THE TRANSFORMING INSTRUCTION BY DESIGN IN
EARTH SCIENCE (TIDES) PROJECT

Report on Technical Quality of 6th and 8th Grade Student Assessments

Prepared for:
Dr. Harold Himmelfarb
Institute of Education Sciences
U.S. Department of Education

Prepared by:
William R. Penuel
Lawrence Gallagher
Patty Kreikemeier
SRI International

SRI Project P16699
ED Grant # R305M050226

May 2006

Introduction and Overview

This report describes activities undertaken to create, pilot, and analyze technical quality of student assessments in sixth and eighth grades for the Transforming Instruction by Design in Earth Science (TIDES) project. For the project, middle school teachers are randomly assigned to one of four conditions to test the efficacy of an Understanding by Design approach to improving teacher quality. The first condition, *Investigating Earth Systems* (IES), will provide assigned teachers with training in the use of a 9-week curriculum unit designed by Earth science experts. The second condition, *Earth Science by Design* (ESBD), will provide assigned teachers with professional development aimed at helping them design their own standards-aligned 9-week curriculum units. The third, *Hybrid* condition will provide assigned teachers with professional development designed to help them adapt IES curriculum materials using principles from the ESBD program. For the fourth *control* condition, assigned teachers will be free to participate in whatever professional development they might have otherwise taken part in for the duration of the study.

Student scores on the tests we have developed will be used, alongside the SAT-9 and FCAT scores (8th grade only) as outcome measures for the project. The goal of these assessments is to measure understanding of Earth science concepts of students whose teachers have volunteered to be part of the study. The particular concepts for which we have developed tests are aligned to Florida's Sunshine State Standards for 6th, 7th, and 8th grades, as well as to Duval County Public Schools' interpretation of those standards and to the National Science Education Standards for inquiry in grades 5-8 (National Research Council, 2000).

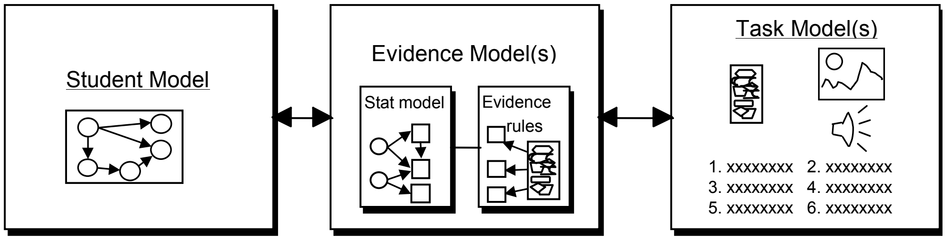
We have developed pilot versions of three separate unit tests, one each for 6th, 7th, and 8th grades. Students are likely to have opportunity to learn only those concepts their teachers are required to teach; therefore, separate tests for each grade are necessary, since most of the Earth science concepts that are in the Sunshine State Standards appear only in one of these grades. Only the concept that energy travels in waves is part of the standards of each of these grades.

As we indicate below, both sets of tests we have piloted have adequate reliability for use in an experimental study and a coherent factor structure suggesting that underlying constructs are being measured well. However, not all items developed and tested were judged to be good indices of deep understanding of constructs. Therefore, some items from our original pilot will be eliminated; others may be adapted, and we will develop some new items for constructs where there were too few items that probed for deeper understanding.

Assessment Design Process

Contemporary approaches to assessment design all point to the need to begin with a clear understanding of the student skills and understandings that instructional activities are designed to develop (Bransford & Schwartz, 2000; Messick, 1992). Consequently, our approach to developing assessments of student learning follows an *evidence-centered* approach (Mislevy, 1994; Mislevy, Steinberg, Almond, Haertel, & Penuel, 2003). Figure 1 shows the key components of an evidence-centered approach to test design.

Figure 1. Evidence-Centered Assessment Design Framework



The **student model** variables approximate selected aspects of the infinite configurations of skill and knowledge real students have, as seen from some perspective about skill and knowledge in the domain, whether drawn from cognitive science or standards, or some combination of the two. The number and nature of the student model variables in an assessment also depend on its purpose. A single variable characterizing overall proficiency might suffice for an assessment just meant to support a pass/fail decision, more detailed variables would be needed for formative assessments that are designed to provide students with detailed feedback to develop proficiency in an area such as inquiry.

