CRESST REPORT 730

Terry P. Vendlinski,
Sam Nagashima,
Joan L. Herman

CREATING ACCURATE SCIENCE BENCHMARK ASSESSMENTS TO INFORM INSTRUCTION

OCTOBER 2007

National Center for Research on Evaluation, Standards, and Student Testing

Graduate School of Education & Information Studies
UCLA | University of California, Los Angeles

Creating Accurate Science Benchmark Assessments to Inform Instruction

CSE Technical Report 730

Terry P. Vendlinski, Sam Nagashima, and Joan L. Herman
CRESST/University of California, Los Angeles

October 2007

National Center for Research on Evaluation,
Standards, and Student Testing (CRESST)
Center for the Study of Evaluation (CSE)
Graduate School of Education & Information Studies
University of California, Los Angeles
301 GSE&IS, Box 951522
Los Angeles, CA 90095-1522
(310) 206-1532

This material is based upon work supported by the National Science Foundation under Grant No. ESI-0119790. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

CREATING ACCURATE SCIENCE BENCHMARK ASSESSMENTS TO INFORM INSTRUCTION[1]

Terry P. Vendlinski, Sam Nagashima, and Joan L. Herman
CRESST/University of California, Los Angeles

Abstract

Current educational policy highlights the important role that assessment can play in improving education. State standards and the assessments that are aligned with them establish targets for learning and promote school accountability for helping all students succeed; at the same time, feedback from assessment results is expected to provide districts, schools, and teachers with important information for guiding instructional planning and decision making. Yet even as No Child Left Behind (NCLB) and its requirements for adequate yearly progress put unprecedented emphasis on state tests, educators have discovered that annual state tests are too little and too late to guide teaching and learning. Recognizing the need for more frequent assessments to support student learning, many districts and schools have turned to benchmark testing—periodic assessments through which districts can monitor students' progress, and schools and teachers can refine curriculum and teaching—to help students succeed. We report in this document a collaborative effort of teachers, district administrators, professional developers, and assessment researchers to develop benchmark assessments for elementary school science. In the sections which follow we provide the rationale for our work and its research question, describe our collaborative assessment development process and its results, and present conclusions.

Background

Despite the attention that NCLB places on state and other accountability testing, research conducted during the last two decades strongly suggests that the mainstay of educational assessment in U.S. primary and secondary school classrooms tends to be teacher-developed or teacher-selected measures rather than externally imposed assessments (Stiggins, 1991). In fact, it seems that most "teachers prefer using their own tests even more than externally developed tests for making educational decisions" (McMorris & Boothroyd, 1993, p. 322). While more recent research has reached similar conclusions, it has also suggested that teacher assessment preferences may be more heterogeneous across grade levels and subject areas than originally thought. For example, recently Rodriguez (2004) found that

[1] The authors would like to thank the CAESL professional development team: Lynn Barakos, Diane Carnahan, Karen Cerwin, Kathy DiRanna, Craig Strang, Jo Topps, and all the district specialists and teachers whose efforts brought this project to fulfillment. We also appreciate the significant contribution of our editors, Ms. Joanne Michiuye and Mr. Fred Moss.

elementary school teachers are much more likely than secondary school teachers to use tests created by textbook publishers, and that such tests are disproportionately used in mathematics. Depending on their use, these tests can have important consequences for the student. Test results provide important data for guiding instruction, inferring understanding, and refining instructional materials. In addition, results can also be used to determine future educational opportunities, and can also have other important and often overlooked consequences. For example, the research of Rodriguez concludes that tests, even those developed by publishers, communicate to students the learning goals and thinking processes valued by teachers, can significantly impact student self-efficacy, and concomitantly, can themselves affect student performance.

Given that the consequences of such assessments are large, one might expect that the quality of accompanying tests would be an essential component in the evaluation of all educational materials. Moreover, given the disproportionately large number of locally developed or selected assessments, it seems logical that essential safeguards should exist to guarantee the validity of the inferences made from such assessments and that teachers should receive a modicum of training to assure the data provided by such assessments will serve their intended purpose.

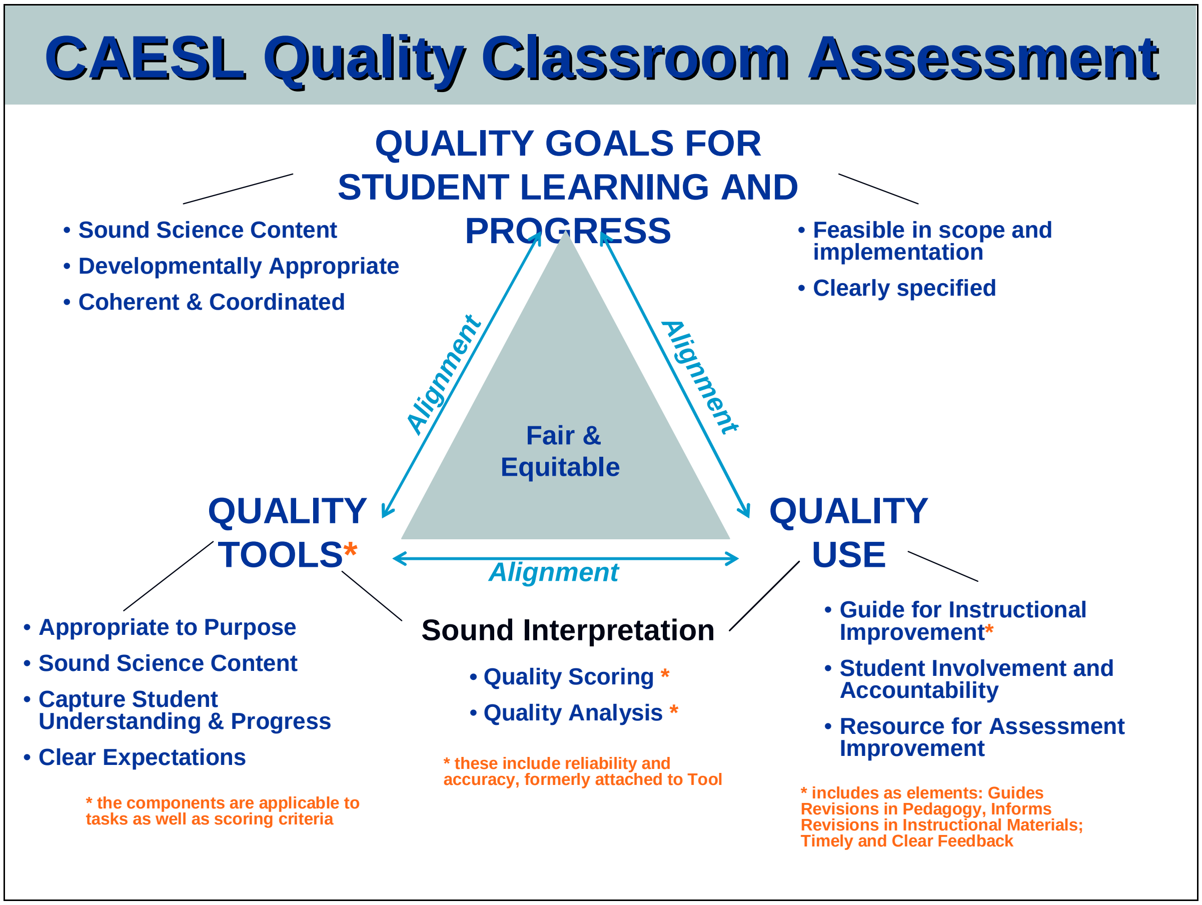
Unfortunately, many teachers have received little training in assessment selection or development (Black, Harrison, Lee, Marshall, & Wiliam, 2003; Black & Wiliam, 1998) and even published tests seldom demonstrate commonly accepted standards of technical quality. Teacher-developed tests often require only that students recall material, are too short to adequately sample the intended knowledge domain, and often evaluate students on criteria irrelevant to the objectives being assessed (McMorris & Boothroyd, 1993). In addition, this same body of research recounts occasions where teachers intentionally gave students hints about a correct solution or strongly favored a particular correct response discernable to students (Kane, Khattri, Reeve, & Adamson, 1997). Assessments supplied with published texts have similar problems with technical quality. Tests published with textbooks were often found to be too narrow (content under-representation), misaligned with objectives (criterion irrelevant), or too simple in the cognitive demands placed on students (Frisbie, Miranda, & Baker, 1993).

While valid inference is often thought to be most critical for interpreting the high-stakes testing data used to make certification, placement, and promotion decisions, accurate inference about student understanding is necessary in formative decisions too (Black & Wiliam, 2004). It is important that teachers have reliable data in order to determine what to teach next, when to move on to the next topic, or when and how to best re-teach a concept. In addition, because high-stakes tests (like state assessments) have achieved a much greater significance for teachers, schools, and states after the passage of the NCLB Act, teachers have a need to know what students understand or don’t understand before the students take such tests. Preferably, the results of local assessments should allow teachers to predict how students will perform when they encounter similar material on state assessments. Ideally, teachers would like assessments that allow them to help each of their students build conceptual understanding in the most efficient way possible.

Locally developed assessments that are aligned with major instructional goals, that accurately predict student performance on high-stakes assessments, that allow precise estimates of current understanding, and that correctly identify current misconceptions should be useful in making instruction both more effective and more efficient. In addition, such assessments could make testing itself more efficient since the results from individual assessments could be used for formative purposes or aggregated to serve a more summative function (Baker, 2004). Locally developed tests of high quality might also have a beneficial effect on student self-concept and motivation to learn.

These concerns, among others, prompted the U.S. Department of Education to recommend that schools and districts implement small pilot studies to ascertain the reliability and validity of the assessments they develop (Kane et al., 1997), a recommendation that mirrors requests we have received from local districts and teachers for help with test development and validation. Implementing such studies, however, is often easier said than done. For educational assessments to meet commonly accepted standards of quality requires that tests adequately sample a domain of interest, measure against explicit criteria, and make verifiable judgments about specific traits of interest. In addition, as suggested earlier, some authors (most notably Messick, 1989) have argued that the consequences of assessment be considered in a review of technical quality.

We and colleagues at the Center for the Assessment and Evaluation of Student Learning (CAESL) have worked to bring this vision of assessment to reality through an extended professional development program for teachers and an applied program of R&D. The CAESL framework guiding the work is shown in Figure 1. The figure broadly communicates a reflective teaching process that starts with significant goals for student learning, continually assesses student understanding relative to those goals, and uses the results to guide and support student progress (DiRanna, in press). At the same time, Figure 1 shows that assessment in support of student learning requires both quality assessments and effective use of results, and that both must be carefully crafted and aligned with goals for student learning (Herman & Baker, 2005; Herman, Osmundson, Ayala, Schneider, & Timms, 2005).



CAESL Quality Classroom Assessment

QUALITY GOALS FOR STUDENT LEARNING AND PROGRESS

- Sound Science Content
- Developmentally Appropriate
- Coherent & Coordinated

- Feasible in scope and implementation
- Clearly specified

QUALITY TOOLS*

- Appropriate to Purpose
- Sound Science Content
- Capture Student Understanding & Progress
- Clear Expectations

* the components are applicable to tasks as well as scoring criteria

Sound Interpretation

- Quality Scoring *
- Quality Analysis *

* these include reliability and accuracy, formerly attached to Tool

QUALITY USE

- Guide for Instructional Improvement*
- Student Involvement and Accountability
- Resource for Assessment Improvement

* includes as elements: Guides Revisions in Pedagogy, Informs Revisions in Instructional Materials; Timely and Clear Feedback

Figure 1. The CAESL Quality Assessment Triangle.

The project described here concentrates on the quality assessment component of the framework. We describe a collaborative test development process we used to create benchmark assessments for elementary school science in coordination with a continuing professional development program to deepen teachers' assessment capacity. Each of the four benchmark assessments that were created (one by researchers and three by participants) addressed a key topic in the fourth or fifth grade curriculum and was intended to provide information that districts could use to monitor student performance toward state standards, and that teachers could use to improve teaching and learning. In line with CAESL and others' criteria for quality assessment (Herman & Baker, 2005), we sought to deepen teachers' capacity and develop assessments that were:

- Aligned with major disciplinary ideas, as well as with state standards and instructional materials.
- Appropriate in levels of cognitive demand to reveal student understanding.

