this type of response is ideally suited for think-aloud examination. The items are of medium difficulty for fifth- and sixth-grade students from a cognitive processing perspective. The item format required that students draw on strategies and knowledge chosen consciously. This allows the student to be aware of his or her thinking process but does not overwhelm the thinking process so that verbalization shuts down.

A uniform protocol was followed for each student interview. Each time the same interviewer introduced the study and provided a warm-up exercise to make the student comfortable with the study setting and requirements. After completing the warm-up exercise, the student was asked to complete the task as previously described. Based on past think-aloud research (Morell, 2008), prompting was minimal and consisted of "keep talking" or "I'm listening." The think-aloud protocol asked students to read each posttest item aloud, verbalize a response to the item, and then write the response. The exit interview was much the same as the student survey. Initially, the exit interview protocol required the interviewer to clarify think-aloud responses. Then students were asked to rate the difficulty and fairness of each posttest item aloud to the interviewer and then answer open-ended questions about posttest items. After the interview was completed, each think-aloud/exit interview was transcribed verbatim, including pauses, emphases, and tones. Next, a coding grid was constructed. The coding grid was based on the scoring guides developed at the Berkeley Evaluation and Assessment Research Center to analyze student posttest responses. The grid was used to score each student's qualitative response to each test item.

A comparison between (a) survey information and (b) think-aloud/exit interview information was conducted to provide further validity evidence. The investigation was conducted to identify where data sources overlap and what the unique contribution(s) of each source is to the validation process. A correlation between student survey responses and think-aloud results was executed.

Based on findings from the IRT analysis of student test data, a back-and-forth between data sources occurred in assembling the validity argument. Instead of guessing why so many or so few students answered an item correctly or incorrectly, as would be the case when examining IRT results in isolation, an investigation regarding the problematic items was conducted by incorporating the qualitative data into this portion of the analysis. This investigation included reviewing the relevant data sources to help accomplish the following goals:

1. Explain the specific issue
2. Help revise/refine the item
3. Help clarify the construct

A final analysis was conducted comparing think-aloud ability estimates and IRT results. For this analysis, students were assigned ability estimates based on their think-aloud sessions. These estimates were loosely based on the categories identified on the construct (progress) map. IRT results of selected students (i.e., those who participated in the think-aloud interviews) were identified, which provided a quantitative ability estimate for each think-aloud student. The two ability estimates were then correlated to see whether the quantitative and qualitative data provided agreement about student proficiency on the posttest. A qualitative discussion about the differences between the categories generated by the think-aloud data and those of the progress map is presented in the next section.

Results

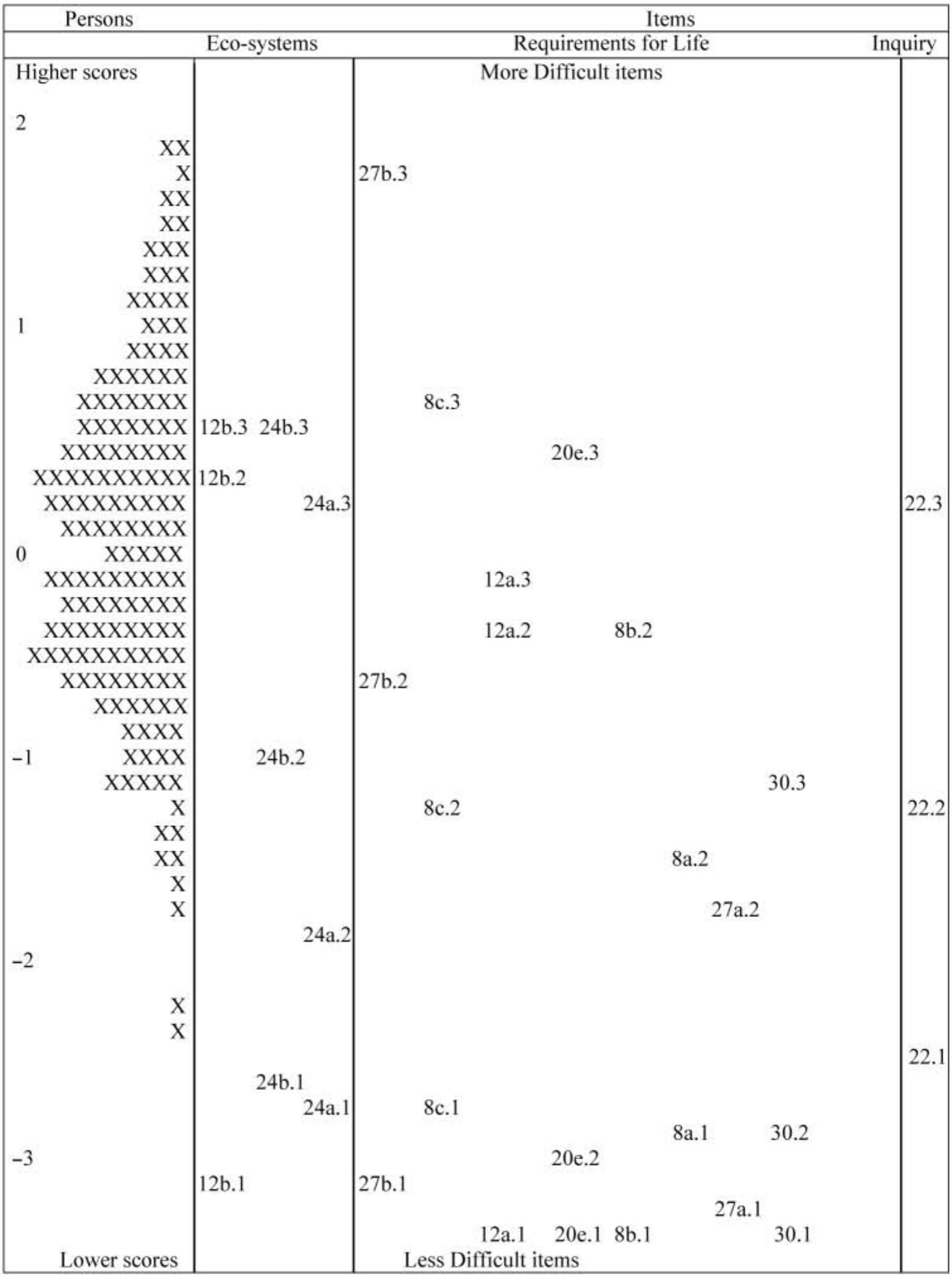
Of fifth- and sixth-graders participating in the Environments module, 216 completed the paper-and-pencil posttest. Of those students, 207 filled out the accompanying paper-and-pencil survey and were eligible for the study (to be eligible, a student's parent or guardian was required to provide written informed consent to allow the child's data to be used). Think-aloud and exit interviews were conducted with 35 students at the completion of the module. The 35 students interviewed were from eight different public schools in two cities. Data gathered from these groups are presented to address the study's research question.