The **evidence model** for an assessment answers the question, "What behaviors or performances should reveal the relevant knowledge and skills described in the student model?" An evidence model specifies how observations for a given task situation constitute evidence about student model variables. Figure 1 shows that there are two parts to the evidence model. One part ("evidence rules") is about evaluating the key features of what the student says, does, or creates in the task situation—the "work product." These are the "observable variables," evaluations of whatever the designer has determined are the key aspects of the performance. These observables may be derived from sources similar to those used to develop the student model, such as research on teaching and learning, socially valued objectives for instruction, or standards. The other part ("stat model") of the evidence model is about the way that the observable variables depend, in probability, on student model variables. This is how we combine evidence across tasks. Familiar psychometric models, such as item response theory and latent class models, can be seen as special cases of these ideas.

Having thought through the behaviors or performances that reveal the constructs the assessment is targeting, we create a **task model** from answering the question, "What tasks or situations should elicit those behaviors?" A task model provides a framework for constructing and describing tasks (i.e., the situations in which examinees act). The variables one uses to describe tasks play many roles, such as guiding task construction, focusing the skills a task elicits, and providing an operational definition of the student model variables (Mislevy, Steinberg, & Almond, 1999). A task model includes specifications for the task environment, including, for example, characteristics of stimulus material, instructions, help, and tools. It also includes specifications for the work product, the form in which what the student says, does, or produces will be captured.

Although the spatial layout of Figure 1 suggests a temporal order to the assessment design process, in practice it is more a matter of iterative bootstrapping than one of discrete sequential steps. We begin with a set of learning outcomes we want to include in the student model, but as we become involved in creating tasks that provide a context for eliciting those learning outcomes (and later as we field test the assessment with students), we often develop new insights into their nature and limitations. These insights may modify the student model, the evidence model, the task model, or all three.

The Student Model for the TIDES Assessments

In our test development process, the primary source for developing a student model of the knowledge, skills, and abilities we expected students to master was Florida's *Sunshine State Standards* and Duval County Public Schools' elaboration of those standards. We chose the state's science standards as the guide for us, because all teachers—including those in the control condition—are expected to teach to those standards. The standards thus provide a fair basis for test design in two respects: (1) students in all conditions are expected to have similar opportunities to learn the content of the standards; and (2) the tests make it possible to measure whether our professional development help teachers do what they are expected to do, whether or not they participate in the study.

Our tests do not attempt to measure the breadth of standards taught in a particular grade level, however. Instead, we have focused on standards in *Earth and Space Science,* since these are the focus of our study. It is possible that there will be "spillover" impacts of professional development on other domains of science; however, the cost of designing assessments prohibited us from developing items in physical and biological science. Further, by focusing on Earth and space science, we were able to develop more items to measure possible impacts on student achievement than we would have been able to develop had we sought to measure all the standards.

The Evidence Model (Rules) for the TIDES Assessments

We chose to develop our evidence model from the theory of understanding outlined in Wiggins and McTighe's (1998) *Understanding by Design*. We used this framework for defining what understanding means, because our study is a study of the impact of three different approaches to preparing teachers to implement a UbD approach in their classroom. This definition is also fair to apply to our control group, because the school district expects all teachers to *teach for understanding* as defined by the UbD framework. According to the UbD definition of understanding, students' can demonstrate understanding of concepts in at least six different (but overlapping ways):

- By **explaining** a phenomenon using the concept;
- By **interpreting** a situation, event, or text using the concept;

- By **applying** what they know about the concept to a novel situation;
- By **taking or identifying a perspective** on the concept;
- By **empathizing** with others who may have discovered or been critical of the concept; and/or
- By **reflecting** critically on what they do and do not know about the concept.

Tests that measure understanding certainly require students to be able to recall facts; but our tests are intended to do more than just test recall. We have aimed to include many items that require students to explain, interpret, and apply concepts. We have also sought to include a few items that require students to take or identify someone else's perspective on a situation, empathize with others, and reflect on what they do or do not know about a concept.