- Consistent with current scientific knowledge.
- Adequate in reliability.
- Fair in expectations for students' opportunity to learn and free from cultural bias or insensitivities.
- Feasible for classroom administration within typical time constraints.
- Useful to teachers.

The ongoing work explores the effectiveness of our efforts, including the following questions:

1. Does a collaborative test development process involving teachers, district staff, professional developers, and assessment researchers focused on the major ideas within a domain produce selected-response, open-ended, and performance-based assessments of demonstrable high technical quality?
2. Do student results on these assessments predict student results on applicable portions of the Grade 5 California Standards Test (CST) in Science?
3. Do the data and analytical results derived from these assessments help teachers make instructional decisions, such as what concepts to teach/re-teach and to determine what alternative conceptions students possess?

The current report will present results of our analysis of pilot tests conducted in order to answer the first of these three questions. The other two questions will be answered after data from the 2005-2006 Science CST and subsequent survey data become available.

Methods

Test Development Strategy

The project involved 18 teachers and district administrators from five California school districts with six professional development leaders who also were expert in science content and pedagogy, and a team of three assessment researchers. The group was divided into three test development teams, each charged with the development of multiple-choice, open-ended, and performance assessment items for one of three intended benchmark tests: fourth grade Ecosystems, fifth grade Properties of Matter, and fifth grade Water Cycle. The fourth benchmark (Electricity and Magnetism) was developed by researchers prior to the first meeting and used as an instructional tool for these test development teams. The test development process involved initial item development, expert content review, piloting and revision of items (and their scoring), selection and field testing of final test forms, and conduct of final technical analyses. The general design for each benchmark assessment anticipated a two-day test window, where the Day 1 assessment would be composed of 20 multiple-choice items and two open-ended explanation items, and Day 2 featured a hands-on performance assessment focusing on an experiment. The next sections elaborate on this test development process.

Test development participants. The teachers and administrators involved in the test development previously had been a part of a three-year professional development program led by the participating professional developers. Year 1 concentrated on large-scale assessment issues and involved teachers in the design, administration, and scoring of a multi-method assessment system designed for state use on the development and scoring of a science assessment, called PASS (Partnership for the Assessment of Standards-based Science). PASS involved innovative multiple-choice formats, open-ended items, and hands-on performance assessment. In Years 2 and 3, teachers applied the CAESL assessment framework described earlier to their classroom assessment practices. Working in grade-level teams and with each team focusing on a single science instructional materials unit (e.g., the water cycle, ecosystems), teachers created an assessment plan by mapping the conceptual flow of their target unit, identifying key juncture points for assessment within the unit, and selecting/adapting from the curriculum materials assessments for these points. The assessment plans included pre- and post-tests and several checkpoints in between. Teachers then implemented the curriculum and assessments with their students. They were guided through a process of analyzing the results and determining implications for: (a) ongoing teaching and learning, and (b) refinement of the assessments and instructional plans for the next time the unit was taught. The participants also learned to evaluate assessments based on their goals for instruction and on how they planned to use assessment results (e.g., a formative or summative purpose). In addition, they learned how to perform a rudimentary evaluation to determine the technical quality of assessment items, scoring rubrics, and the inferences made from these items. Topics such as validity, reliability, and generalizability were discussed, although participants were never required to determine quantitative parameters of technical quality (e.g., item difficulty, point biserial coefficients, or reliability coefficients). See DiRanna (in press), and Gearheart et al. (2006) for additional details about the professional development program. Representing each of five participating districts, a core group of teachers and district staff were invited to participate in a follow-on benchmark assessment development institute by virtue of their subject matter expertise, as demonstrated by the fact that they had taught an assessment topic for at least five years.

Initial assessment development process. Initial item development was accomplished during a two-and-a-half-day summer institute. Participants gathered on the first morning to hear presentations on critical quality issues in the development of benchmark assessments. They also heard presentations on the value of aligning assessment items with the big ideas (as we called the major principles and concepts) within a domain, and of purposively developing items to address specified content and cognitive demand to assure adequate coverage of the domain. The approach was consistent with recent research suggesting that isolating item content in small domains has positive effects on the validity of the inferences drawn from a collection of items. In particular, content-related evidence generated during the development of an item can provide support for the reasonableness of the item's target domain, the appropriateness of the scoring rules and procedures, the adequacy of the sampling of the target domain, and the generalizability of the resulting scores (Kane, 2006). Moreover because cognitive research shows the learning value of engaging students in applying and explaining key disciplinary principles in a variety of contexts, assessments so structured can contribute directly to student learning (Herman & Baker, 2005).

The remainder of the day involved participants in hands-on exercises to expand on these concepts, and to deepen educators' understanding of the alignment of assessment and learning goals and elements of item quality. Prior to the day, researchers had developed a conceptual flow, and assessment items in the area of electricity and magnetism. Working in small groups, teachers reviewed and revised the map of the big ideas of electricity and magnetism, reviewed and revised the content and cognitive demands (e.g., factual recall, conceptual understanding, problem solving) inherent in a pool of test items of that content, and also reviewed and revised other items for technical flaws (violations of given item-writing rules and principles) and for bias, including unwarranted linguistic complexity. Each of these activities was an occasion for continuing discussion of the criteria for good test items and for creating tests aligned with the understanding of key ideas in a domain.

At the beginning of the second day, the group was divided into three development teams. Each team was composed of one or two professional development specialists, four to six educators (including district staff), and an assessment research specialist. The professional development specialist functioned as both facilitator and team member. Teams first developed the conceptual flow of big ideas and subtopics in their particular topic areas, based on state standards, reviews of common instructional materials, and their own content and pedagogical knowledge. Teams designated concepts essential for testing and the cognitive level at which students should understand each concept — factual recall, understanding, or problem solving. Generally, the understanding level was given priority. The Appendix shows the conceptual flow for the water cycle as an example of the conceptual flows developed by each group.

Each team then broke into two subgroups of three, and each threesome then developed a set of assessment items for each of the ideas represented in the conceptual flow of their domain. The intent was to develop twice the number of items that would be needed for the eventual benchmark test to enable poor items to be weeded out during the pilot and field test stages of the project.

As noted earlier, the plan was for a test composed of a majority of multiple-choice format items, as they are relatively fast for students to take and relatively easy for teachers to score; however, while multiple-choice items may be efficient, they "may not provide the appropriate information for identifying students' misconceptions with respect to the given subject matter" (Birenbaum & Tatsuoka, 1987, p. 392). In addition, multiple-choice items tend to have less diagnostic value than open-ended responses, unless there are multiple, parallel items (Birenbaum, Tatsuoka, & Gutvirtz, 1992). Consequently, each subgroup was tasked to develop a set of open-ended and performance task items in addition to their multiple-choice items. In general, each group of three developed 20 selected-response items, 2 open-ended items, and 1 extended performance task. As a result, this procedure generated a total of approximately 40 selected-response items, 4 open-ended items, and 2 performance tasks that were aligned to the conceptual flow for each of the content areas. In most cases, however, fewer than 40 selected-response items were piloted because some items were weeded out prior to pilot testing.

Participants were encouraged to develop their multiple-choice and open-ended items using the Assessment Design and Delivery System (ADDS). One of the tools the ADDS provides is the assessment designer—a structured assessment development template that focuses the assessment developer's attention on key aspects of assessment design. In particular, the ADDS suggests that designers specify key attributes of the item such as the grade level(s) and linguistic ability the item is appropriate for, the subject area or domain of the test item, the appropriate standard being assessed, the big ideas or topics from the domain being tested, and the level of cognitive demand the item requires of the test taker. These levels—transfer, explain, complex problem solving, make connections, application, and recall—follow the cognitive demand levels developed at CRESST and are research based (see for example, Baker, 1998). While item format is often associated with cognitive level (e.g., multiple choice is equated to recall), we and others encourage developers not to make that association (Stiggins, 1994). Rather, the ADDS asks item developers to first consider the instructional goal, the use, and the cognitive demand of the required item, then to choose a format that best meets these needs.

The ADDS also encourages developers to consider some basic guidelines in the development of items using each type of format. In particular, developers were encouraged to consider the following in developing multiple-choice tasks:

- Items should address important content accurately. They should not focus on the memorization of isolated or inconsequential facts.
- Both item prompts, and item answers should be clear and concise. Avoid double negatives (or even negatives) and overly complex wording, especially when such wording is not required to assess the content of interest.
- Use “all of the above” or “none of the above” sparingly. The answer “all of the above” can be eliminated by knowing any one selection is incorrect or verified if any two answers are correct. As such, they can adversely affect the technical quality of the item because students who may know very little about the content of interest may be able to answer the question correctly. In addition, questions, especially complex questions, with “none of the above” as an answer, may disproportionately attract student responses because of minor misconceptions or procedural errors on the part of the student (Docy, Moerkerke, DeCorte, & Segers, 2001).
- Choose distractors in order to pinpoint common misconceptions and provide educators information, which will be useful in follow-on instruction.
- Write all choices using parallel construction.
- Address a single concept or big idea with a single question.

These guidelines roughly parallel those of Haladyna and Downing (1989); however, as these authors point out, there is contradictory evidence that suggests these guidelines alone do not produce quality selected-response items. In particular, Traub (1992) argued that these guidelines gave technical quality measures too much weight and that developers need to provide the arguments that support the interpretation of a particular test score. For this reason, item developers in this study were encouraged to constantly match multiple-choice items, including distracters, with the conceptual flow, and to develop items using the ADDS template.

A slightly different method was used to generate open-ended prompts and rubrics. The designers of open-ended items were encouraged to create prompts that contained information a test taker could use to explain, analyze, recall or apply their conceptual understanding. Here again, cognitive levels were not explicitly linked to an item’s format.

As with the CAESL model (Herman, Osmundson, et al., 2005), the ADDS encouraged designers to integrate open-ended prompt and rubric design into a single activity. In other words, designers were encouraged to design rubrics in conjunction with prompts and to engage in an iterative refinement of each as the design process illustrated shortcomings of one or the other. For example, a prompt on an ecosystems item initially asked a test taker to “Look at the ‘beaks’ and the types of ‘food’ that are available. Predict which type of food each beak can pick up the best. Explain your prediction.” Rubric developers indicated that the response required the student to explain which type of beak was most appropriate for which type of food. This, however, was identified as unclear since the prompt could have been interpreted as asking which type of beak is best for all three types of food. What the test developers wanted was a hypothesis based on scientific concepts. Consequently, the prompt was modified to "Look at the beaks and the types of food that are available. Select one type of beak and predict which type of food that beak can pick up best. Explain your prediction." The rubric, in turn, was evaluated against the goals of instruction. Ultimately, the prompt might ask respondents to not only identify the best beak for a particular type of food, but what similarities the beaks in the model have to real beaks in the wild. It might also identify the need to separate the prompt into multiple parts as the complexity of the scoring (and the prompt itself) becomes more difficult to understand.

After developing initial items, each group of three participants exchanged their items for review with their mirror group within each content domain. For example, the three participants who created water cycle items shared these items with the other three participants creating water cycle items. If required, the items were again revised in order to clarify their meaning or to focus their assessment objective.

Item review and preparation for the pilot testing. After the items were initially developed, they and the conceptual flows on which they were based were subjected to expert review. Colleagues from the Lawrence Hall of Science, who were collaborators in the professional development, enlisted scientists from the Hall with relevant subject matter expertise to do a content review, resulting in some revisions. Curriculum experts also did a review and suggested minor item revisions. The item sets were then returned to the assessment researchers for final formatting and editing. Two test forms were created for each content area and prepared for pilot testing. Because time limitations are known to have an adverse effect on test taker (and item) performance (Livingston, 2006), especially near the end of a test, tests were not timed and were designed to take less than 45 minutes. An average class period was known to be 55 minutes in the population tested.

Each of the five participating districts was tasked with selecting one to two classrooms for participation in the pilot testing of each assessment, with the general charge that participating students should be typical of the district, while also assuring that the pilot test included some high-performing students. That is, we wanted the pilot test to provide data on whether the items were appropriate, but also wanted to be sure that we had a full range of performance and particularly had examples of high-performing students in order to hone the scoring rubrics for the open-ended items. Together, the five districts represented both northern and southern California, and a mix of demographic and language characteristics.