Internal Structure Evidence: Contribution to the Validation Argument

The research question for this study necessitated a mixed method approach be taken; multiple sources of evidence needed to be collected in different ways to capture a richer understanding of how the assessment functioned. Both qualitative and quantitative data were collected through these instruments. Results from a sample of items will be presented by data source. As a starting point, it is necessary to present the quantitative analysis of posttest data. First, the partial credit results will be presented and then other analyses of quantitative and qualitative data will illuminate issues affecting the validity argument.

Along with depicting a visual representation of person abilities and item difficulties on a common scale, a Wright map (Figure 1) provides a graphical representation of the construct map's levels. In FOSS theory, the levels specified on the construct map include the following: notions (the lowest level of understanding, which often includes misunderstandings), recognition (the next level, which suggests that students know or can memorize facts), emerging relationships (the third of five levels in which students show simple signs of relational thinking regarding the environment), constructing relationships (the fourth level defined on the construct map in which students demonstrate an understanding of the relationships between and among environmental factors), and knowledge integration (the final level which fifth- and sixth-grade students generally do not reach through the Environments module). A student at the knowledge integration level would have a complete understanding of the environment as a dynamic system including different components and how more or less of different elements affect multiple aspects of the environment. The Wright map in Figure 1 displays the difficulties of all the Environments module posttest items, separated according to whether they test students' knowledge of "ecosystems," "requirements for life," or "inquiry." The numbers on the far left of the map are logits, which represent the log odds that a student will answer an item correctly at a given

Figure 1
Latent Distributions and Thresholds for the Environments Module Posttest Items



threshold; it provides a constant scale on which to compare student ability and item difficulty. Each “X” represents 1.4 cases or students. The students’ placement on the Wright map indicate their ability in logits; students closer to the bottom of the map are less proficient in the posttest material, whereas students closer to the top of the map are more proficient in the material.

Items appear to the right of the student distribution, organized by item threshold (represented by the item number followed by a “.1,” “.2,” or “.3”)—their placement on the Wright map indicates their difficulties in logits; items closer to the bottom of the map are easier for students, whereas items closer to the top of the map are more difficult to answer correctly. If an “X” appears above a threshold, then that student has greater than a 50% chance of answering the item at that threshold. If an “X” appears below a threshold, then that student has a less than 50% chance of answering the item correctly. If an “X” appears next to an item threshold that indicates the student has a 50% chance of answering the item at this threshold; this is the most desirable scenario because they provide the most information about student ability. These thresholds represent logit values as the 50% chance of scoring at a given level or below and at the next level or above. In Figure 1, the “.1” thresholds represent the point at which students have a 50% chance of scoring at the recognition level and above versus the notions level; the “.2” thresholds represent the point at which students have a 50% chance of scoring at the emerging relationships level and above versus the recognition level or below; and the “.3” thresholds represent the point at which students have a 50% chance of scoring at the constructing relationships level versus all the lower levels. Threshold “3” of Item 27b (the point at which it becomes more likely that the student will answer this question at the constructing relationships level) is the hardest to reach at almost 2 logits. Average student ability is around Threshold 3, the constructing relationships level, which is expected of students taking the posttest who have completed all instruction.

It should be noted that the Wright map can be linked to the FOSS theory of construct map levels. The scoring guide for each item was created so that the lowest score reflected a student response at the notions level, the next lowest score reflected a student response at the recognition level, and so on. When examining the Wright map, all the “.1”s represent item thresholds at the notions level, the “.2”s represent item thresholds at the recognition level, the “.3”s represent item thresholds at the emerging relationships level, and the “.4”s represent item thresholds at the constructing relationships level. As can be seen in Figure 1, there appears to be a “banding” or clustering together of the different thresholds. This provides support for the FOSS theory that there are distinct levels of student understanding. Furthermore, student “X”s next to each threshold cluster show which level of understanding these students have achieved regarding the Environments module content.