The Task Model for the TIDES Assessments

In late spring 2005, for the sixth grade test we relied on Sunshine State Standards and the Duval County Public Schools' interpretation of those standards to construct a test blueprint. The test blueprint identified the constructs to be tested, how many items would be needed to measure the construct, and item formats. Initially, our test blueprint identified four broad concepts in the sixth grade standards for which we planned to develop items:

1. Earth is comprised of interacting systems, which are the geosphere, hydrosphere, atmosphere, and biosphere.
2. Earth's surface is constantly changing and its history can be studied through evidence of the changes.
3. Heat energy is a driving force in the changes and constant moving of earth.
4. Vibrations in materials in the Earth set up wave disturbances that can be described and that behave differently in different media.

In eighth grade, we sought to develop items for the following four constructs:

1. Gravity is a universal force that connects all matter and is related to mass and influenced by the distance between objects.
2. The relative positions of the sun, Earth, and moon, account for phases of the moon and tides.
3. The Earth is just one planet among nine in a solar system, in a galaxy filled with billions of stars, in a universe of billions of other galaxies.
4. Technology is essential to science for access to outer space.

For each concept, our goal was to develop enough items to yield 3-4 items that could be used in the unit test. For this purpose, we anticipated needing a minimum of 6-9 items for each concept. Further, for each construct, we sought to develop items using a mix of item formats, including multiple-choice, short response, gridded response, and constructed response items.

The construction of pilot items was undertaken in three steps. On the basis of the test blueprint, a senior member of the SRI team, Ms. Patty Kreikemeier, developed a notebook and training session intended to help current ESBD teachers in Duval County Public Schools (who will not be part of the study) to develop draft items. Involving current ESBD teachers in the process, we believed, would help us to ensure that items would be aligned to content taught in Duval County and that would also test for different dimensions of understanding. On the basis of input from teachers in the training session, Ms. Kreikemeier developed a total of 77 items for sixth grade and 77 for eighth grade. At this time, she also constructed a key and draft rubrics for open-ended items.

Before these items were piloted, we sent them out for review. A geoscientist employed as a consultant to the US Geological Survey acted as a reviewer for scientific accuracy. In addition, we asked the district staff involved in our project and professional development providers to review the items. We asked this group to review items for the accuracy of science content and graphics; accuracy of our judgments of the links to standards; and for suggestions on how to improve the item's overall ability to measure student understanding. On the basis of both sets of reviews, we constructed four separate test booklets, each targeting a particular construct.

The Procedure for Piloting the Assessments

A researcher conducted informal think-aloud interviews with a small sample of students in northern California before piloting the items in Duval County Public Schools. The purpose of these interviews was to establish whether items could be understood by sixth grade items and to establish that when answering items, appropriate cognitive skills were employed by the students. On the basis of these items, the pilot interviews were revised before being implemented with a small group of students in Duval County.

Each of the grade teachers who agreed to be part of the pilot received one of four of revised items to use to administer to their students before they had taught their Earth science unit. Each test booklet was completed by between 30-40 students at that time. The objectives of this first pilot were to determine if items needed serious revision because too few students could understand them and to establish whether the items appeared to tap the same construct, as determined by initial reliability estimates (Cronbach's alpha).

On the basis of this initial pilot, we reduced the number of items to 41 in sixth grade and to 46 in eighth grade. We eliminated items that were too easy ($p > 0.80$) before the item was taught, since these items would not likely be sensitive to instruction. We also eliminated items in which too few students seemed to understand what the question was asking. Finally, on the basis of initial reliabilities, we eliminated some items that detracted from the overall reliability of the scale.

Once teachers had completed their units, we asked the teachers to test all of their students using the revised test. The revised test of 41 items measured all four concepts that sixth grade teachers were expected to teach as part of their Earth science units. A total of 219 students took the sixth grade test. In eighth grade, the revised test included 46 items; a total of 170 students took that test.

Four researchers then scored the unit tests. Several of the items required a 3 or 4 point scoring rubric. A percentage of items were double-scored by two raters, and in these cases, the scores were averaged to produce a final score for that particular item. The item scores were not standardized, and in computing total scores, a 4 point item would count twice as much as a 2 point item.

