The Transforming Instruction by Design in
Earth Science (TIDES) Project

Report on Technical Quality of 7th Grade Student Assessment

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

June 2006

Introduction and Overview

This report describes activities undertaken to create, pilot, and analyze technical quality of student assessments in seventh grade 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, the seventh grade test has adequate reliability for use in an experimental study and a coherent factor structure suggesting that two underlying constructs are being measured well. However, the test was more difficult than earlier tests developed for 6th and 8th grades, perhaps due to the age of students who completed the pretest and alignment of items to students' opportunity to learn. The difficulties are not problematic from a test quality perspective, however, since they will prevent ceiling effects and proved to have adequate reliability. As with our other tests, we plan to eliminate and adapt some items to better capture deep and enduring understandings of student learning.

Assessment Design Process

We discussed the assessment design process in an earlier report; we will not describe the design process in this report but refer the reader to the following report:

Penuel, W. R., Gallagher, L., & Kreikemeier, P. (2006). *Report on the technical quality of the 6th and 8th grade assessments.* Menlo Park, CA: SRI International.

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 students 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 level teachers who agreed to be part of the pilot received one of three sets of revised items to use to administer to their students before they had taught their Earth science unit, which for seventh grade was a test on students' understanding of the hydrosphere (water). 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 27 for the seventh grade test. 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.

Unlike with our previous pilots, we did not conduct posttests with ESBD pilot teachers. By the time we were able to revise items and solicit feedback from reviewers of items. Instead, we conducted a posttest locally with 6th grade students in California. We chose 6th rather than 7th grade for the pilot because the California standards require teachers to cover material tested on the 7th grade TIDES test. A total of 171 students took our revised 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.

Results: Analysis of the Seventh 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 high 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. Our other tests for 6th and 8th grades met this expectation, but the 7th grade test did not. 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 Seventh Grade Unit Test (Pilot)

Item	*N*	*P*
1	171	0.28
2	171	0.17
3	171	0.16
4	171	0.22
5	171	0.14
6	171	0.29
7	171	0.44
8	171	0.57
9a	170	0.21
9bc	170	0.14
10	170	0.23
11	170	0.33
12	170	0.30
13	170	0.27
14	170	0.44
15	170	0.60
16	170	0.07
17	170	0.44
18	170	0.40
19	170	0.55
20	170	0.32
21	170	0.15
22	170	0.35
23	169	0.30
24	28	0.18
25	169	0.14
26	170	0.29

Why were these items more difficult for students? One possible explanation was that we included our professional development providers more actively in writing and revising 7th grade items to better measure "deep understanding." There were more open-ended items on this test than on previous tests, and the items may have been more difficult for students to answer correctly. A second explanation has to do with the context for the posttest. We tested 6th graders, who may have found the test language more difficult. Further, the California standards give much more emphasis to a wide range of scientific facts and vocabulary than the UbD approach gives. Therefore, a test that tests for deep understanding would have been not as closely aligned to an approach that emphasizes teaching for deep understanding.

Despite the difficulty of the test, there was good overall reliability of the test as a whole. As a whole, the entire 27 item scale showed good internal reliability (Cronbach's alpha = 0.84) with this group of students.

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 either a 1, 2, or 3 factor solution is viable under the criterion of selecting eigenvalues greater than 1. A two-factor solution yielded a solution with the highest reliability for items clustered within factors, and since we hope to develop some subscales for the analysis, we selected a two-factor solution.

Figure 2. Scree plot of Eigenvalues from Factor Analysis

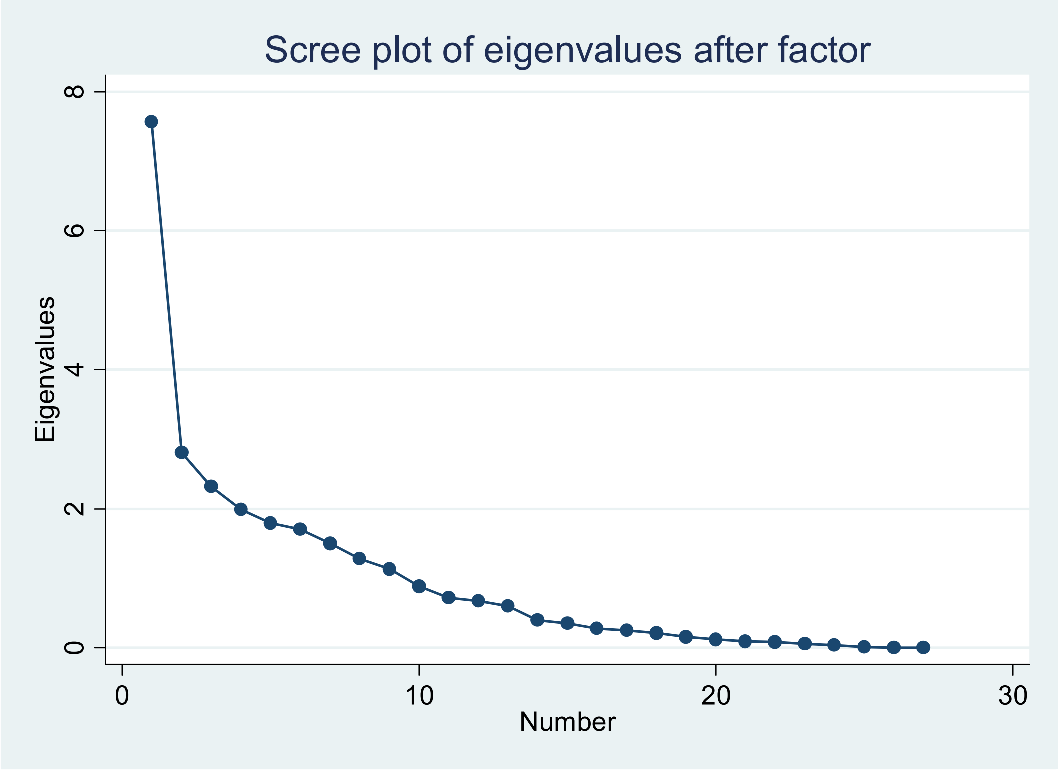

Table 2 on the following page shows the item loadings for each factor. The table shows only those loadings that are greater than 0.30. Exploratory factor analysis suggested a two-factor model was the best fit, although there were several items with high uniqueness values (that is, they did not load heavily on either factor). Increasing the number of factors did not help this, and in fact introduced negative loadings on some items, a sign that we are possibly trying to fit too many factors to the model.