From this map, it can be seen that some of the thresholds are closer together and some of the thresholds are farther apart. This information is important because it can inform the cognitive theory embedded in the scoring guide and indicate points that need further investigation and attention. If the cognitive theory used to develop the outcome categories represented the developmental ranges accurately it would be anticipated that the thresholds would be fairly evenly spaced with clear gaps between each level. A large gap between thresholds for one item usually means there is a small threshold gap somewhere else for that item. It is helpful to have a variety of data sources so that when unusual threshold gap patterns are encountered, data can provide insight as to why there is a gap and where the problem is—there could either be an issue with the theory or the scoring guide.

An examination of the Wright map in Figure 1 reveals that several of the Environments module posttest items need closer examination and explication. If the cognitive theory used to score the data functions correctly, it would be expected that the item thresholds would be spaced fairly equally. As can be seen in the first item, Item 12b, the thresholds are not equally spaced. The space between 12b.1 and 12b.2 is quite large, whereas the gap between 12b.2 and 12b.3 is quite small. In addition, based on the underlying constructs from which items were written, all the first thresholds (.1), second thresholds (.2), and third thresholds (.3) should consistently appear in the same general area of the Wright map across items forming bands.

As Figure 1 shows, several items (12b, 12a, 20e, and 30) are in need of further investigation. Item 12a shows the same issue as Item 12b. The threshold between 12a.1 and 12a.2 is large, whereas the gap between 12a.2 and 12a.3 is small. Item 20e also shows a different pattern. The threshold between 20e.1 and 20e.2 is small, whereas the space between Item 20e.2 and 20e.3 is large. The threshold gap between 30.1 and 30.2 is small, whereas the gap between 30.2 and 30.3 is large. In addition, all thresholds for Item 30 are low, indicating that the item may be too easy for students and not aligned to the underlying construct. Item 12b (at threshold 12b.2), Item 20e (at threshold 20e.2), and Item 30 (thresholds 30.2 and 30.3) are interesting because they have thresholds that fall outside their anticipated logit ranges. As this research suggests, the Wright map clearly identifies the item thresholds and clarifies which items and scoring levels need scrutiny.

Partial Credit, Survey, and Interview Results

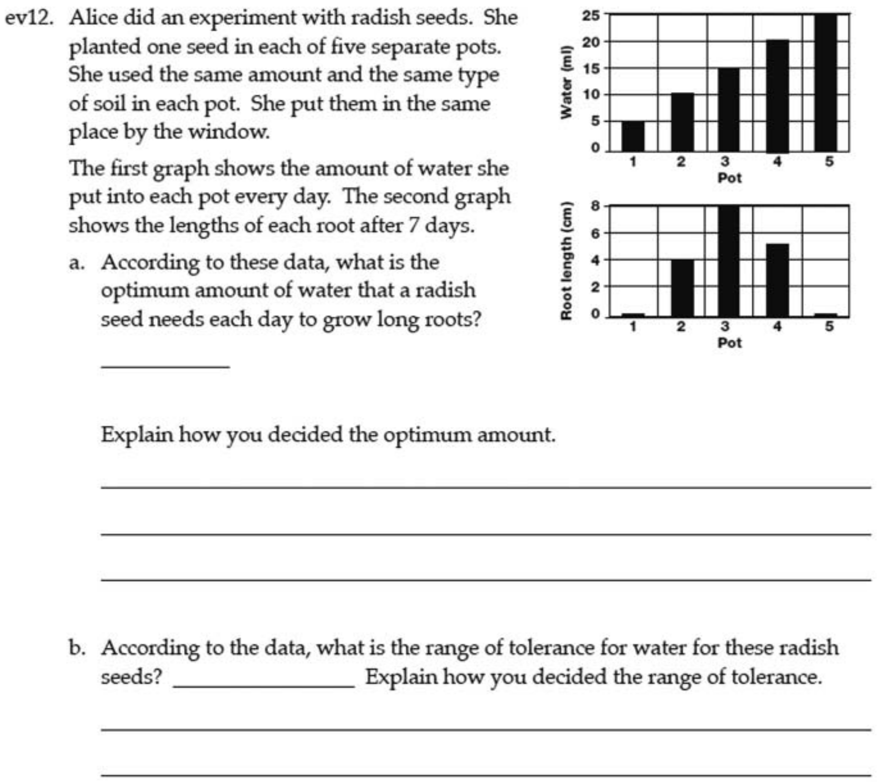
Based on the partial credit analysis, Items 12b, 12a, 20e, and 30 needed further investigation because their thresholds were not consistently spaced according to the developmental theory. Specifically, the second threshold of Items 12a and 12b appeared too difficult, and the second threshold of Items 20e and 30 appeared too easy. These conclusions were drawn based on the IRT partial credit analysis, surveys, and think-aloud and exit interview results.

According to the data sources, Item 12b was one of the more difficult items appearing on the posttest. Items 12a and 12b required specific knowledge about the concepts of optimum amount (for 12a) and range of tolerance (for 12b). Students had to interpret two graphs to arrive at an answer and then they had to justify that answer with an explanation.

Item 12b asks, “According to the data, what is the range of tolerance for water for these radish seeds? __________ Explain how you decided the range of tolerance.” To complete this item successfully, students need to interpret two graphs to arrive at an answer and then explain their answer. Figure 2 shows the item as it originally appeared on the test.