All items had at least 3 levels of scoring. 0 = blank or not attempted, 1 = attempt but completely wrong response, and levels 2 and above indicated varying degrees of correctness. For the sake of our analysis, we differentiated blank vs. attempted-but-wrong scores. An initial analysis indicated that collapsing the 0 and 1 scores into a single "no credit" score did not improve reliability of the test, which would have happened if students were choosing randomly to leave an item blank or making wild guesses. In fact, the reliability was lowered if these scores were collapsed, supporting the hypothesis that taking a guess was more indicative of latent science knowledge than leaving an item blank.

For the sixth grade test, we did not score eight of the items and two parts of one item (#41), either because well below 20% of the student responses were comprehensible to scorers or because the scoring rubric proved unreliable.

For the eighth grade test, just two items proved too difficult to score at the time of the posttest.

Results: Analysis of the Sixth Grade Unit Test

To analyze the items, we calculated basic difficulty statistics, overall test reliability, and an exploratory factor analysis to determine whether the constructs we intended to test emerged as underlying factors that could explain a significant portion of the variability in students' knowledge.

Difficulty Levels and Overall Reliability

Overall, the items on the test were of low to moderate difficulty. We would expect most items on a test with sensitivity to instruction to be of low to moderate difficulty at the time of a post-test. Table 1 below shows the p-values for each item scored. The p-values correspond to the percentage of students who got the item correct.

Table 1. *P*-values for Items on Sixth Grade Unit Test (Pilot)

Item	*N*	*P*		Item	*N*	*p*
1	219	0.85		23	217	0.33
2	219	0.83		24	217	0.39
3	219	0.50		25	217	0.43
4	219	0.73		26	217	0.84
5	218	0.69		27	217	0.83
6	219	0.53		28	216	0.78
7	217	0.82		29	217	0.75
8	219	0.56		30	217	0.80
9	219	0.76		31	217	0.76
10	219	0.68		32	218	0.65
11	219	0.88		33	218	0.62
12	219	0.95		34	217	0.55
14	219	0.61		35	217	0.42
17	216	0.53		40	218	0.27
19	214	0.51		41a	218	0.31
20	219	0.57		41c	218	0.28
21	214	0.52		41e	219	0.34

After calculating difficulty levels for items, we tested the overall reliability of these items as a single test. As a whole, the entire 34 item scale showed good internal reliability (Cronbach's alpha = 0.82) with this group of students. We expect these students to be similar to students in at least one of the experimental conditions in our study.

Factor Analysis

The raw item scores were then subjected to a factor analysis in order to discern latent knowledge domains tapped by groups of items. One difficulty inherent in this analysis is the confounding of knowledge domains and item bundles. Some groups of items, for example, all share a common graph as a prompt; the items then ask the student to identify varying features of the same graph. In these cases (without much more sophisticated analysis) it is difficult to tease out the common factor of the item bundle from the knowledge tapped.

There is no hard-and-fast rule regarding the appropriate number of factors to extract from a factor analysis. Some have suggested extracting the factors whose corresponding eigenvalues are greater than 1. Others suggest looking at the sorted list of eigenvalues and picking a natural break-point, where adding factors contributes little variation to the total score. The graph below suggests that both criteria converge on a common recommendation: begin by extracting 4 factors. Based on the scree plot (which is merely a sorted list of eigenvalues plotted in a series), it appears that there is one dominant factor and possible 3 smaller factors worth noting. These 4 factors account for 81% of the score variance. Under relaxed conditions, the factors correlated at little more than 0.2, which suggests an orthogonal factor structure may be warranted.