Our professional development colleagues coordinated the production, mailing, and follow-up of the pilot test administration, while the tests themselves were returned to the assessment researchers. Approximately 180 students completed the assessments in each area. Teachers participating in the pilot testing were asked for feedback on the items and individual students within the selected classes also were asked individually about problems they may have encountered with individual items on the test.

Review of pilot test results. After pilot testing was completed, the item developers met to select and, as necessary, revise the items that would undergo field testing. In preparation for that meeting, the assessment researchers were charged with entering student responses to and conducting psychometric analyses of the multiple-choice items. With our professional development colleagues, they selected a range of responses to the open-ended items and refined initial rubrics for scoring them.

As suggested by Traub (1992), the choice of items should consider what the item measures, not just the technical quality of the item. Consequently, items were selected based on three criteria. First, items were eliminated if they did not perform as expected. While this often meant that an item did not fit appropriate statistical models or expectations (e.g., a 1-pl IRT model in the case of selected-response items, and frequency descriptives for open-ended responses), occasionally items did not elicit the expected student responses or the range of expected student responses. In other words, an item was eliminated if it did not seem to measure the appropriate construct or did not allow valid, criterion-based discrimination. Second, in an effort to ensure content validity, items were selected based on their coverage of key ideas in the conceptual flow. Finally, items were selected to ensure a diverse range of difficulty.

Because of the considerable time required to score all the answers to the open-ended items, the assembled assessment teams scored only a sample of student responses from the pilot testing to ensure that the items performed as expected and that scoring rubrics were suitable. In each case, raters were given 10 actual student responses in order to calibrate their scoring on the given rubric. Raters then scored a randomly selected set of responses in a round robin format. At that point, participants determined which two of the open-ended items in each content area performed best in eliciting student understanding of the intended construct and which best matched the conceptual flow priorities for that area. Available testing time, reliability, and content coverage suggested that two open-ended items were optimal. Consequently, only responses to two of the four open-ended items developed in each content area were scored completely, and those two items subsequently field tested.

Field testing. Similar to the pilot testing, the field test versions of the test items were subjected to additional content review and refinement prior to assembly and final editing by the research team. The professional developers again coordinated the production and distribution of the test form materials and assured their return. Districts again were responsible for identifying field test classrooms and were asked to select six to eight classrooms that represented the range of typical performance in their districts. Approximately 500 students participated in the field testing of each of the four benchmark assessments. As with the pilot testing, completed tests were returned to the assessment researchers for entry and psychometric analyses of the multiple-choice items, and assessment researchers collaborated with their professional development colleagues to prepare for a final assessment team meeting and the scoring of students' open-ended responses.

Technical Evaluation of Items

Multiple-choice items were evaluated using a standard one parameter logistic (1-pl) model and skew statistics. In each case, we also examined the relationship between a student's performance on a particular item and the student's performance on the test using the point biserial correlation. To do so, we developed two test forms for each of the four domains by randomly distributing multiple-choice items between two forms so that students were not required to attempt to complete an assessment that was too long. Both test forms in each domain, however, shared some common items that allowed us to compare items using a common logit scale.

As defined in Brown (1988), we have used the point-biserial correlation coefficient rather than a biserial correlation to estimate the degree of relationship between a student's response to an item (correct or incorrect) and percent score on the overall test. It should be noted, however, that the point-biserial can only indicate the correlation between an item and other items that actually appeared on the test, not on all items that could be written to cover the concepts tested.

We evaluated the quality of the selected open-ended items using three different measures. First, in addition to asking both teachers and students to indicate if the questions were unclear, we also used item descriptive statistics such as frequency of response and skew of student responses. Because items were generally written so that a score of 5 was above "mastery," we expected few, if any, students to achieve that score. We did, however, want to allow for that possibility, and to provide students and teachers alike with high expectations. Likewise, a score of 0 allowed raters to indicate no response or a response that was unintelligible. While scores were considered ordinal in nature (e.g., a score of 3 was of higher quality than a score of 1 or 2), developers did not intend that student responses be equally distributed over all score points, just that there be some distribution. In fact, it was generally agreed that very few students would score “4” and that most questions should not have a high percentage of “0” scores (possibly indicating that a question was unclear to students). Consequently, participants generally wrote items expecting that the median response would be between 2 and 3.

Secondly, and in addition to the open-ended items themselves, we also considered scoring rubrics in our determination of item “goodness.” According to the CAESL model, scoring criteria (rubrics) should be considered along with assessment prompts when determining whether an item is consistent with instructional goals, whether an item is appropriate for use in a given situation, and whether an item performs as expected (i.e., is accurate and reliable). Although our pilot study design prevents us from dissociating rater and rubric variance, and interactions, we can estimate the proportion of score variance attributable to differences in student responses as opposed to between and within rater scores. In addition, we checked to ensure that the items are measuring interconnected content. Each of these measures is described below.

There are a number of correlation coefficients that can serve as an indicator of rater reliability. We have avoided using the Pearson correlation coefficient since, in this case, the direction of correlation is not important (Rater 1 and Rater 2 are arbitrary designations). Consequently we use the Intraclass Correlation Coefficient (ICC). As suggested by Shrout and Fleiss (1979), we have chosen to use a 1 way random ICC. This estimate of reliability takes the form $\dfrac{BetweenMS - WithinMS}{BetweenMS + (k-1)WithinMS}$. Given that the number of judges rating each open-ended student response in this study is 2, the (k-1) term in the denominator drops out, and the interpretation of the ICC becomes the proportion of the total variance that is associated with differences between students. For our purposes then, ICC values closer to one mean that a greater proportion of the variance in scores is associated with actual differences in student responses rather than with the way different raters interpreted those responses. Because of the design of this study, rater–student interactions cannot be isolated from other sources of error or rater variability. Again as suggested by Shrout and Fleiss, since we are interested in reliability across raters, we compute ICC using differences between individual scores (an absolute agreement model) rather than differences between individual and average scores (a consistency model). Although we only report absolute ICCs in this report, we did compute both values for each of the open-ended items. In every case, a consistency model produced ICC values closer to unity for our data.

In addition to ICC we have also computed Kappa scores as a statistical measure of exact agreement between raters once chance agreement is removed. Although the Kappa statistic can be problematic in cases that mirror the present study (Fleiss, Nee, & Landis, 1979), we have used it not as the sole measure of inter-rater reliability, but as a measure of rater agreement above and beyond mere chance. Specifically, the Kappa statistic is defined as: (percent observed agreement – percent expected agreement) / (1 – percent expected agreement). Expected agreement and observed agreement are defined in a way similar to the familiar Chi-square measure in that expected agreement (from random chance) is the sum of the products of the percentages of observations in row_i and column_i of a square, two dimensional matrix, where i varies from 1 to n (the number of rows / columns). The percentage of agreement is calculated by summing the number of times the raters agree (the diagonal of the matrix) and dividing by the total number of observations. Table 1 shows how we computed the Kappa statistic for the first open-ended Electricity and Magnetism item.

Table 1

Computing Kappa Example (Elect 1)

First rater data	Second rater data						
	Score						Totals
Score	0	1	2	3	4	5	
0	8[a]	2	0	0	0	0	10
1	4	65[a]	27	1	0	0	97
2	0	45	181[a]	15	6	0	247
3	0	0	20	32[a]	15	0	67
4	0	1	3	5	9[a]	13	31
5	0	0	1	2	4	45[a]	52
Totals	12	113	232	55	34	58	504

[a] Kappa statistic for this data is computed by first summing these cells.

The Kappa statistic for this data is computed by first summing the indicated cells (340) and dividing by the total number of observations (504) to determine the percentage of observed rater agreement (.675). Next we find the percent expected agreement by summing the products of the percentages in the i^{th} row and the i^{th} column: ((12/504)(10/504) + (113/504)(97/504) + … (58/504)(52/504)) = .2997. Finally, we compute Kappa: (.675 – .2997) / (1 – .2997) = .536.

Given the way Kappa is defined, if two raters (or the same rater when finding intrarater reliability) never differed when scoring identical responses, the number of exact agreements would equal the total number of scores and Kappa would equal 1. Conversely, if the raters never agreed, the number of exact agreements would equal zero and Kappa would be negative. Kappa would be zero if raters agreed no more often than expected by chance. By convention (see Landis & Koch, 1977; Viera & Carrett, 2005), Kappa scores of .2 to .4 show fair agreement, .41 to .59 are considered to indicate moderate inter-rater agreement (reliability), Kappas of .6 to .79 represent substantial agreement, and scores of .8 or larger indicate outstanding agreement (reliability). One must use caution, however, because while Kappa is indicative of exact agreement among raters over and above chance agreement, it does not indicate how rater scores are distributed. In other words, if all raters consistently scored student responses as a one on a zero-to-five-point scale, Kappa would equal 1, but the item might still be of little inferential value. This again argues for the need for the more descriptive measures of assessment item quality we include below.

As a final measure of rubric (and rater) reliability, we calculated intrarater reliability. It is commonly accepted that a rater's scores are likely to drift over time, especially if a scoring rubric imposes a scoring system on raters that differs from the rater's own beliefs. The effect is likely to be even more profound if the rubric appears arbitrary or unclear to the rater. The greater the difference between a rater's "internal rubric" and the imposed rubric, the more likely within-rater (intrarater) variance is apt to be large. In addition, this variability is prone to increase as the length of scoring sessions increases. Intrarater reliability is calculated using the ICC (exact agreement of a rater with him or herself over time) to monitor such drift and to serve as a proxy of the clarity of each scoring rubric. While an exact agreement of 100% is desirable, we realize that exact agreement is nearly impossible to achieve and that, as before, we cannot isolate the various sources of drift given our study design. Consequently, intrarater reliability statistics are reported only as a rough measure of the soundness of each scoring rubric.

Finally, as suggested by Welch (2006) we examine the generalizability of the questions to the general knowledge domain by calculating the inter-prompt correlations to determine the extent of content coverage. We do so by comparing median ratings for each of the students who had scores for both open-ended questions in each knowledge domain. As the data was ordinal, we used Spearman's rho (r) to determine the correlation for items across student responses. This coefficient will range in value from –1 (a perfect negative linear relationship) to +1 (a perfect positive linear relationship). As they address a common domain of knowledge, we expect to see a significant correlation between the two open-ended items in each domain; however, a very large intercorrelation is not desirable. A very high correlation would suggest that each question is measuring not just the same domain but, in fact, identical content or a narrow range of student ability. In addition, we have conducted a generalizability study (G-study) for each set of open-ended items to estimate the proportion of variance attributable to differences in students, differences in items and, for three of the four domains, differences attributable to variability in the raters. Because not every rater scored each student response on these items, raters are nested within students. Our G-study for the Electricity domain was slightly different. Since the Electricity items were used for training purposes, there was no overlap in raters between the two open-ended items for Electricity. One group of six raters rated Electricity Item 21 and another group of six raters rated Electricity Item 22. Then, in each group, raters scored a subset of student responses. Consequently, raters are nested within item for the analysis on this section. In all cases, students, raters, and items are considered random facets for our purposes since each is a random sample of the larger student, rater, and item population, respectively.

Results

The process described above allowed us to pilot test and score up to 40 multiple-choice items, 2 open-ended items, and one performance task in each of four knowledge domains, the three developed by the assessment development teams (ecological systems, water cycle, and properties of matter) and the fourth, electricity and magnetism, developed for the initial training process. Logistical problems with the performance tasks delayed pilot testing and contaminated the data collected from these tasks, so performance task results will not be reported here. Consequently, we evaluated and report here the results of pilot testing the multiple-choice and open-ended items in each domain.

Selected-Response (Multiple-Choice) Items

The selected-response items in each of the four domains were multiple-choice items consisting of a prompt (either a complete question or stem) and, in general, four possible responses. All but five (over 96%) of the 140 multiple-choice items developed by participants fit the standard Rausch (1-pl IRT) model. Using this model, we calculated difficulty parameters for each item. The difficulty of the items was generally normally distributed from easy to difficult (-2.00 to +3.00 logits).

Table 2 presents the difficulty (in logits) and the point bi-serial correlation coefficient for each of the items pilot tested in the domain of fifth grade Electricity and Magnetism. These results are based on 250 valid student responses.