The scoring guide awarded the maximum amount of credit for an answer of “10 ml to 20 ml” with an adequate explanation, such as “the radishes grew in Pots 2 to 4, but did not grow in Pots 1 and 5. Pots 2 to 4 got between 10 and 20 ml of water.” Although the student survey analysis identified this item as one of the more difficult ones, it did not clarify why it was more difficult. However, the think-aloud and exit interviews identified one

Figure 2
Middle School Science Posttest Items 12a and 12b, Original Format



ev12. Alice did an experiment with radish seeds. She planted one seed in each of five separate pots. She used the same amount and the same type of soil in each pot. She put them in the same place by the window.

The first graph shows the amount of water she put into each pot every day. The second graph shows the lengths of each root after 7 days.

a. According to these data, what is the optimum amount of water that a radish seed needs each day to grow long roots?

____

Explain how you decided the optimum amount.

____

____

____

b. According to the data, what is the range of tolerance for water for these radish seeds? ________________ Explain how you decided the range of tolerance.

____

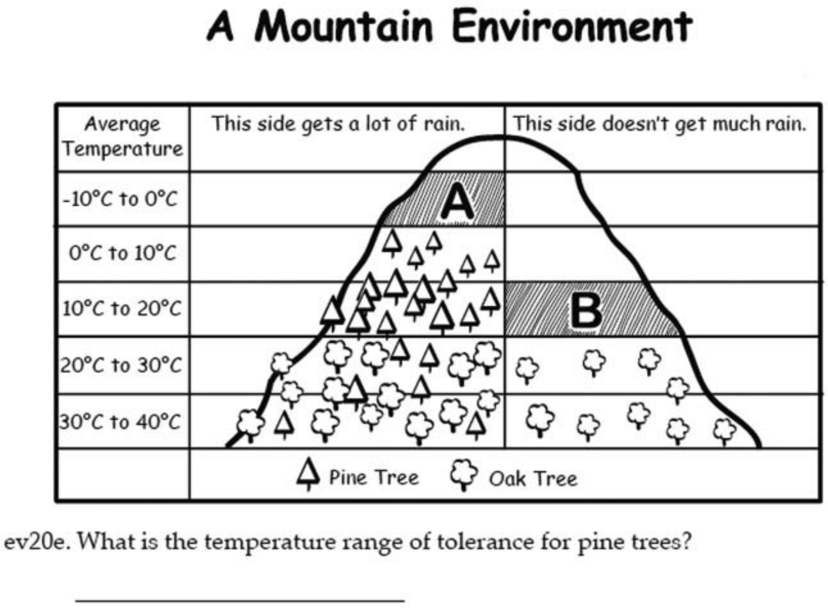
____

issue with the item. That is, students misread the graph and included 5 and 25 ml within the range of tolerance (because the roots in Pots 1 and 5 showed very small amounts of growth). From the think-aloud and exit interviews, it was also noticed that students read the second graph but failed to transfer information from that graph (root length by pot) to the one above (water by pot) to arrive at the correct answer. The scoring guide’s Level 2 disallowed for responses that were not provided in milliliters. Students who responded that Pots 2 to 4 grew, whereas Pots 1 and 5 did not received a Level 1 score. Although this is a correct interpretation of the second graph, it fails to relate the information back to the watering graph provided first.

Item 20e showed a problematic second threshold as well. This item’s second threshold was easier than anticipated based on the partial credit analysis. All three student data sources (student surveys, think-aloud and exit interviews, and the partial credit analysis) indicated that Item 20e was easy. Item 20e asked students to identify the range of tolerance for pine trees based on the picture (see Figure 3 for the item as it originally appeared on the test).

According to student surveys, this item was easy “because it gave you the picture” or “you just needed to read off the picture.” These responses were somewhat vague and

Figure 3
Middle School Science Posttest Item 20e, Original Format



A Mountain Environment

ev20e. What is the temperature range of tolerance for pine trees?