Figure 2. Scree plot of Eigenvalues from Factor Analysis

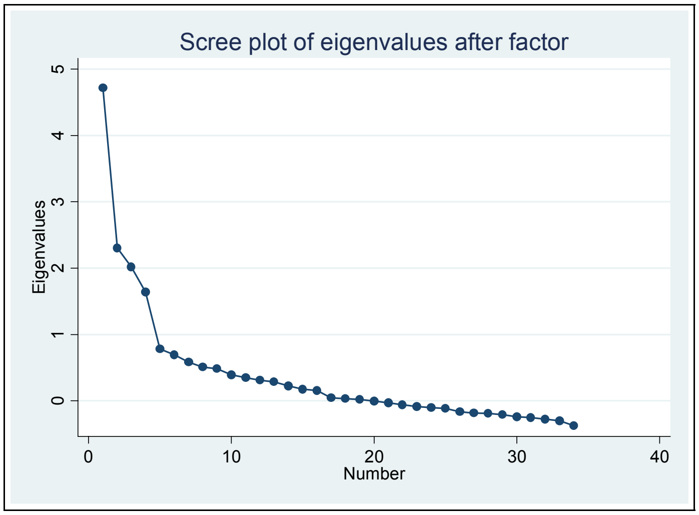


Table 2 on the following page shows the item loadings for each factor. The table shows only those loadings that are greater than 0.30. In only one case did an item load on more than one factor.

Table 2. Factor Loadings and Uniqueness by Item: Sixth Grade Test

Item	Factor 1	Factor 2	Factor 3	Factor 4	Uniqueness
1			0.56		0.63
2			0.33		0.85
3			0.46		0.76
4			0.46		0.74
5			0.55		0.69
6			0.51		0.68
7			0.59		0.64
8			0.55		0.63
9			0.52		0.70
10					0.91
11					0.97
12			0.33		0.83
14				0.38	0.85
17				0.65	0.58
19				0.46	0.76
20				0.42	0.80
21				0.46	0.71
23	0.31			0.38	0.75
24				0.47	0.71
25	0.55				0.63
26		0.62			0.57
27		0.34			0.85
28		0.59			0.62
29		0.59			0.63
30		0.59			0.62
31		0.58			0.65
32		0.49			0.74
33		0.46			0.77
34		0.40			0.76
35		0.35			0.81
40a	0.87				0.24
41a	0.70				0.48
41c	0.76				0.42
41e	0.80				0.35

We then analyzed the reliability of each factor, if items associated with that factor were considered part of a scale. All of the scale had moderate to high reliability. Factor 1 had the highest reliability ($\alpha = 0.85$). Factor 2 had a reliability of $\alpha = 0.76$, and Factor 3 had a reliability of $\alpha = 0.80$. Factor 4 had the lowest reliability, with $\alpha = 0.66$.

As a final step to our analysis, we examined the items associated with particular factors to establish whether the content tested was aligned with the initial set of concepts we set out to test. Interestingly, although there were four factors, they did not map directly onto the four concepts we sought to measure. None of the factors could be said to be measuring the interaction of different spheres within the Earth system. Two of the factors (Factors 1 and 3) related to changes in Earth's surface; one factor included items about plate tectonics and earthquakes, and a second pertained more broadly to analysis of changes in Earth's surface over time. The effects of heat energy were captured by a set of items relating to volcanoes and their effects (Factor 4), and characteristics of waves were measured in another (Factor 2). Because each of these concepts is part of the Sunshine State Standards, we believe our preliminary analysis suggests that our unit test has good reliability with respect to measuring concepts sixth grade teachers are expected to teach.

Further Review of the Items for Content Validity and Alignment with the Evidence Model

Just because a test has good reliability does not ensure that it is a valid measure. It is possible that the items with greater reliability still do not test important science content and may not measure deeper aspects of understanding. We therefore asked a panel of USGS scientists with expertise in different Earth systems—atmospheric, hydrologic, and geologic—to analyze items for their scientific importance. We also asked our team of internal reviewers that included 2 district staff members and 2 representatives from the professional development team to rate the depth and aspects of understanding for each item scored in the pilot test. As Table 3 shows, some further revisions are necessary to add items that test for particular aspects of understanding not tested—specifically perspective, empathy, and self-knowledge. In addition, some items will be dropped either because their scientific accuracy was questioned or because their scientific importance to the field was judged to be peripheral rather than central.