Table 2

Results for 1-pl IRT Analysis of Electricity and Magnetism Items

Item	Difficulty	Pt. Bi-serial
1[a]	+ 0.88	+0.23
2[a]	-1.06	+0.34
3[a]	+1.62	+0.21
4	-0.25	+0.13
5[a]	+0.32	+0.35
6[a]	+0.47	+0.23
7	+1.02	-0.09
8	No Ans.	NA
9	-1.00	+0.27
10[a]	+0.35	+0.43
11[a]	+0.67	+0.43
12	-0.46	+0.37
13[a]	-1.17	+0.21
14[a]	+0.45	+0.37
15	+0.13	+0.48
16	+0.30	+0.12
17[a]	-0.38	+0.67
18[a]	+0.04	+0.46
19[a]	-0.34	+0.39
20	-1.21	+0.38
21[a]	-0.06	+0.66
22[a]	-0.23	+0.42
23[a]	-0.50	+0.54
24	-1.96	+0.26
25[a]	-0.14	+0.60
26[a]	+0.17	+0.37
27[a]	+1.13	+0.13
28	+1.45	+0.10
29[a]	-0.23	+0.30

[a] Items chosen for inclusion on the final field test.

The items as a group had a slight positive skew (+0.12) indicating that the item difficulties are approximately normally distributed, but that individuals are correctly answering slightly more questions than, on average, would normally be expected. In addition, 2 of the 29 items (Items 7 and 8) do not exhibit the technical quality generally considered acceptable. Closer inspection revealed problems: In the case of Item 7, students who did less well on the test as a whole were more likely to correctly answer this question than students who did better on the test as a whole. A reexamination of this item revealed two potentially correct answer choices with the less likely of the two selected as the correct choice on the answer key. Item 8 did not have a single acceptable answer. All other items were considered technically sound.

Of these items, the 20 items that best addressed the content specified in the conceptual flow and that allowed for a range of student abilities (difficulty levels) were chosen for inclusion on the final field test. In the table, these item numbers are followed by an “a” footnote. Although not specifically considered when the items were chosen, the technical quality of the selected items is comparable to the technical quality of the items on the fifth grade California Standards Test (CST) for Science. Specifications for the 2005 Administration of that CST stipulated a target difficulty rating of -.19, a mean point biserial greater than 0.34 and minimum point biserial of 0.14 (Educational Testing Service, 2006). Of the 20 Electricity and Magnetism items selected for field testing, average difficulty is +.1, mean point biserial is 0.39, and the minimum point biserial is 0.13. Pilot items not used on the field test were archived for later use by teachers and districts.

Table 3, based on the valid responses of 287 students, shows the difficulty (in logits) and the point bi-serial correlation coefficient for each of the items pilot tested in the domain of the Water Cycle for fourth grade.

Table 3

Results for 1-pl IRT Analysis of Water Cycle Items

Item	Difficulty	Pt. Bi-serial
1[a]	+0.12	+0.43
2	-1.78	+0.30
3[a]	-0.83	+0.32
4	+0.70	+0.17
5[a]	+0.25	+0.41
6[a]	-0.35	+0.30
7	+1.72	+0.07
8	-0.61	+0.37
9	+1.10	+0.17
10[a]	-0.72	+0.37
11[a]	-0.57	+0.23
12[a]	-0.12	+0.38
13[a]	+1.01	+0.26
14	-1.47	+0.35
15	+1.60	+0.08
16	+1.29	+0.12
17[a]	+0.56	+0.27
18	-1.44	+0.50
19	-2.24	+0.41
20	+0.11	+0.31
21[a]	+1.00	+0.56
22[a]	+0.70	+0.36
23	+0.92	+0.17
24[a]	-0.46	+0.31
25	-0.07	+0.27
26	-0.09	+0.48
27	-0.70	+0.33
28[a]	+0.22	+0.24
29	-0.09	+0.39
30[a]	+0.53	+0.40
31[a]	-0.69	+0.43
32	+1.93	-0.09
33[a]	+0.45	+0.44
34	+2.46	+0.03
35	-0.70	+0.32
36[a]	-0.81	+0.28
37[a]	+0.86	+0.30
38[a]	-0.18	+0.25
39[a]	-1.58	+0.57

[a]Items chosen for inclusion on the final field test.

The items as a group had a slight negative skew (-0.24) indicating that the item difficulties are approximately normally distributed, but individuals are correctly answering slightly fewer questions, on average, than would normally be expected. Thirty-eight of the 39 items display acceptable levels of technical quality. Of these items, 19 were chosen both to cover the content reflected in the conceptual flow, and to include items of various difficulty levels and various degrees of correlation with the overall test (point bi-serial statistics). These items are indicated by an "a" footnote following their item number in Table 3. The multiple-choice items selected for the final version of the Water Cycle test had a mean point biserial of 0.36 and a minimum point biserial of 0.23. Mean difficulty level of the selected items was -.03. The other 19 items were archived for later use by teachers and district personnel.

Table 4 shows the difficulty (in logits) and the point bi-serial correlation coefficient for each of the items pilot tested in the domain of Ecosystems for fourth grade. One hundred eighty-eight students provided valid responses for our analysis.

Table 4

Results for 1-pl IRT Analysis of Ecosystems Items

Item	Difficulty	Pt. Bi-serial
1[a]	+0.53	+0.21
2[a]	+0.90	+0.37
3[a]	-4.34	+0.07
4	-0.21	+0.38
5	-0.51	+0.43
6[a]	-1.21	+0.23
7[a]	-0.20	+0.33
8	-1.83	+0.38
9[a]	+0.58	+0.47
10	+1.56	+0.03
11	-0.01	+0.42
12[a]	+2.00	+0.11
13	+0.74	+0.30
14	0.00	+0.31
15	+0.79	+0.49
16	+1.00	+0.28
17	+0.73	+0.25
18[a]	-1.35	+0.38
19	+0.02	+0.39
20[a]	+0.79	+0.13
21	+0.19	+0.38
22[a]	+0.17	+0.44
23[a]	-0.53	+0.54
24	-0.45	+0.57
25[a]	-1.42	+0.24
26[a]	+0.53	+0.24
27	-1.42	+0.23
28[a]	+0.19	+0.37
29[a]	0.00	+0.23
30	-0.20	+0.33
31[a]	+0.49	+0.42
32	+0.44	+0.19
33	+0.17	+0.27
34[a]	-0.19	+0.32
35[a]	-0.62	+0.47
36	+0.16	+0.09
37[a]	+0.09	+0.32
38	+0.39	+0.32
39	+1.04	+0.10
40	+0.96	+0.30

[a] Items chosen for inclusion on the final field test.

Every multiple-choice item developed to assess student competency on Ecological Systems met minimum standards for technical quality. The 19 multiple-choice items selected for the final version of the Ecosystem test (as indicated by an "a" footnote in Table 4) have a mean difficulty of -.19, a mean point biserial of 0.31, and a minimum point biserial of 0.07. While this minimum point biserial coefficient is well below the CST target minimum (0.14), it should be noted that the vast majority (70%) of items on the four assessments had point biserials of 0.30 or greater. For various reasons, we are unable to calculate a skew for these items.

The 1-pl technical quality parameters for the fifth grade Properties of Matter multiple-choice items are given in Table 5. The parameters in Table 5 are based on responses from 287 students. The Properties of Matter items, as a group, have virtually no skew (-0.02) indicating that the item difficulties are normally distributed, and individuals are correctly answering the number of questions, on average, that would normally be expected.

Table 5

Results for 1-pl IRT Analysis of Properties of Matter Items

Item	Difficulty	Pt. Bi-serial
1[a]	+1.18	+0.32
2	+3.32	-0.10
3[a]	+0.54	+0.37
4[a]	+0.57	+0.35
5	-0.10	+0.31
6[a]	+0.07	+0.33
7	-0.76	+0.38
8	+0.23	+0.35
9[a]	+0.50	+0.27
10[a]	-1.65	+0.38
11[a]	-0.10	+0.26
12[a]	-0.87	+0.51
13	+0.79	+0.44
14[a]	-1.42	+0.48
15[a]	-0.80	+0.37
16	-0.38	+0.25
17[a]	+0.36	+0.32
18[a]	-0.28	+0.41
19[a]	-0.86	+0.27
20[a]	-0.71	+0.42
21	+0.64	+0.17
22	No Ans	NA
23[a]	-0.46	+0.47
24[a]	-0.23	+0.51
25	+0.96	+0.17
26	-0.62	+0.49
27	-0.46	+0.51
28[a]	+1.20	+0.17
29[a]	+0.71	+0.37
30	-0.80	+0.25
31	-0.52	+0.37
32[a]	-0.07	+0.54

[a] Items chosen for inclusion on the final field test.

All but two items (Items 2 and 22) demonstrated acceptable levels of quality. Item 2 seemed to assess content or against criteria that differed from that assessed by each of the other acceptable items and Item 22 had no acceptable answer and so was not evaluated.

Overall, the 19 Properties of Matter multiple-choice items selected for inclusion in the field test (denoted by an "a" footnote in Table 5) were similar in technical quality to multiple-choice items developed for the fifth grade California Standards Test for Science. The items selected for inclusion in the field test version of our Matter assessment had a mean point biserial value of 0.37 and a minimum value of 0.17. The mean difficulty of the selected items is -.12. These values are very close to CST specifications. The remaining 11 items were archived for future use by participating teachers and districts.

Open-Ended Items

Since open-ended items not only typically take longer to administer, but often also take longer to score, these items were selected and scored in a slightly different way from the process used for multiple-choice items. As explained above, only a minimal number of open-ended items were scored to assure the basic quality of each item. Two open-ended items in each domain were identified to completely score. All responses for the two open-ended items identified in each domain were then scored to determine technical quality using the measures described in the methods section.

The first open-ended question involving Ecological Systems asked students to explain the relationship between plants and animals in a given food web. Figure 2 shows the distribution of approximately 165 student responses to this question after scoring by four raters. In general student responses were as expected, although more than 35% of the students did not answer or supplied answers that were unintelligible. The fact that the inter-rater reliability was high (70% exact agreement among raters) and that the Kappa statistic (.549) was moderate suggests that variability in scores was due more to differences in student performance rather than rater differences. In fact, the single measure ICC statistic (.757) indicates that almost 76% of the variance in scores comes from student (rather than rater) differences. On the other hand, there was considerable drift in ratings over time. Raters matched their previous scores only about 59% of the time.

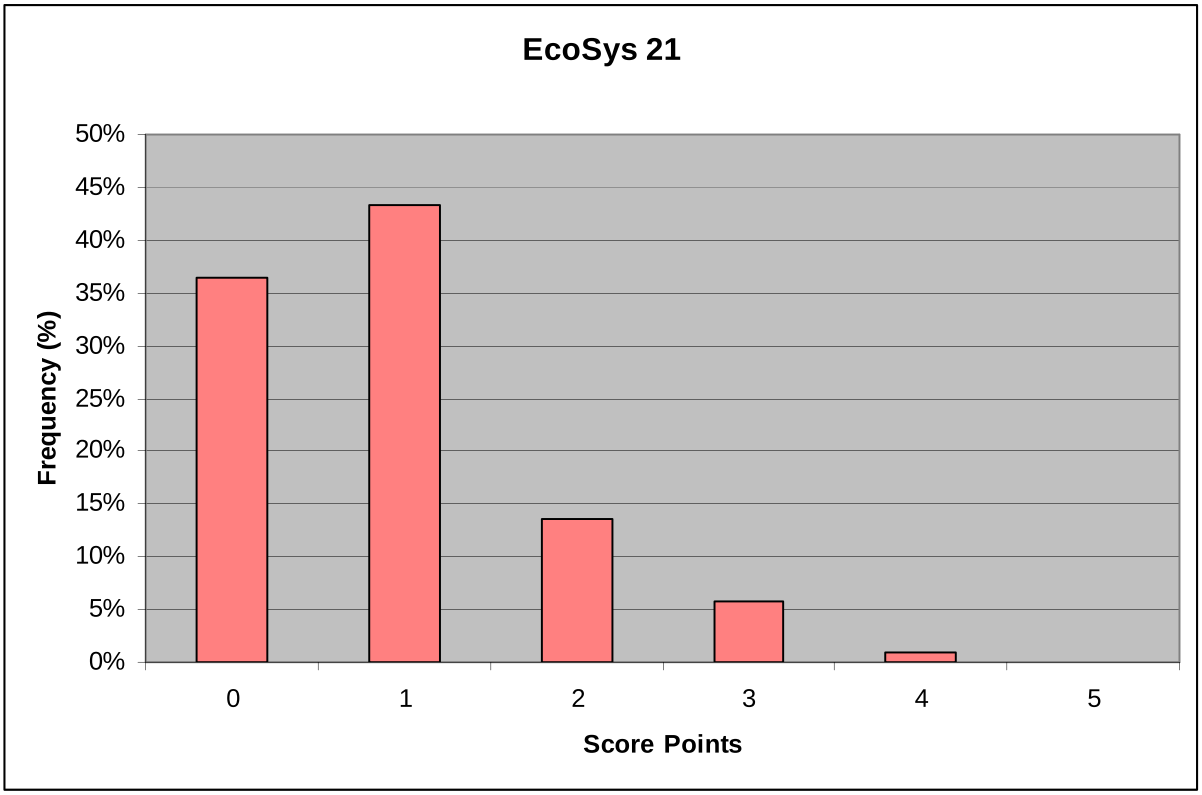


Figure 2. Distribution of student responses to Ecological Systems OE Question 1.