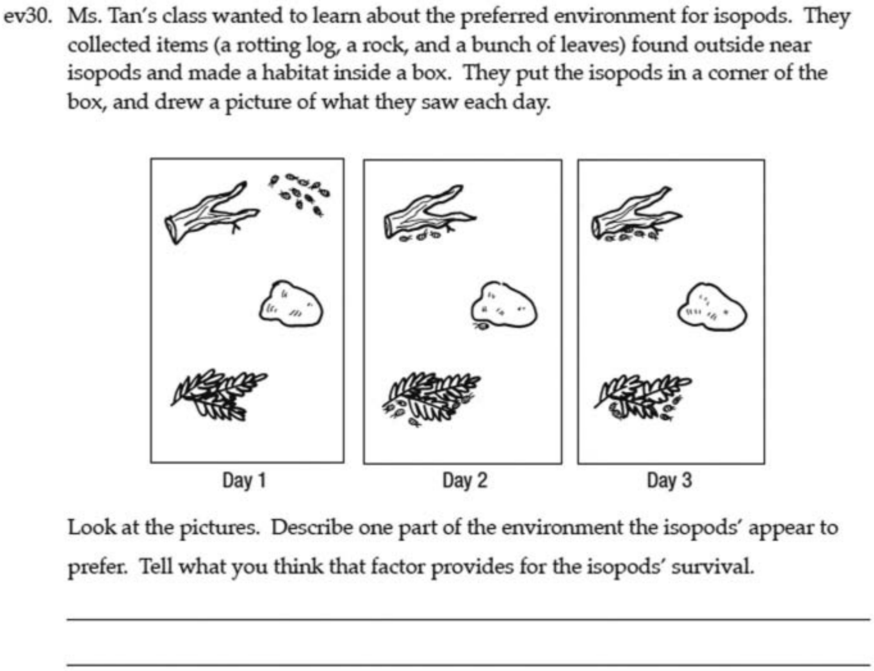
____

provided little insight into why the second threshold was so easy to reach. The think-aloud and exit interview analyses identified why it could have been so easy for students to reach the second threshold. According to the scoring guide, a Level 3 response would be "0°C to 40°C." A Level 2 response was defined as "writes any partial range of tolerance included within the range of 0°C to 40°C." According to think-aloud and exit interview findings, many students chose one 10° band instead of identifying the entire range. Based on this information, it is recommended that the picture be changed to list only one temperature at each mountain level. Therefore, students would not confuse bands of temperature ranges with the range of tolerance. Students thought that they needed to pick the band where most of the pine trees grew instead of the beginning and ending temperature where pine trees grew.

Item 30 exhibited the same problematic low second threshold as Item 20e, and also an unusually low third threshold. This indicates that the thresholds were too easy to reach according to the partial credit analysis. Item 30 was a straightforward item that asked students to read the question, examine three pictures provided, and write a response to the question about which environment isopods appear to prefer and what does the environmental factor provide for the isopods survival (see Figure 4 for the item as it originally appeared on the test).

Results of the student survey and partial credit analyses were similar indicating that this item was easy. Although the partial credit analysis merely revealed that the threshold was

Figure 4
Middle School Science Posttest Item 30, Original Format



ev30. Ms. Tan's class wanted to learn about the preferred environment for isopods. They collected items (a rotting log, a rock, and a bunch of leaves) found outside near isopods and made a habitat inside a box. They put the isopods in a corner of the box, and drew a picture of what they saw each day.

Day 1 Day 2 Day 3

Look at the pictures. Describe one part of the environment the isopods' appear to prefer. Tell what you think that factor provides for the isopods' survival.

____

____

too easy to reach, the think-aloud and exit interviews provided evidence as to why this was the case. It was too easy because students could have no idea about the answer and guess anything but the rock and receive a Level 2 score. Students could say either "the rotting log and/or leaves" or "they provide moisture, shelter, protection, or food" and get a Level 3 score. To make this item more difficult, it would need to be completely revised. One teacher mentioned that this item had a "practical validity" problem because students commented to her that they found isopods underneath rocks all the time, and they (the rocks) provided moisture, shelter, and protection.

The think-aloud and exit interview analyses indicated ways in which the scoring guide for Item 12a (see Figure 2) could be improved. Level 2 for Item 12a (just like the scoring guide for Item 12b) disallowed for responses that were not provided in milliliters. However, some students provided answers in terms of "pots" (correctly) not "milliliters." The interpretation of the second graph is correct but fails to incorporate the information back to the water graph provided first. To receive a score above Level 1 (attempts to answer but takes wrong focus), a student must give their answer in terms of "pot" instead of "milliliter" as these students understand how to read graphs at the recognition level and are not at the notions level. These students failed to use the two-step analysis process required to receive full credit.

Student Survey Results Compared With Think-Aloud and Exit Interview Responses

Between 200 and 206 students provided responses to difficulty and fairness ratings for posttest items through the surveys administered after the test, and 35 students provided this information during exit interviews. Information obtained from these two sources of data overlapped in two content areas (difficulty ratings and fairness ratings). The correlation between the student opinions of item difficulties on each instrument was quite high, $r(12) = .770$, $p < .01$. Although not as highly correlated as item difficulty opinions, the item fairness correlation was still high, $r(12) = .664$, $p < .05$.

Students reported information on the same scales for each of the posttest items. Item difficulty ratings for Items 30, 22, 20e, 8a, 27a, 12b, and 24a were similar (e.g., within 0.05 units of each other). However, difficulty ratings for Items 8b, 8c, 27b, 12a, and 24b differed (e.g., were greater than 0.05 units different) between data sources. Exit interview results for Items 24b, 24a, and 8c indicate that students rated the items as more difficult than they did on the survey. The difference between difficulty ratings on the exit interview and survey were most pronounced with Item 8c: 57% of survey takers said Item 8c was easy, 36% said it was of medium difficulty, and 7% said it was hard; 27% of interviewed students said Item 8c was easy, 56% said it was of medium difficulty, and 18% indicated that it was hard. Of all the posttest items, results regarding Item 8c differed most sharply between the two data sources. It could be speculated that this difference was due to students identifying the word ‘‘viable’’ as difficult to understand during the exit interview.

The beginning of the exit interview provided an opportunity for the researcher to follow up with questions that might have arisen during the think-aloud protocol. This opportunity did not exist within the student survey administration. Important information was obtained from asking students to clarify and explain their thinking behind item responses. For example, Item 8a asked students to use a graph to determine the optimum number of spoons of salt used to hatch the largest number of brine shrimp. Most students were able to read the graph correctly and arrive at the correct response of 5 (1-ml) spoons of salt. However, one student arrived at the correct conclusion using incorrect means. During the think-aloud interview, it was observed that the student simply counted the number of vertical sections present on the highest bar. During the exit interview, the interviewer asked the student to review Item 8a and explain how she arrived at 5 ml spoons of salt. The student explained that she counted the highest bar’s sections and got 5 (sections). The correct answer happened to be 5 ml spoons of salt and coincidentally, that bar had five sections or blocks. This exemplifies the benefit of using think-aloud interviews with students; the written answer alone indicated that the student understood how to read from the graph, but the exit interview and survey revealed that the student did not.