Table 3. Second Review of Items for Scientific Content and Depth of Understanding Tested

Concept
Plate Tectonics and Earthquakes Number of Usable Items: 10 Items Rating Medium or High on Facets of Understanding: 1 (Explanation, Interpretation, Explanation) Items Needing Revision of Scientific Content: 1
Volcanoes Number of Usable Items: 5 Items Rating Medium or High on Facets: 3 (Explanation, Interpretation, Explanation) Items Needing Revision of Scientific Content: 1
Characteristics of Waves Number of Usable Items: 3 Items Rating Medium or High on Facets: 2 (Explanation, Interpretation, Explanation) Items Needing Revision of Scientific Content: None, but 4 items must be dropped entirely
Analyzing Changes to the Earth's Surface over Time Number of Usable Items: 6 Items Rating Medium or High on Facets: 6 (Explanation, Interpretation, Explanation) Items Needing Revision of Scientific Content: 1

Over the summer, we plan to develop additional items on earthquakes and on the characteristics of waves intended to tap more facets of understanding. In addition, for each of the four constructs, we plan to develop 1 to 2 items that measure either perspective or empathy. To analyze self-knowledge, we will develop survey items for students aimed at having them reflect on their overall level of understanding of the concepts included on the test.

Results: Analysis of the Eighth Grade Unit Test

To analyze the items, we calculated basic difficulty statistics, overall test reliability, and an exploratory factor analysis to determine whether the constructs we intended to test emerged as underlying factors that could explain a significant portion of the variability in students' knowledge.

Difficulty Levels and Overall Reliability

Overall, the items on the test were of low to moderate difficulty. We would expect most items on a test with sensitivity to instruction to be of low to moderate difficulty at the time of a post-test. Table 4 below shows the p-values for each item scored. The p-values correspond to the percentage of students who got the item correct (or for polytomously scored items, the mean partial credit score).

Table 4. *P*-values for Items on Eight Grade Unit Test (Pilot)

Item	N	p		Item	N	p
1	170	0.77		25	170	0.79
2	170	0.47		26	170	0.45
3	170	0.45		27	170	0.42
4	170	0.50		28	170	0.35
5	170	0.30		29	170	0.39
6	170	0.40		30	170	0.30
8	170	0.66		31	170	0.74
9	170	0.62		32	170	0.72
10	170	0.57		33	170	0.71
11	170	0.74		34	170	0.66
12	170	0.69		35	170	0.71
13	170	0.75		36	170	0.69
14	170	0.77		37	170	0.48
15	170	0.74		38	169	0.42
16	170	0.38		39	170	0.49
18	170	0.63		40	170	0.33
19	170	0.54		41	170	0.40
20	170	0.42		42	170	0.78
21	170	0.84		43	170	0.60
22	170	0.78		44	170	0.72
23	170	0.86		45	170	0.70
24	170	0.38		46	170	0.70

After calculating difficulty levels for items, we tested the overall reliability of these items as a single test. As a whole, the entire 44 item scale showed good internal reliability (Cronbach's alpha = 0.93) with this group of students. We expect these students to be similar to students in at least one of the experimental conditions in our study (the ESBD condition).

The raw item scores were subjected to a principal components factor analysis. An analysis of the eigenvalues suggests one dominant factor, after which secondary factors contribute little variance. We found interpretable item groupings with a 3 factor solution, accounting for 66% of the score variance. However, when allowing the factors to correlate we find factor inter-correlations ranging from .33 to .51, reinforcing the idea that there is a significant degree of common variance among the 3 factors.

Figure 3. Scree plot of Eigenvalues from Factor Analysis of 8th grade items

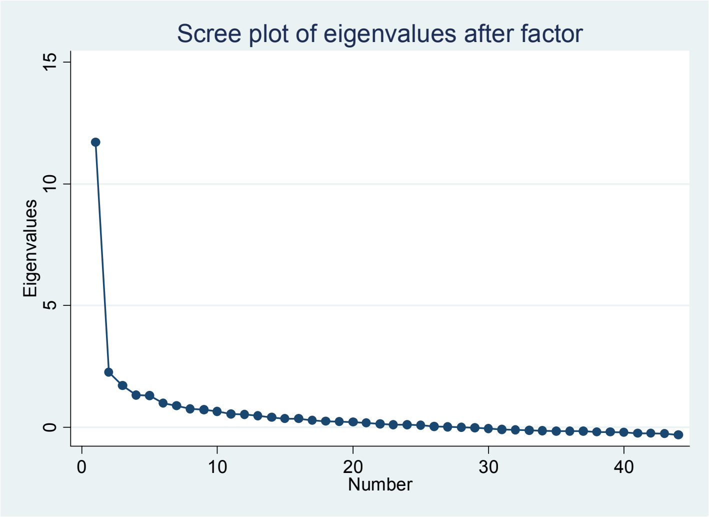


Table 5 on the following page shows the item loadings for each factor. The table shows only those loadings that are greater than 0.30.