The second open-ended question involving Ecological Systems asked students to explain how a given set of living things depended on a given set of non-living things for survival. Figure 3 shows the distribution of about 168 student responses to this question after scoring by four raters. In general student responses were generally as expected, although a relatively large number of students (27%) did not answer or supplied answers that were unintelligible. The inter-rater reliability was high (63% exact agreement among raters) and the single measure ICC statistic (.83) indicates that most of the variance in scores comes from student (rather than rater) differences. Moreover, there was little drift in these ratings over time. Raters matched their previous scores about 83% of the time.

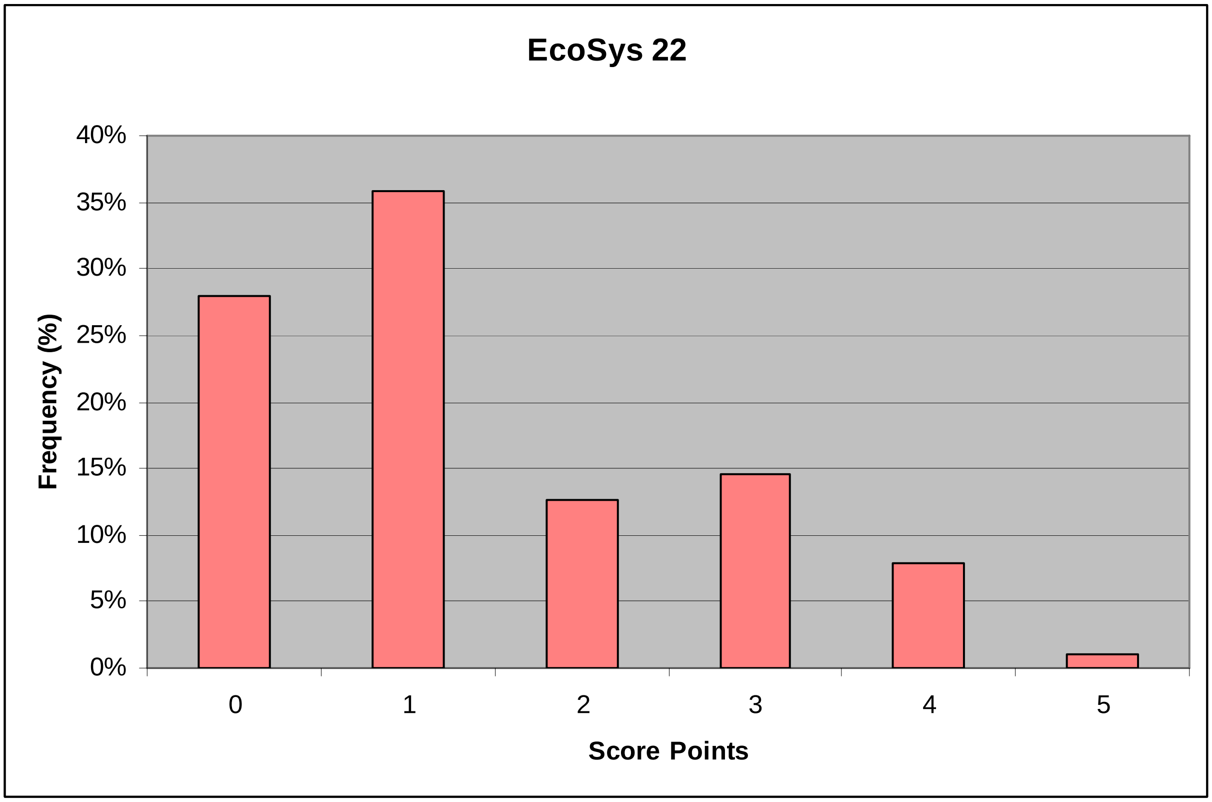


Figure 3. Distribution of student responses to Ecological Systems OE Question 2.

Student median ratings on the two open-ended Ecological Systems questions are significantly correlated (r = .567, p < .001) suggesting that these two items are assessing related, but not identical, content.

Table 6 reports the variance components for the open-ended items appearing on the Ecological Systems assessment. Approximately 36% of the total variance in our results is explained by differences in the students themselves. Differences between the two items and between raters account for small amounts of the variability (8% and 3%, respectively) in our results. The interaction of the student with the items accounts for over one third (37%) of the variability in our results. The remaining variance (15%) comes from measurement error and from interactions we cannot separate given our experimental method (e.g., rater[student] interaction with item). The variance components for these items suggest that they are, in fact, responding strongly to differences in students and the interaction between these student differences and the items.

Table 6

Variance Estimates for the Two Open-ended Ecological System Items

Component	Variance estimate	% of variance
Student	.49	36%
Item	.11	8%
Rater (Student)	.05	3%
Student × Item	.51	37%
Error	.21	15%
Total variance	1.36	100%

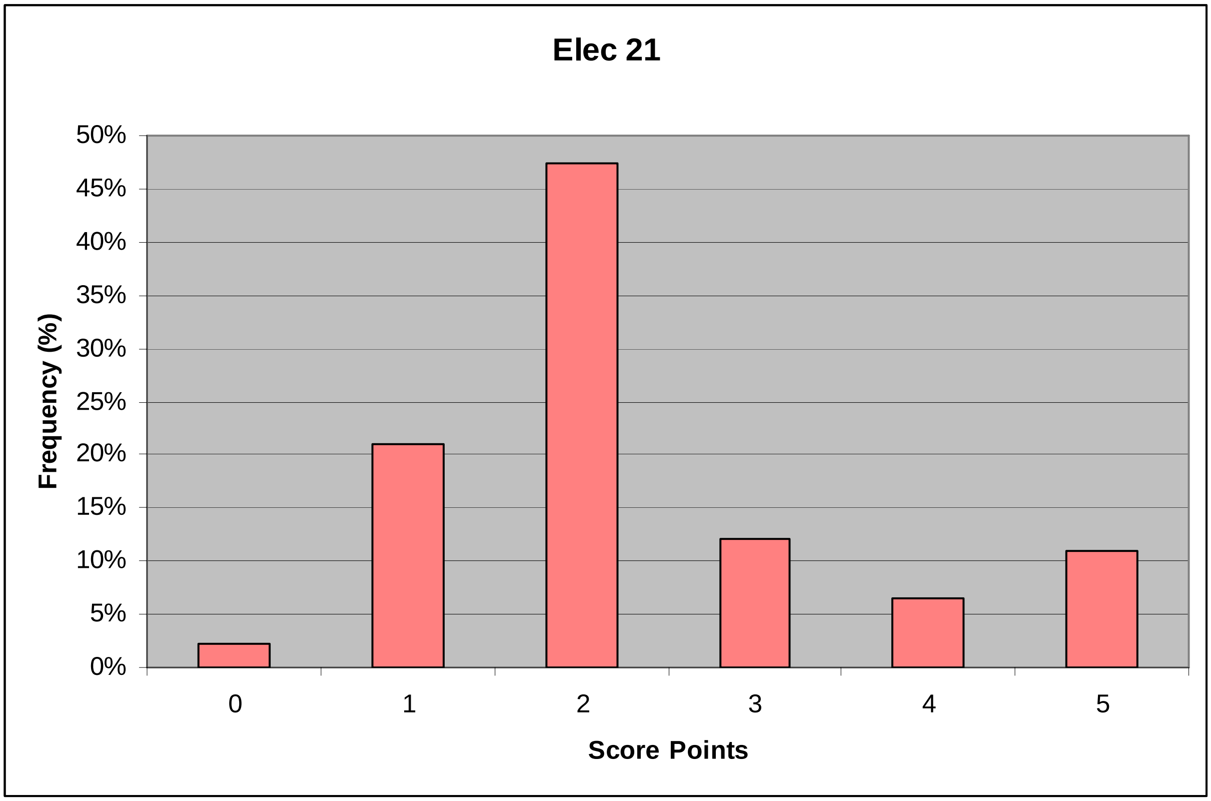


Figure 2. Distribution of student responses to Electricity OE Question 1.

The first open-ended question involving the understanding of Electricity asked students to explain how they would connect a battery and light bulbs to form a parallel circuit. Figure 4 shows the distribution of approximately 180 student responses to this question after scoring by six raters. The student responses for this item were as expected, and virtually none of the responses were blank or unintelligible. The inter-rater reliability was high (67% exact agreement among raters) and the Kappa statistic (.535) was moderate, suggesting that variability in scores was due more to student rather than rater differences. This is confirmed by the single measure ICC statistic (.862) suggesting that more than 86% of the variance in scores comes from student (rather than rater) differences. Moreover, there was only limited drift in ratings over time. Raters matched their previous scores 81% of the time.

The second question involving Electricity asked students to explain what would happen when one of two bulbs in a parallel circuit is removed. As shown in Figure 5, the distribution of almost 160 student responses to this question after scoring by six raters was as expected, and relatively few of the responses were blank or unintelligible. Here again, the inter-rater reliability was high (77% exact agreement among raters) and the Kappa statistic (.631) was substantial, suggesting that scores on this item overwhelmingly reflected student rather than rater differences. As expected, the single measure ICC statistic (.782) suggests that almost 80% of the variance in scores comes from student (rather than rater) differences. Repeated scoring of student responses by raters shows that these raters matched their previous scores 90% of the time.

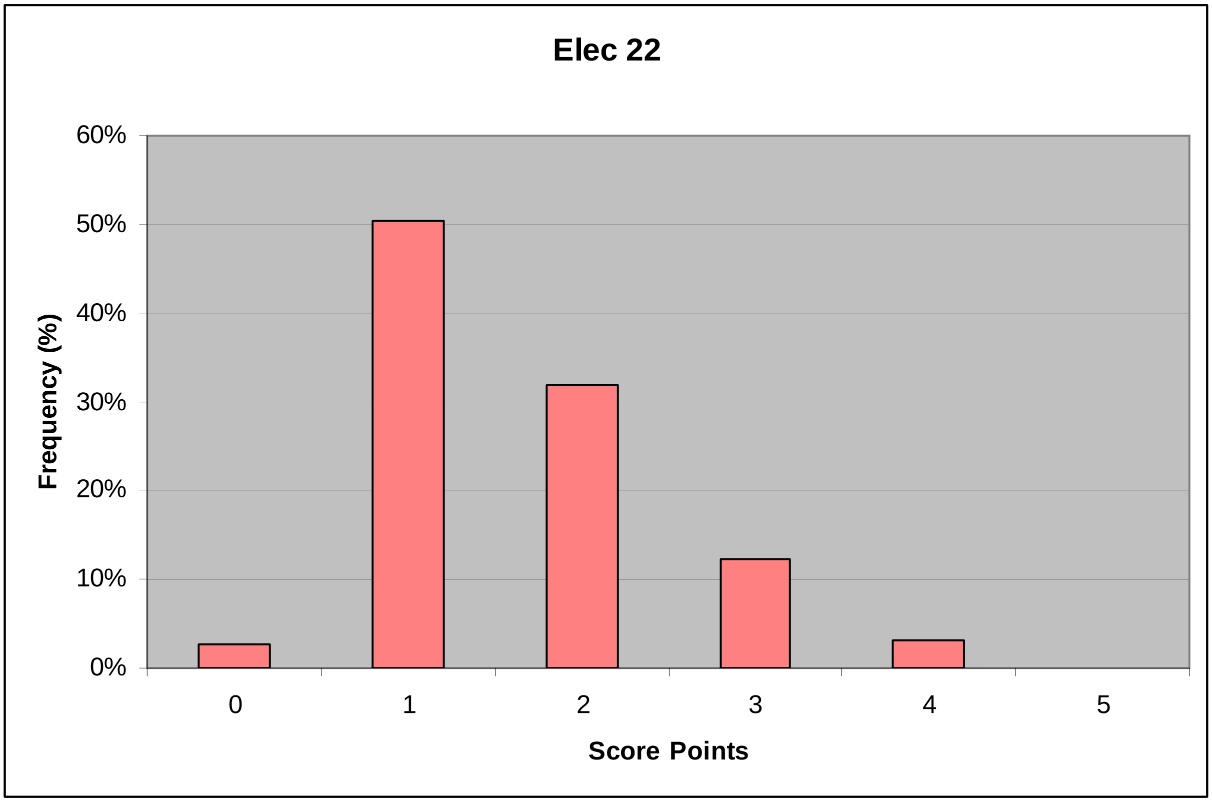


Figure 3. Distribution of student responses to Electricity OE Question 2.