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

Item	Factor 1	Factor 2	Uniqueness
1	0.67		0.54
2	0.53	0.50	0.47
3		0.54	0.64
4	0.79		0.38
5	0.75		0.36
6	0.50	0.59	0.40
7	0.40	0.42	0.66
8		0.69	0.52
9	0.37	0.38	0.72
9bc	0.40	0.42	0.67
10	0.57		0.64
11	0.32		0.88
12		0.65	0.58
13	0.41	0.45	0.62
14			0.92
15		-0.31	0.87
16	0.43		0.82
17			0.91
18			0.88
19		0.40	0.78
20	0.37	0.55	0.56
21	0.43	0.37	0.68
23	0.81		0.35
24	0.90		0.18
25	0.52	-0.39	0.58
26	0.70		0.51

We then analyzed the reliability of each factor, if items associated with that factor were considered part of a scale. All of the scales had moderate reliability. Factor 1 and Factor 2 both had the same reliability ($\alpha = 0.72$).

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. Although the alignment was good, we identified more than one construct tested by factor. The first factor appeared to test students' two constructs, namely their knowledge of *groundwater* and of *sand dune creation and erosion.* The second factor appeared to test three different constructs: *freshwater conservation, salt marsh ecosystem dynamics,* and *human* *impacts on the hydrosphere.* Given the better reliability for the overall scale and diversity of constructs measured by each factor, it may be advisable for our team to treat the 7th grade test as a single scale measuring the content of the seventh grade Earth science unit in Duval.

IRT Analysis of the Items

In an effort to extract the most discriminating items from the grade 7 pilot test, the data were re-analyzed under an IRT model. While a conventional classical reliability analysis would have given us sufficient information for the extraction of better-performing items (i.e., items that would combine into more reliable scales), the IRT method does lend itself neatly to graphical representations of the discrimination data.

There are some problems with using an IRT approach to these data that are important to note. First, these items violate a key tenet of a standard IRT model—that item responses are conditionally independent of one another, given student ability. The violation occurs here because many items share a common stem, prompt, or other resource. Responses to one item are inextricably linked to responses on other items, beyond what student ability would predict. One work-around to such conditional dependence is to bundle items into small "test-lets" (essentially creating a combined score for the entire bundle), which defeats the purpose of this analysis. A second misfit has to do with the possibility of guessing on multiple choice items. While there are IRT models that explicitly model guessing, one typically needs much more data for estimating these models, and they generally cannot be combined with the graded-response models.

To partly address these concerns, we partially collapsed tem scores, with the "0" (item completely blank) and "1" (item completely wrong) score categories combined. When kept separate, the IRT model failed to converge, suggesting that there was not enough information in the data to create separate step difficulty parameters for each score. Combining these two score levels led to a quickly converging model, and did not significantly alter the factor analysis or overall item difficulties.

A second way to address these concerns is by comparing the IRT analysis with the classical test analysis of the data. When the assumptions of IRT are met (namely, conditional independence) then the item-test correlation and IRT discrimination parameters should be well correlated. As in our dataset, that is not necessarily the case—item dependencies inflate the IRT-generated discrimination parameters for obviously bundled items. Therefore, we have chosen to report classical and IRT statistics side-by-side for the items in Table 3 on the following page. In addition to the discrimination parameter, we report on the item-test correlation and overall difficulty of the item.

Table 3. IRT and Classical Test Statistics: Seventh Grade Test

Factor	Item	Discrimination	Item-test correlation	Overall *p* (Difficulty)
1	1	0.29	0.22	0.28
	2	0.94	0.43	0.17
	4	0.96	0.52	0.22
	5	0.82	0.45	0.14
	8	0.36	0.21	0.57
	10	0.97	0.54	0.23
	11	0.96	0.50	0.33
	16	0.54	0.38	0.07
	21	1.55	0.57	0.15
	22	0.65	0.47	0.35
	23	0.74	0.49	0.30
	24	0.31	0.22	0.18
	25	0.61	0.51	0.14
2	3	0.94	0.42	0.16
	6	0.73	0.46	0.29
	7	0.77	0.43	0.44
	9a	0.61	0.40	0.21
	9bc	0.52	0.35	0.14
	12	0.96	0.51	0.30
	13	0.60	0.42	0.27
	19	1.05	0.41	0.55
	20	1.80	0.62	0.32
	26	0.80	0.45	0.29
unique	14	1.01	0.57	0.44
	15	0.50	0.31	0.60
	17	1.40	0.44	0.44

We do plan to use results from these analyses to remove items that appear to be both too difficult and also have poor ability to discriminate among students with different abilities. Item 1, for example, is a good candidate for elimination because of its poor ability to discriminate between students and low *p*-values.

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 have asked a panel of USGS scientists with expertise in different Earth systems—atmospheric, hydrologic, and geologic—to analyze items for their scientific importance. We are still analyzing their suggestions for improving the test.

Reference

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