Table 5. Factor Loadings and Uniqueness by Item: Eighth Grade Test

Item	Factor 1	Factor 2	Factor 3	Uniqueness
1	0.45		0.36	0.65
2	0.31	0.44		0.64
3	0.42		0.32	0.64
4	0.37		0.45	0.65
5		0.33	0.30	0.80
6		0.51		0.61
8			0.41	0.78
9				0.84
10			0.40	0.70
11	0.56			0.62
12		0.32		0.79
13			0.80	0.33
14			0.76	0.38
15			0.78	0.35
16		0.57		0.66
18	0.36	0.32	0.32	0.67
19		0.60		0.57
20		0.55		0.63
21		0.47		0.70
22		0.40		0.77
23		0.40		0.75
24		0.45		0.77
25		0.60		0.56
26		0.74		0.42
27		0.55		0.62
28		0.60		0.58
29		0.58		0.63
30		0.52		0.69
31	0.77			0.38
32	0.77			0.36
33	0.69			0.50
34	0.72			0.46
35	0.66			0.52
36	0.65			0.53
37	0.45			0.71
38	0.45	0.31		0.69
39	0.40	0.36	0.31	0.61

Table 5 (Cont'd). Factor Loadings and Uniqueness by Item: Eighth Grade Test

Item	Factor 1	Factor 2	Factor 3	Uniqueness
40		0.49	0.32	0.57
41		0.50		0.64
42	0.64			0.49
43	0.54			0.59
44	0.54			0.64
45	0.43			0.70
46	0.47			0.72

The item data and preliminary item grouping across the 3 factors were submitted to a content analysis. On the basis of this analysis, we concluded that two of the factors tapped multiple topics that are part of Florida's *Sunshine State Standards.* Factor 1 appears to measure students' understanding of waves; it also appears to measure students' understanding of the characteristics of stars. Factor 2 appears to measure student understanding of three related topics: the gravitational pull of planetary bodies; the relationship between Earth and its moon; and uses of technology for space exploration. Factor 3 covers one topic: the relationship between the Sun and Earth.

Further Review of the Items for Content Validity and Alignment with the Evidence Model

As part of the content analysis of the different factors, we decided to eliminate items with less direct ties to the *Sunshine State Standards*. After pruning several items that did not address core concepts, the remaining item pool was re-analyzed as a candidate pool for a final instrument. In this case, with a pool of 22 items (9 for factor 1, 9 for factor 2, 4 for factor 3) the overall score reliability is 0.87, nearly that of the 44-item test with half the number of original items.

We have not yet submitted these items for secondary review by scientists or by the professional development providers, due to time limitations. The professional development providers are focused currently on planning for their summer workshops, which begin June 5, 2006. We plan to complete review of these items and write any additional items needed by June 30, 2006.

References

Bransford, J. D., & Schwartz, D. (2000). Rethinking transfer: A simple proposal with interesting implications. *Review of Research in Education, 24*, 61-101.

Messick, S. (1992). The interplay of evidence and consequences in the validation of performance assessments. *Educational Researcher, 23*(2), 13-23.

Mislevy, R. J. (1994). Evidence and inference in educational assessment. *Psychometrika, 59*, 439-483.

Mislevy, R. J., Steinberg, L. S., & Almond, R. G. (1999). *Evidence-centered assessment design: A brief overview*. Princeton, NJ: Educational Testing Service.

Mislevy, R. J., Steinberg, L. S., Almond, R. G., Haertel, G. D., & Penuel, W. R. (2003). Improving educational assessment. In B. Means & G. D. Haertel (Eds.), *Evaluating educational technology: Effective research designs for improving learning*. (pp. 149-180). New York: Teachers College Press.

National Research Council. (2000). *Inquiry and the National Science Education Standards*. Washington, DC: National Academy Press.

Wiggins, G., & McTighe, J. (1998). *Understanding by design*. Alexandria, VA: ASCD.