Student median ratings on the two open-ended Electricity questions are significantly correlated ($r = .275$, $p < .001$) indicating that the two questions are measuring similar, but not identical, content.

As can be seen in Table 7, the variance components for the open-ended Electricity items suggest that nearly half of the total variance in our results (43%) comes from differences in the student test takers, while virtually none of the variance in student scores comes from the two items and very little of the variance (2%) is contributed by raters. The interaction of the student with the items accounts for more than one third (36%) of the variability in our results, while the remaining variance (19%) comes from measurement error and from interactions we cannot separate given our experimental method (e.g., rater[student] interaction with item). The variance components for these items suggest that the items are, in fact, responding very strongly to differences in students, and the interaction between these student differences and the items.

Table 7

Variance Estimates for the Two Open-ended Electricity Items

Component	Variance estimate	% of variance
Student	.40	43%
Item	.00	0%
Rater (Item)[a]	.02	2%
Student × Item	.33	36%
Error	.18	19%
Total variance	.93	100%

[a] Since the Electricity items were used for training purposes, there was no overlap in raters between the two open-ended items for Electricity. Consequently, raters are nested within item for this analysis.

The first open-ended question involving Properties of Matter asked students to determine where on the periodic table a shiny, malleable, and conductive element would be found and why they answered in the way that they did. We received almost 150 student responses. As shown in Figure 6, the distribution of student responses after scoring by four raters was generally as expected, although 8% of the students did not answer or supplied answers that were unintelligible. Inter-rater reliability was high (78% exact agreement among raters) and the Kappa statistic (.713) was substantial, suggesting that variability in scores was due more to student rather than rater differences or chance agreement.

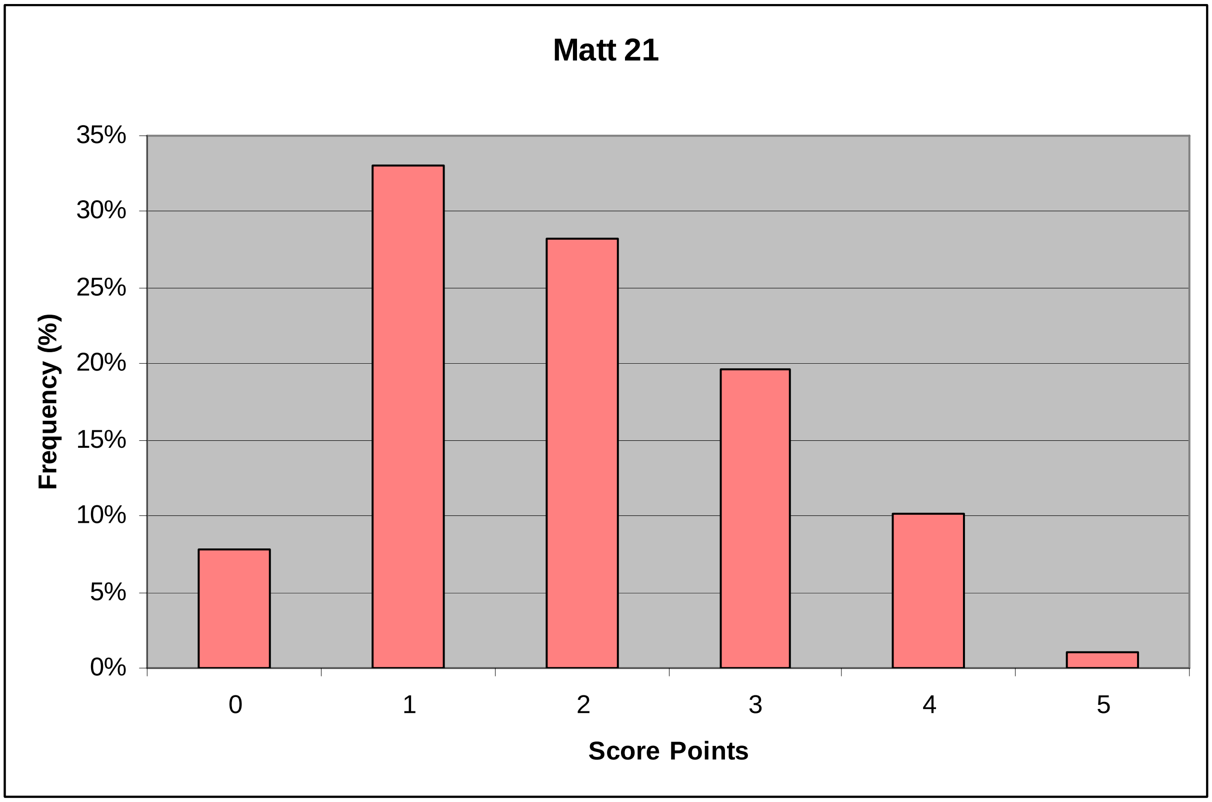


Figure 4. Distribution of student responses to Properties of Matter OE Question 1.

In fact, the single measure ICC statistic (.892) implies that almost 90% of the variance in scores comes from student (rather than rater) differences. There was little drift in ratings over time as raters matched their previous ratings 87% of the time.

The second open-ended question involving Properties of Matter asked the 170 students who responded to explain how they would separate a mixture of different compounds. As shown in Figure 7, the distribution of student responses to this question after scoring by four raters was largely distributed among score points 1-3; however, a number of students (almost 10%) did not supply scoreable responses and very few students earned a “mastery” score (4) on their response.

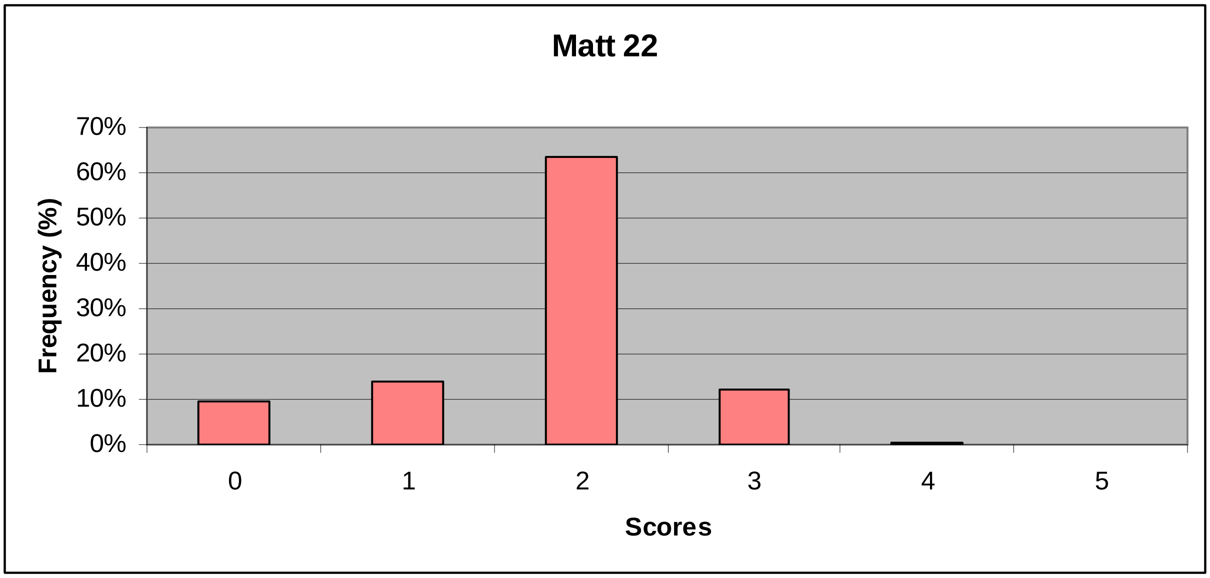


Figure 5. Distribution of student responses to Properties of Matter OE Question 2.

The raters agreed on exact scores 72% of the time and the Kappa statistic (.631) suggests this was moderate. In fact, the single measure ICC statistic (.739) suggests that approximately 74% of the variance in scores comes from student (rather than rater) differences. Repeated scoring of student responses by each rater shows, however, that these raters only matched their previous scores 73% of the time. Student median ratings on the two open-ended Properties of Matter questions are significantly correlated (r = .365, p < .001), suggesting that the two items were measuring similar content.

Table 8 reports the variance components for the open-ended Properties of Matter items. Approximately 29% of the total variance in our results arises because of student differences. Differences between the two items and between raters contribute virtually no variability to the results. The interaction of individual students with the items accounts for more than half (55%) the variability in our results. The remaining variance (17%) comes from measurement error and from interactions we cannot separate given our experimental method. The variance components for the items suggest that they are, in fact, responding strongly to differences in students and the interaction between these student differences and the items.

Table 8

Variance Estimates for the Two Open-ended Properties of Matter Items

Component	Variance estimate	% of variance
Student	.31	29%
Item	.00	0%
Rater (Student)	.00	0%
Student × Item	.58	55%
Error	.18	17%
Total variance	1.06	100%

The first open-ended question involving the Water Cycle asked students to explain where the condensation on the side of a glass “comes from” and how that condensation might be increased. Figure 8 shows the distribution of approximately 220 student responses to this question after scoring by four raters.

While all the responses were scoreable, within the expected range of 1–4, and exact agreement among raters was 74%, the Kappa score showed below moderate correlation (.368). Moreover, the ICC statistic (.497) suggests that more than half of the variability in scores was due to rater rather than student differences. This was the case even though raters matched their previous scores 86% of the time. These results could indicate that the rubric left great leeway for interpretation of student responses among scorers.