Think-Aloud Ability Estimates With Partial Credit Model Results

To compare empirical student ability estimates generated through the partial credit model, it was necessary to estimate student abilities based on think-aloud and exit interviews. The ability estimates based on think-aloud and exit interview sessions were loosely created from the categories identified on the construct (progress) map—notions, recognition, emerging relationships, and constructing relationships.

An analysis of the IRT results and think-aloud ability estimates reveals a high correlation, $r(32) = .846$, $p = .01$. IRT ability estimates ranged from −1.19 to 1.74 logits. IRT estimates of students placed at the notions level ($n = 2$) by think-aloud interviews ranged from −1.19 to −1.04 logits from the partial credit analysis. IRT estimates of students estimated to be at the recognition level by think-aloud interview responses ($n = 6$) ranged from −0.74 to 0.02. Students categorized at the emerging relationships level by think-aloud interview responses ($n = 23$) had ability estimates between 0.18 and 1.74, whereas the student ($n = 1$) categorized into the constructed relationships level by think-aloud interview responses showed an ability estimate of 1.45. The fact that these ranges do not overlap confirms that there are distinct differences in knowledge among the levels.

Students receiving the lowest ability scores on both estimates showed confusion over one of the most important aspects of the Environments module. These students could not define or describe the difference between two key terms: range of tolerance and optimum conditions. Range of tolerance is the range (of a particular environmental factor) in which a living organism can exist. Optimum conditions are those most favorable to an organism's survival, growth, and reproduction. The relationship between the two is that the optimum condition for an organism is within the range of tolerance. These terms were problematic for students at the notions level. Students at this level often equated the terms or showed no understanding that they were different in some way. According to think-aloud and exit interviews, comments from students regarding these terms include "they mean the same thing" or "they just asked that question."

Students scoring at the recognition level showed a basic understanding of these terms. Items 8a and 12a asked students to identify the optimum condition or optimum amount of an environmental factor, whereas Items 8b and 12b asked students to identify the range of tolerance of an organism for an environmental factor. As observed during think-aloud sessions, students at this level often read the question and immediately defined the term. For example, a student read, "What is the optimum amount of water that a radish seed needs each day to grow long roots? Optimum, optimum the best condition . . ." After defining the optimum amount, students at this level reviewed evidence presented, identified the optimum condition or amount, and wrote the correct response. Then, the students at this level went on to read the "b" part, which asked about the range of tolerance and immediately defined that term and went about identifying the range of tolerance and recording the answer. Students at this level did not reflect back or mention "optimum conditions" after leaving the "a" items.

In comparing responses with think-aloud and exit interviews, a difference was observed between students at the recognition level and students at the emerging relationships level. Students at the recognition level read each item independent of other items encountered on the posttest. Students at the emerging relationships level would often link between terms present in items other than the one of most pressing interest. For example, when students at this level got to the "b" item, they would often comment, "Range of tolerance not optimum amount. Range of tolerance is bigger than optimum amount." Or they would say something such as, "Optimum was the best here (15 ml of water) and the range is from here (10 ml of water) to here (20 ml of water) because radishes grew roots here to here.” Students at this level took into account both terms as they answered the items and knew the difference between the two terms and showed signs of relating the two terms.

There were also anomalies observed between the results of the think-aloud interviews and IRT analyses. According to the think-aloud interviews, one student was placed in the constructed relationships level, although he did not display the highest ability estimate from the IRT analysis. This student’s ability level was estimated to be 1.45. One student had a higher ability estimate of 1.74. During the think-aloud and exit interview session, this student easily answered each question often detailing not only the correct response but rationale based on the given scenario and slight variations. Although this student provided elaborate and detailed information during the think-aloud and exit interview session, he gave a brief and vague response to Item 24b. This exemplifies another benefit of using think-aloud interviews in conjunction with IRT analyses; think-aloud interviews can reveal what students know about concepts that they are unable to express on paper.

Conclusions

This study provides examples of how evidence gathered to investigate different aspects of validity can be used to inform and contribute to the overall validity argument. Specifically, this question was asked: How can evidence gathered regarding respondents’ opinions inform the validation argument? This is a question that researchers in the areas of psychology and education have been exploring to try to understand the intersection among aspects of validity, educational measurement, and cognitive theory. Many researchers have contributed promising ideas to these areas (e.g., Ayala et al., 2002; Embretson & Gorin, 2001; National Research Council, 2001; Pellegrino, 1988; Snow & Lohman, 1989; Wilson, 2005). The study presented here builds on previous work in the field, and puts forth novel ideas to be studied and methodologies to be used. Angoff (1988) wrote that “construct validation is a process, not a procedure; and it requires many lines of evidence, not all of them quantitative” (p. 26). Based on this comment and others with a similar message, a mixed methods approach to this study was used to capture and maximize both quantitative and qualitative data types.

Furthermore, this research presents a practical example of how to integrate disparate methods in mixed research studies, an area within the field which is lacking. A new test validation process, combining one quantitative and several qualitative methodologies, is proposed that integrates disparate data sources in such a way as to contribute to the conversations about validity as related to both assessment use and mixed method integration.