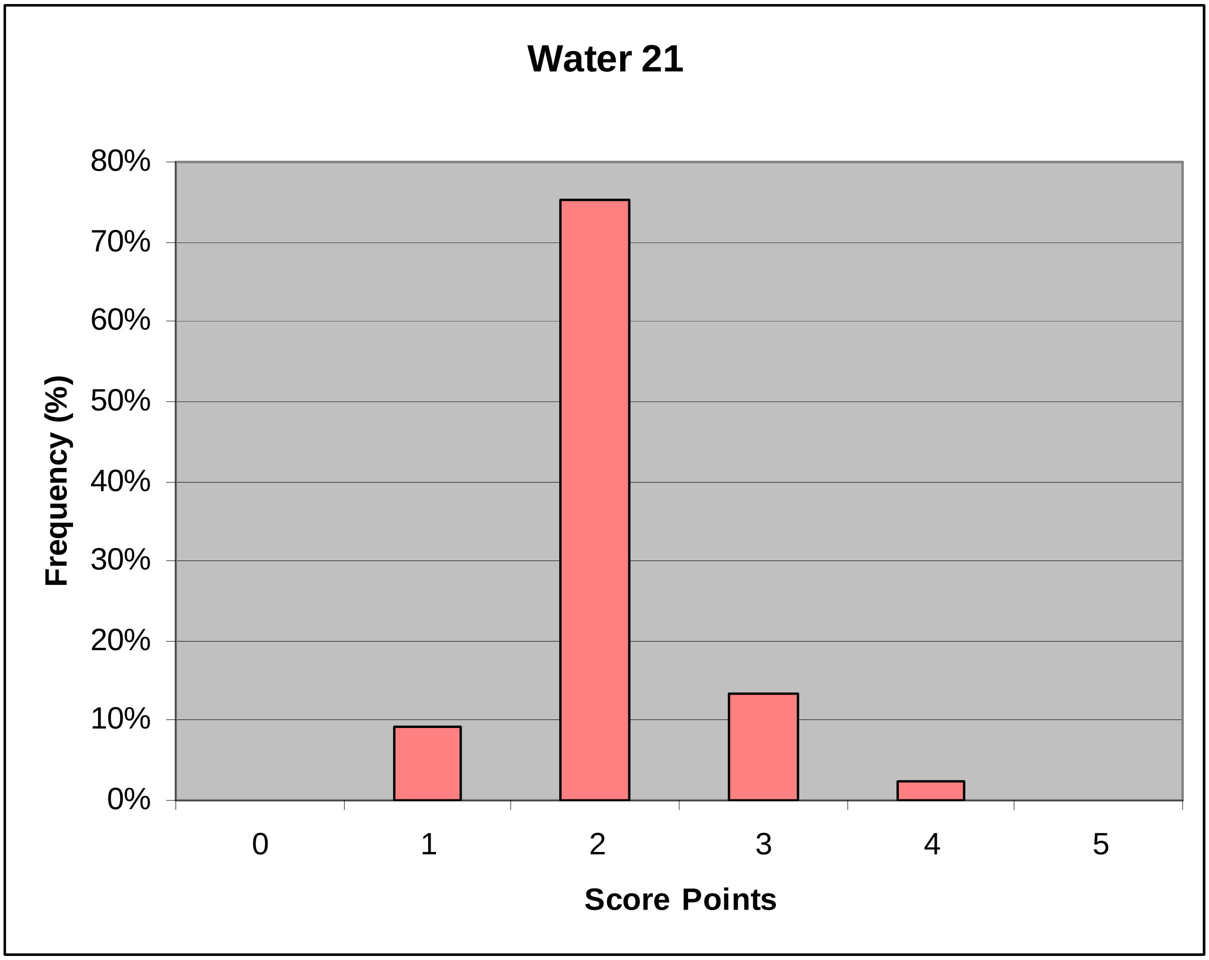


Figure 6. Distribution of student responses to Water Cycle OE Question 1.

The second open-ended question involving the Water Cycle asked students to use their knowledge of erosion and deposition to explain the causes of changes in a given river over a period of 55 years. Figure 9 shows the distribution of approximately 220 student responses to this question after scoring by four raters. As was the case in the first open-ended Water Cycle question, student responses were almost entirely distributed among score points 1–4. Nevertheless, exact agreement among raters only occurred for 53% of the student responses and the single measure ICC statistic (.612) indicates that only a little more than half the variance in scores comes from student (rather than rater) differences. Kappa (.289) suggests only fair agreement among raters when chance agreement is removed. There was little drift in these ratings over time. As was the case in the first open-ended water cycle question, raters matched their previous scores about 80% of the time.

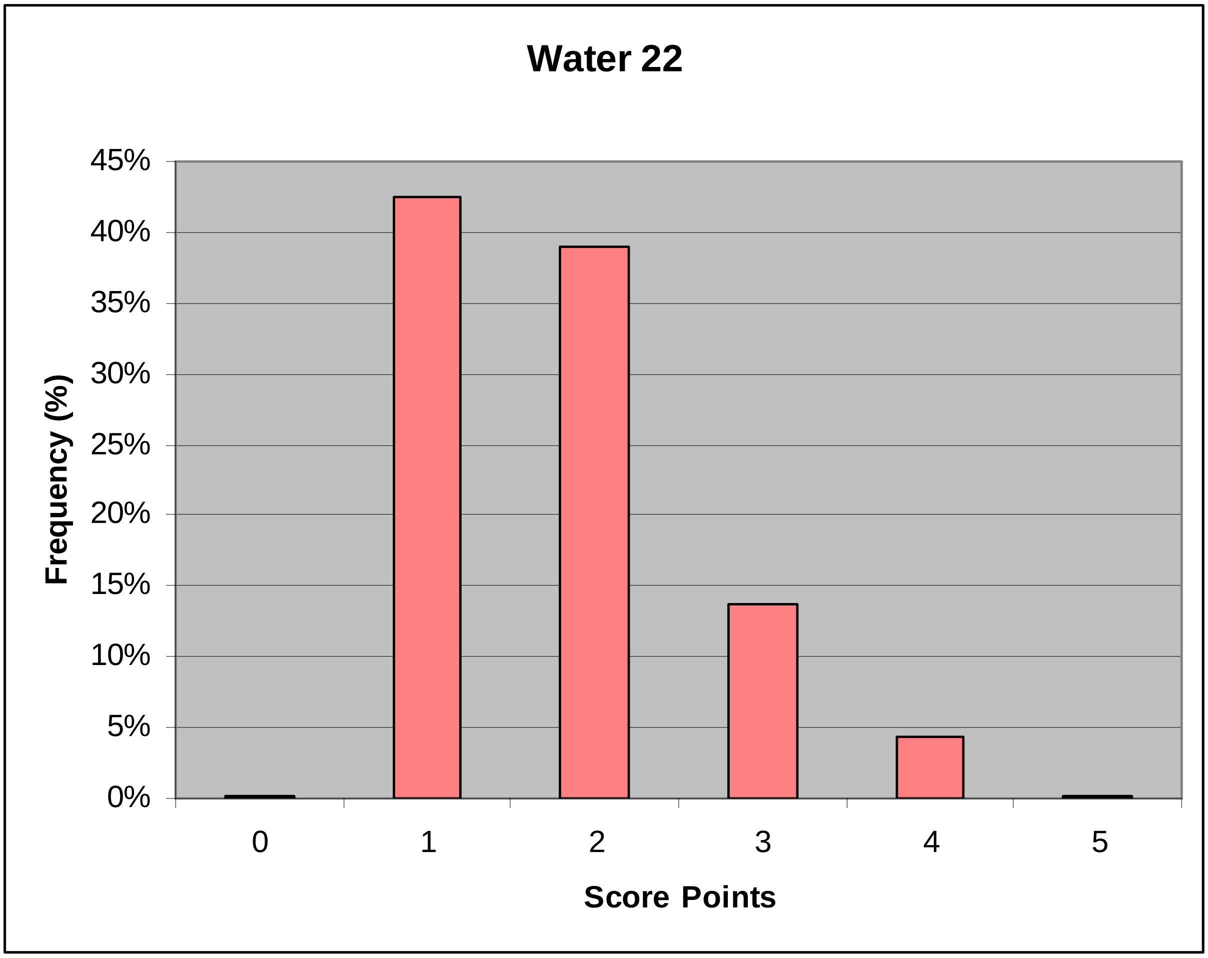


Figure 7. Distribution of student responses to Water Cycle OE Question 2.

The two open-ended questions on the Water Cycle are significantly correlated (*r* = .350, *p* < .001) suggesting that the two items were measuring similar, though not identical, content.

As can be seen in Table 9, approximately 16% of the total variance in our results arises because of student differences, 5% of the total variance is explained by differences between the two items (e.g., the items might be testing different aspects of the water cycle or may be more difficult for all students), 12% of the variability results from the raters (given the students they looked at), and 41% of the variance results from an interaction of students and items. The remaining variance comes from measurement error and from interactions we cannot separate given our experimental method (e.g., rater[student] interaction with item). The variance components for these items suggest that the items are, for the most part, responding to differences in students, and the interaction between these student differences and the items.

Table 9

Variance Estimates for the Two Open-ended Water Cycle Items

Component	Variance estimate	% of variance
Student	.10	16%
Item	.03	5%
Rater (Student)	.07	12%
Student × Item	.24	41%
Error	.15	26%
Total variance	.59	100%

Discussion and Conclusion

Decades of research suggest that the overwhelming majority of assessments used in American classrooms are selected or developed by teachers. This same research also suggests that such tests often lead to inaccurate inferences and can have detrimental effects on student self-efficacy. Locally developed assessments that accurately predict student performance on high-stakes assessments, that allow precise estimates of current understanding, and that correctly identify current misconceptions should be useful in making instruction both more effective and more efficient. The question investigated by this study was whether a collaborative process involving teachers in a systematic assessment development process could be effective in developing selected-response and open-ended assessments of demonstrable high technical quality.

Based on the analysis of the data from our pilot of the multiple-choice items, we found that CAESL-trained teachers had little difficulty collaborating to develop multiple-choice test items of high technical quality for knowledge domains in which they had teaching experience. In the case of our participants, this teaching experience did not necessarily equate to their current teaching assignment. Of the 140 multiple-choice items developed, only 5 (less than 4%) did not meet minimum technical quality specifications (1-pl model, etc.). Of these 5 items, 2 could easily have been modified to reflect a single answer. Consequently, developers had little difficulty choosing items and assembling a test that both covered the conceptual flow and generally matched the state CST specifications for item difficulty, mean biserial correlation, and minimal biserial correlation. As the experience of other researchers suggests, we found that tying assessment development to instructional goals (as specified in both the state standards and the conceptual flow) and considering the ultimate use of the assessment were key in focusing developers to identify the major ideas that needed to be assessed. Prior to developing any assessment items, participants clearly established key conceptual goals for their instruction (the conceptual flow), matched standards to those goals, and focused on the cognitive demands required to demonstrate proficiency on each assessed concept. This precursory effort seemed to result in clearly written item prompts and distracters that focused on assessing a particular concept at a specific cognitive level.

The open-ended items developed were of similar technical quality. While only the items needed to assess concepts in the conceptual flow were ultimately scored due to time constraints, the scored items generally functioned well. Aside from the second open-ended Water item, the raters were able to score student responses with a high degree of reliability (.63 to .78 exact agreement). The relatively low inter-rater reliability on the second open-ended Water item suggests that the rubric for that item might need clarification. Our G-study results point to a similar conclusion. While the rater(student) facet contributed only about 12% of the total variability in this case, this percentage was much larger than that contributed by the rater facet on the other three open-ended problems. Moderate to high Kappa statistics for each of the open-ended items lead us to a similar conclusion, suggesting that rater agreement on scores is probably not caused by mere coincidence but by a combination of raters and rubrics. The two minor exceptions to this conclusion are, again, the open-ended Water items. Unlike the other open-ended items, each of these items demonstrated only fair inter-rater agreement beyond what mere chance would explain.

Our analysis of scores for the open-ended items in general indicated that a very large proportion of the variance in scores is the result of student variation rather than variability introduced by the raters or scoring rubrics. For all but two of the items, student variation accounted for more than three fourths of the total variability in scores. The one major exception to this trend was the first open-ended Water item. Although raters agreed on how to score student responses to this item a high percentage (74%) of the time, slightly more than half of the variability in scores seems to come from sources other than student differences. This suggests that the item and rubric probably need to be rewritten in order to more accurately assess student understanding and ability. Scores from the second open-ended Water item also seemed to be affected, to a greater degree than the non-Water items, by something other than student variability. Although about 60% of the variance in scores can seemingly be explained by actual student differences, this is well below the proportion of variance explained by other open-ended items in the study.

Finally, we found that when raters applied the scoring rubrics to rescore previously scored open-ended items, they generally agreed with their first score about 80% of the time. While acceptable, this intrarater reliability statistic might be improved by clarifying the rubrics so they more clearly articulate the differences between score points. In addition, it suggests that raters using these rubrics might need to score fewer papers in a single session or undergo recalibration for scoring sessions lasting four hours (the length of our typical scoring session) or more.

The distribution of student scores was generally as expected. With the exception of the two open-ended items that addressed Ecological Systems, the items had a wide range of student responses with most of the responses in the 1-3 score point range. We noticed no ceiling or floor effects, suggesting that the items were able to distinguish student abilities. While we are not presently able to conclude that these distinctions are pedagogically important, we do know that the results discriminate student abilities based on criteria that our content experts felt were relevant to conceptual understanding in each of the four domains. Although the two open-ended items addressing Ecological Systems did seemingly function to categorize student understanding, the fact that each question generated a significant number of non-responses indicates that the questions may have been difficult for students to understand, that time may have been an issue, or that the topic may have not yet been taught in classrooms that pilot tested the items.

Finally, we conducted a measurement of item intercorrelation on each pair of open-ended items in each domain. In each domain, the two open-ended items were significantly correlated, suggesting that the items were, in fact, measuring similar conceptual knowledge. Of note however, was that the intercorrelation coefficient for the two open-ended Electricity and Magnetism items was lower than that coefficient in the other domains. Since both Electricity and Magnetism items dealt with parallel circuits, we expected the rho coefficient to be the highest between items in this domain. The G-study, however, did suggest that item differences between these two items contributed virtually nothing to the variability in student scores.

In the end, this process yielded one benchmark assessment and an archive of additional assessment items in each of the four knowledge domains. Each assessment was composed of 20 selected-response and 2 open-ended items. Although previous research has suggested that locally developed assessments are often of low technical quality, the results presented here suggest that teachers and district personnel can develop high-quality multiple-choice and open-ended items *if provided an appropriate framework and supportive context for doing so*. In part, the item quality described seems to result from the fact that developers were very clear about instructional goals (e.g., conceptual flow, standards, and the degree of student conceptual understanding) and how the assessments were to be used in the classroom before designing the items (the CAESL framework). While important in designing multiple-choice items, this scaffolding seems to be especially important when developing open-ended items. In our study, the ADDS system appeared to facilitate this process. Although open-ended items developed without using the ADDS generally demonstrated acceptable quality, the three open-ended items (in the domains of Ecological Systems, Electricity and Magnetism, and Properties of Matter) developed using the ADDS framework demonstrated the highest overall quality. In addition, these items did not require substantial revision before the field test. This could mean substantial savings in time and cost (less developer time, fewer pilot tests, etc.) if large numbers of items were being developed. As the number of items compared in this study is small, this conclusion is preliminary and requires further investigation.

Similarly, while the results presented here suggest that the CAESL process itself was an important factor in the development of high-quality assessment items, we are unable to separate out the effects that the three years of prior professional development had on these results. Our findings do suggest that it might be profitable to attempt a similar study using teachers with five or more years of teaching experience in a subject area, but no prior professional development in assessment selection and use.

Using the benchmark assessments developed in this study, results from the fifth grade CST, and responses from teachers, we hope to address the question of whether student results on these tests can reliably predict performance on high-stakes, state assessments. Such predictive power would allow teachers to use these benchmark assessments to identify and redress learning deficiencies, and could have positive results not only for student test scores, but for student self-efficacy as well.

References

Baker, E. L. (1998). Model-based performance assessment. Los Angeles, UCLA.

Baker, E. L. (2004). Assessing and monitoring performance across time and place. U.S. Department of Education Secretary's No Child Left Behind Leadership Summits "Empowering Accountability and Assessment Using Technology," St. Louis, MO.

Birenbaum, M., & Tatsuoka, K. K. (1987). Open-ended versus multiple-choice response formats—it does make a difference for diagnostic purposes. Applied Psychological Measurement, 11(4), 385-395.

Birenbaum, M., Tatsuoka, K. K., & Gutvirtz, Y. (1992). Effects of response format on diagnostic assessment of scholastic achievement. Applied Psychological Measurement, 16(4), 353-363.

Black, P., Harrison, C., Lee, C. , Marshall, B., & Wiliam, D. (2003). Assessment for learning: Putting it into practice. Buckingham, UK: Open University Press.

Black, P., & Wiliam, D. (1998). Assessment and classroom learning. Assessment in Education, 5(1), 7-74.

Black, P., & Wiliam, D. (2004). The formative purpose: Assessment must first promote learning. In M. Wilson (Ed.), Towards coherence between classroom assessment and accountability (pp. 20-50). Chicago, IL: University of Chicago Press.

Brown, J. D. (1988). Understanding research in second language learning: A teacher's guide to statistics and research design. Cambridge, UK: Cambridge University Press.

DiRanna, K. (Ed.). (in press). Assessment centered teaching: A reflective practice. Thousand Oaks, CA: Corwin Press.

Docy, F., Moerkerke, G., DeCorte, D., & Segers, M. (2001). The assessment of quantitative problem-solving skills with "none of the above" (NOTA) items. European Journal of Psychology of Education, 16(2), 163-177.

Educational Testing Service. (2006). California Standards Tests Technical Report Spring 2005 Administration. Retrieved from http://www.cde.ca.gov/ta/tg/sr/documents/startechrpt05.pdf

Fleiss, J. L., Nee, J. C. M., & Landis, J. (1979). Large sample variance of kappa in the case of different sets of raters. Psychological Bulletin, 86, 974-977.

Frisbie, D. A., Miranda, D. U., & Baker, K. (1993). An evaluation of elementary textbook tests as classroom assessment tools. Applied Measurement in Education, 6(1), 21-36.

Gearhart, M., Nagashima, S., Pfotenhauer, J., Clark, S., Schwab, C., Vendlinski, T., et al. (2006). Developing expertise with classroom assessment in K-12 science: Learning to interpret student work interim findings from a two-year study. Educational Assessment, 11(3 & 4), 237-263.

Haladyna, T. M., & Downing, S. M. (1989). Validity of a taxonomy of multiple-choice item-writing rules. Applied Measurement in Education, 2, 51-78.

Herman, J., & Baker, E. (2005). Making benchmark testing work. Educational Leadership, 63, 48-54.

Herman, J. L., Osmundson, E., Ayala, C., Schneider, S., & Timms, M. (2005, April). The nature and impact of teachers' formative assessment practices. In J. L. Herman (Chair), Building science assessment systems that serve accountability and student learning: The CAESL model. Symposium conducted at the annual meeting of the American Educational Research Association (AERA), Montréal, Canada.

Kane, M. (2006). Content-related validity evidence in test development. In S. M. Downing & T. M. Haladyna (Eds.), Handbook of test development (pp. 131-154). Mahwah, NJ: Lawrence Erlbaum Associates.

Kane, M. B., Khattri, N., Reeve, A., & Adamson, R. (1997). Assessment of student performance: Studies of education reform, U.S. Department of Education. Washington, DC: Office of Educational Research and Improvement.

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. Biometrics, 33, 159-174.

Livingston, S. A. (2006). Item analysis. In S. M. Downing & T. M. Haladyna (Eds.), Handbook of test development (pp. 421-444). Mahwah, NJ: Lawrence Erlbaum Associates.

McMorris, R., & Boothroyd, R. A. (1993). Tests that teachers build: An analysis of classroom tests in science and math. Applied Measurement in Education, 6(4), 321-342.

Messick, S. (1989). Validity. In R. L. Linn (Ed.), Educational measurement (pp. 13-103). New York: Macmillan Publishing Co.

Rodriguez, M. C. (2004). The role of classroom assessment in student performance on TIMSS. Applied Measurement in Education, 17(1), 1-24.

Shrout, P. E., & Fleiss, J. L. (1979). Intraclass correlations: Uses in assessing rater reliability. Psychological Bulletin, 86(2), 420-428.

Stiggins, R. J. (1991). Facing the challenges of a new era of educational assessment. Applied Measurement in Education, 4, 263-273.

Stiggins, R. J. (1994). Student-centered classroom assessment. New York: Macmillan College Publishing.

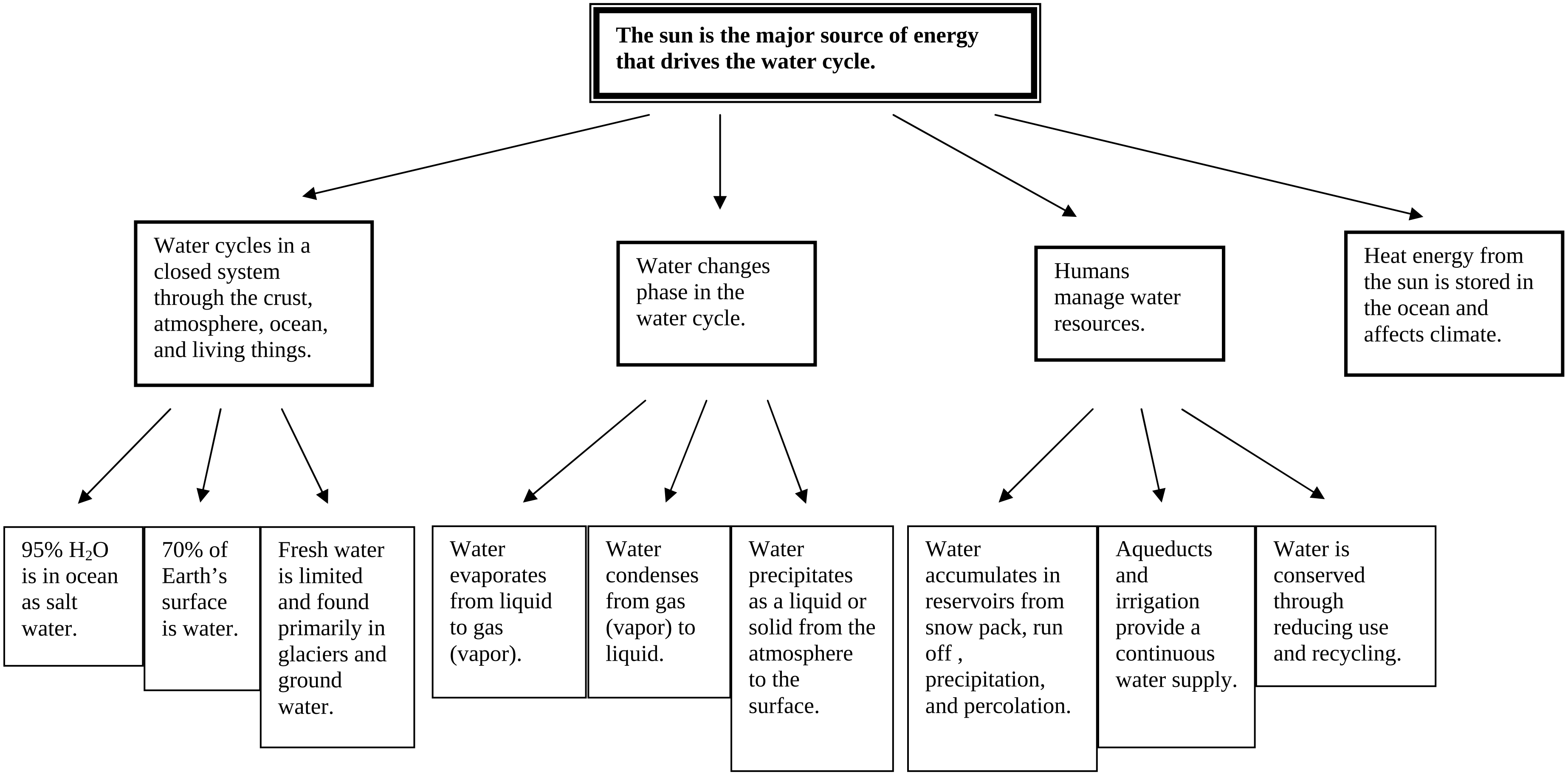
Traub, R. E. (1992). On the equivalence of the traits assessed by multiple-choice and constructed-response tests. In R. E. Bennett & W. Ward (Eds.), Construction versus choice in cognitive measurement: Issues in constructed response, performance testing and portfolio assessment (pp. 236-256). Hillsdale, NJ: Lawrence Erlbaum Associates.

Viera, A. J., & Carrett, J. M. (2005). Understanding interobserver agreement: The kappa statistic. Family Medicine, 37(5), 360-363.

Welch, C. (2006). Item and prompt development in performance testing. In S. M. Downing & T. M. Haladyna (Eds.), Handbook of test development (pp. 303-327). Mahwah, NJ: Lawrence Erlbaum Associates.

Appendix

Hierarchy of Key Ideas / Conceptual Flow



- **The sun is the major source of energy that drives the water cycle.**
  - Water cycles in a closed system through the crust, atmosphere, ocean, and living things.
    - 95% H_2O is in ocean as salt water.
    - 70% of Earth's surface is water.
    - Fresh water is limited and found primarily in glaciers and ground water.
  - Water changes phase in the water cycle.
    - Water evaporates from liquid to gas (vapor).
    - Water condenses from gas (vapor) to liquid.
    - Water precipitates as a liquid or solid from the atmosphere to the surface.
  - Humans manage water resources.
    - Water accumulates in reservoirs from snow pack, run off , precipitation, and percolation.
    - Aqueducts and irrigation provide a continuous water supply.
    - Water is conserved through reducing use and recycling.
  - Heat energy from the sun is stored in the ocean and affects climate.