A mixed methods approach for the study was necessary because no single data source could provide the range of data necessary to address the research questions. From the conception of the study to reporting study results, the mixed methods approach was used to provide the framework for planning, conducting, organizing, analyzing, and reporting the research findings. The approach wove in theories of cognition at various levels to provide a basis for questioning the student’s current role in assessment development, to provide a methodological rationale for conducting think-aloud interviews with fifth- and sixth-graders, and to interpret the Environments module construct and provide meaning or understanding to the variety of data sources used.

This research highlights how one aspect of validity evidence (evidence based on response processes) can inform the validation argument. In some cases, evidence based on response processes is collected, but often not presented or its value is minimized (see, e.g., Floyd, Phaneuf, & Wilczynski, 2005; Morell, 2002). This article also provides a methodology for exploring how response processes validity evidence can be collected and presented. A cornerstone of this research is the use of think-aloud and exit interviews to collect data.

Results from this study support the assertion that evidence based on response processes from multiple sources contributes to the appropriate use and interpretation of assessment information. Based on findings from the IRT analysis of student test data, a back-and-forth among data sources occurred in assembling the validity argument. Instead of guessing why so many or so few students answered an item correctly or incorrectly, more data were collected (implementing a different methodology) on items identified as needing further investigation based on the partial credit analysis.

Overall the evidence presented via partial credit analysis, think-aloud and exit interview analyses, and survey results combined to identify issues, help explain or understand technical problems, describe item characteristics, and identify possible solutions regarding the interpretation and use of the assessment information. The partial credit analysis provided a starting point in the validity investigation. It supplied technical descriptions of evidence regarding the Environments module construct that was not provided through think-aloud and exit interviews or surveys. Surveys provided information regarding the perceived item difficulty and fairness. Think-aloud and exit interviews provided context and information to clarify and explain issues. This methodology and process of gathering, analyzing, and establishing validity evidence is of great value in interpreting and understanding the different aspects of the concept as well as providing concrete evidence to ensure the quality of an assessment. Construct validation is multifaceted and requires multiple sources of evidence. These demands necessitate a mixed methods approach to ensure that an adequate validity investigation has been conducted.

This study has several important limitations, and findings should be considered in light of these limitations. First, regarding think-aloud and exit interview data collected, students were not randomly chosen to participate. The researcher asked teachers to choose students from his or her class that met the required criteria. The criteria for inclusion in the study were that each student's parent(s) were required to provide explicit informed consent to participate (in the form of a signed consent form), each student needed to be fluent verbally and in written English, a competent reader, and sufficiently comfortable with a stranger so that he or she could read, think, and respond aloud to test items while being audiotaped. The students chosen by teachers had higher ability levels than all students included in the quantitative portion of the study. Unfortunately, it was impossible to overcome these limitations when using think-aloud interviews. If students who were not fluent verbally and in written English were chosen to participate, poor test results could not be attributed to low ability in the Environments module—perhaps students were very proficient and not able to express themselves in an understandable way by the researcher. This is a drawback of assessment research, because all tests that require students to read and write are in some part a test of their knowledge of the English language.

Comparisons between and among different sources of data should continue. One of the most time-intensive aspects of this project was the collection, transcription, coding, categorization, analysis, and interpretation of think-aloud and exit interview data. Cost effectiveness, cost utility, and/or cost–benefit research should be conducted to determine the utility of both methods relative to common priorities. Research should also be conducted to identify the extent to which students can answer accurately the "why" questions when they are presented in a survey format.

Further research should be conducted to ensure that participants at a variety of age ranges can contribute to the validity process. Studies should be conducted with younger participants to see if they can provide validity evidence or if there are developmental hindrances for younger participants. Research should be conducted in other areas of science as well as other subjects. Research based in other areas such as mathematics, reading, and social studies should be conducted to provide breadth to these findings.

Although much of the research on testing occurs in schools, the validation process exemplified in this study is not limited to the realm of educational testing. Many other fields would benefit from the use of such techniques to ensure assessment validity. For example, within the field of licensure, a test can determine whether a person is allowed to practice within a profession. This application includes diverse vocations such as accounting, law, nursing, massage therapy, information security, or personal training. It is therefore of utmost importance that exams are subjected to a validation process because they have a direct impact on people's lives. An extensive study such as the one described is sometimes not feasible from a practical standpoint, but a small study, focusing perhaps on a few items that content experts have identified as essential to practice within the profession, would be invaluable to lending credence to the use of an exam and ensuring that those individuals passing the test and receiving a license are qualified to practice within their field.

This research suggests that evidence based on response processes can inform the validity argument by identifying troublesome items; describing what makes an item need attention; and why the items, issues, or larger context need additional attention. This study describes powerful methodologies that can be combined under a mixed model framework to strengthen the validity of any assessment. Investigating the issues of cost-effectiveness and data limitations on different age groups will help facilitate the application of the methodologies used in this research to other studies, making the extremely important test validation process more accessible.

References

American Psychological Association, American Educational Research Association, & National Council on Mathematics Education. (1999). *Standards for educational and psychological testing*. Washington, DC: American Psychological Association.

Angoff, W. H. (1988). Validity: An evolving concept. In H. Wainer & H. I. Braun (Eds.), *Test validity*. Hillsdale, NJ: Lawrence Erlbaum.

Ayala, C. C., Yin, Y., Shavelson, R., & Vanides, J. (2002, April). *Investigating the cognitive validity of science performance assessments with think-alouds: Technical aspects*. Paper presented at the annual meeting of the American Educational Research Association, New Orleans.

Boren, M. T., & Ramey, J. (2000). Thinking aloud: Reconciling theory and practice. *IEEE Transactions on Professional Communication, 43,* 261-278.

Campbell, D. T., & Fiske, D. W. (1959). Convergent and discriminant validation by the multitrait–multimethod matrix. *Psychological Bulletin, 56,* 81-106.

Dellinger, A. B., & Leech, N. L. (2007). Toward a unified validation framework in mixed methods research. *Journal of Mixed Methods Research, 1*(4), 309-332.

Embretson, S., & Gorin, J. (2001). Improving construct validity with cognitive psychology principles. *Journal of Educational Measurement, 38,* 343-368.

Ericsson, K. A., & Simon, H. A. (1993). *Protocol analysis: Verbal reports as data*. London: A Bradford Book, MIT Press.

Floyd, R. G., Phaneuf, R. L., & Wilczynski, S. M. (2005). Measurement properties of indirect assessment methods of functional behavioral assessment: A review of research. *School Psychology Review, 34,* 58-73.

Full Option Science System. (2005). *Environments posttest*. Berkeley: Lawrence Hall of Science, Regents of the University of California.

Greene, J. C., & Caracelli, V. J. (1997). Defining and describing the paradigm issue in mixed-method evaluation. In J. C. Greene & V. J. Caracelli (Eds.), *Advances in mixed-method evaluation: The challenges and benefits of integrating diverse paradigms* (pp. 5-18). San Francisco: Jossey-Bass.

Greene, J. C., Caracelli, V. J., & Graham, W. F. (1989). Toward a conceptual framework for mixed-method evaluation designs. *Educational Evaluation and Policy Analysis, 11,* 255-274.

Greene, J. C., & McClintock, C. (1985). Triangulation in evaluation: Design and analysis issues. *Evaluation Review, 9,* 523-545.

Greeno, J. G. (1980). Some examples of cognitive task analysis with instructional implications. In R. E. Snow, P. A. Federico, & W. E. Montague (Eds.), *Aptitude, learning and instruction: Vol. 2. Cognitive process analyses of learning and problem solving* (pp. 1-21). Hillsdale, NJ: Lawrence Erlbaum.

Hamilton, L. S., Nussbaum, E. M., & Snow, R. E. (1997). Interview procedures for validating science assessments. *Applied Measurement in Education, 10,* 181-200.

Hanson, W. E., Creswell, J. W., Plano Clark, V. L., Petska, K. S., & Creswell, J. D. (2005). Mixed methods research designs in counseling psychology. *Journal of Counseling Psychology, 52,* 224-235.

Harvey, R. J., & Hammer, A. L. (1999). Item response theory. *The Counseling Psychologist, 27,* 353-383.

Howe, K. R. (1988). Against the quantitative–qualitative incompatibility thesis, or, dogmas die hard. *Educational Researcher, 17,* 10-16.

Johnson, R. B., & Onwuegbuzie, A. J. (2004). Mixed methods research: A research paradigm whose time has come. *Educational Researcher, 33,* 14-26.

Lawrence Hall of Science. (2003). *Assessing science knowledge (ASK): Project summary* (Unpublished report). Berkeley: Regents of the University of California.

Leighton, J. P. (2004). Avoiding misconception, misuse, and missed opportunities: The collection of verbal reports on educational achievement testing. *Educational Measurement: Issues and Practice, 23,* 6-15.

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13-103). New York: Macmillan.

Messick, S. (1993). *Foundations of validity: Meaning and consequences in psychological assessment*. Princeton, NJ: Education Testing Service.

Morell, L. (2002). *Analysis of the Buck Institute for education's high school economics test*. Unpublished manuscript, University of California, Berkeley.

Morell, L. (2008). Contributions of middle grade students to the validation of a national science Assessment study. *Middle Grades Research Journal, 3,* 1-22.

National Research Council. (2001). *Knowing what students know: The science and design of educational assessment*. Washington, DC: National Academies Press.

Neilson, J. (1993). *Usability engineering*. Cambridge, MA: AP Professional.

Onwuegbuzie, A. J., & Johnson, R. B. (2006). The validity issue in mixed research. *Research in the Schools, 13,* 48-63.

Patton, M. Q. (2002). *Qualitative research and evaluation methods*. Thousand Oaks, CA: Sage.

Pellegrino, J. W. (1988). Mental models and mental tests. In H. Wainer & H. I. Brown (Eds.), *Test validity* (pp. 49-59). Hillsdale, NJ: Lawrence Erlbaum.

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests*. Copenhagen, Denmark: Danmarks Paedogogiske Institut.

Snow, R. E., & Lohman, D. F. (1989). Implications of cognitive psychology for educational measurement. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 263-331). New York: Macmillan.

Taskakkori, A., & Teddlie, C. (Eds.). (2003). *Handbook of mixed methods in social and behavioral research*. Thousand Oaks, CA: Sage.

Taylor, K. L., & Dionne, J. P. (2000). Accessing problem-solving strategy knowledge: The complementary use of concurrent verbal protocols and retrospective debriefing. *Journal of Educational Psychology, 92,* 413-425.

Tzuriel, D. (2000). Dynamic assessment of young children: Educational and intervention perspectives. *Educational Psychology Review, 12,* 385-435.

Wilson, M. R. (2005). *Constructing measures: An item response modeling approach*. Mahwah, NJ: Lawrence Erlbaum.

Wright, B. D., & Masters, G. N. (1982). *Rating scale analysis*. Chicago: Mesa Press.