Evaluating the Consistency of Test Content across Two Successive Administrations of a State-Mandated Science and Technology Assessment $^{1, 2}$

Timothy O’Neil$^3$ and Stephen G. Sireci

University of Massachusetts Amherst

Kristen L. Huff

The College Board

April 2002

$^1$ Center for Educational Assessment MCAS Validity Report No. 2. (CEA-454). School of Education, University of Massachusetts Amherst, Amherst, MA

$^2$ The research was completed as part of a contract between the Massachusetts Department of Education and the University of Massachusetts Amherst. However, the Department of Education is not responsible for any errors that remain in the paper nor should it be assumed that they agree with the findings and conclusions.

$^3$ Timothy O’Neil is now at Harcourt Educational Measurement. The authors thank Mark Gierl for sharing his insights on uncovering the cognitive skills measured by test items, and Joan Herman and three anonymous reviewers for their helpful comments on an earlier version of this paper.
Abstract

Educational tests used for accountability purposes must represent the content domains they purport to measure. When such tests are used to monitor progress over time, the consistency of the test content across years is important for ensuring that observed changes in test scores are due to student achievement rather than due to changes in what the test is measuring. In this study, expert science teachers evaluated the content and cognitive characteristics of the items from two consecutive annual administrations of a tenth grade Science and Technology assessment. The results indicated the content area representation was fairly consistent across years and the proportion of items measuring the different cognitive skill areas was also consistent. However, the experts identified important cognitive distinctions among the test items that were not captured in the test specifications. The implications of this research for the design of Science and Technology assessments and for appraising the content validity of state-mandated assessments are discussed.
Evaluating the Consistency of Test Content Across Two Successive Administrations of a State-Mandated Science and Technology Assessment

In the United States today, large-scale educational tests are consistently used to make important decisions regarding everything from individual student performance to more global decisions about entire schools, districts, or states. Several states currently mandate passing one or more tests to obtain a high school diploma. Furthermore, the recent reauthorization of the Elementary and Secondary Education Act (i.e. No Child Left Behind Act of 2001 (NCLB), 2002) mandates assessing students annually in reading and math in grades three through eight, and at least once in high school. Clearly, educational assessments are an important component in contemporary educational reform and accountability movements.

For large-scale educational assessments to be useful for accountability and instructional purposes, the content and cognitive domains represented by the tests must be clearly defined and the test items must adequately represent these domains (Popham et al. 2001). Furthermore, when such tests are used to monitor student progress, the consistency of test content across administrations is a critical validity issue. Evaluations of test content are often framed within the context of content validation, but some validity theorists argue against the use of that term.4 Setting aside issues related to nomenclature, Sireci (1998b) borrowed from Cronbach (1971), Lennon (1956), and others to define the

---

4 Some measurement specialists (e.g., Messick, 1989) argue against use of the term content validity because it is does not directly describe score-based inferences. Although this position has theoretical appeal, in practice, content validity is a widely endorsed notion of test quality (Sireci, 1998b). Thus, the position taken here is similar to Ebel (1977) who claimed “content validity is the only basic foundation for any kind of validity ... one should never apologize for having to exercise judgment in validating a test. Data never substitute for good judgment” (p. 59).
four critical components of content validity: domain definition, domain representation, domain relevance, and appropriateness of test development procedures. The study summarized in the current paper focuses on domain representation and relevance.

Most state-mandated K-12 educational assessments are targeted to statewide curriculum frameworks. NCLB requires states to (a) have statewide curriculum frameworks in place for math and reading by the spring of 2004 and for science by 2008. Students’ performance on the math and reading tests will be used to assign students to proficiency categories and monitor the “adequate yearly progress” of schools, districts, and states for reaching the goal of all students being “proficient” by 2014 (NCLB, 2002). Thus, NCLB reinforces the importance of ensuring that state-mandated educational assessments represent the domains they are designed to measure (Crocker, 2003).

In this paper, we evaluate the content (and cognitive) domain representation of a state-mandated tenth-grade Science and Technology assessment and investigate the consistency of this representation over two, successive, annual administrations. Such analyses are important for evaluating the degree to which annual statewide accountability tests are appropriate for accomplishing their objectives.

Importance of Content Validity

Although the Standards for Educational and Psychological Testing (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 1999) call for a variety of forms of validity evidence to support the use of a test for a specific purpose, validity studies often receive low priority in state-mandated assessments (Popham et al., 2001; Sireci, 1998b). This lack of validity research is unfortunate because statewide tests are typically constructed to
measure curriculum frameworks that were developed by the state to improve student learning and instruction. The need to adequately demonstrate content or curricular validity is arguably more important today than ever before. Evaluations of test content that are independent of the test developer are particularly informative for gauging the utility of a test for accomplishing its goals.

Evaluating the consistency of test content is one way to ensure that student performance differences over time are due to student or instructional characteristics, rather than due to changes in the test. Although test forms from different years are statistically equated to maintain a common scale across years, parallel content is prerequisite to score equating (Kolen & Brennan, 1995).

**Purpose of the Current Study**

This study was motivated by an earlier study where we evaluated the dimensionality of a state-mandated Science and Technology assessment (O’Neil & Sireci, 2001) across two successive annual administrations. In the earlier study, we evaluated the dimensionality of the item score data from two independent groups of about 60,000 students who took the test in each year. The results indicated the data were multidimensional, but we were unable to interpret the dimensions, even after considering item characteristics such as content area measured, cognitive complexity, item difficulty, and item format. Furthermore, there was little overlap across the tests with respect to common items, and so we were unable to evaluate the equivalence of the dimensionality. Given that we are psychometricians and not science experts, we decided to convene a group of expert science teachers in the state to help us interpret similarities and differences among the test items. This idea involved into a larger effort to evaluate the
content quality of the test and the consistency of the test content across years. The primary purposes of the current study were to (a) evaluate the content domain representation of the tests, (b) better understand the cognitive skills measured by the tests, and (c) evaluate the consistency of the tests with respect to the content and cognitive domains measured.

Method

Description of Test and Data

The test we examined is a criterion-referenced Science and Technology test from a state-mandated, comprehensive assessment system used to evaluate students, schools, and districts. The 1999 and 2000 versions of the test were examined. The grade 10 Science and Technology test was chosen for this study because previous research has shown that both cognitive and content dimensions are important for measuring students’ science proficiency (Sireci, Robin, Meara, Rogers, & Swaminathan, 2000; Sireci, Rogers, Swaminathan, Meara, & Robin, 2000) and because content changes were being considered for this test, which would make the results of this study informative for improving future Science and Technology assessments within the state.

The 1999 and 2000 Grade 10 Science and Technology tests consisted of 36 dichotomously scored multiple-choice (MC) items and 6 polytomously scored constructed-response (CR) items. Test data were available for approximately 60,000 students in each year. The coefficient alpha reliability estimates for the total scores on the two tests were similar (.90 for 1999 and .89 for 2000).

The technical manual listed five content areas under which the test items were categorized: Earth and Space Sciences, Life Sciences, Physical Sciences, Technology,
and Inquiry (Massachusetts Department of Education, 2000). The numbers of items in each content area for the 1999 and 2000 tests are presented in Table 1, stratified by item format. The percentage of test score associated with each content area is also reported.

With respect to item format, all content areas were measured with one or two CR items, with the exception of Inquiry in 1999, which was measured using only three MC items.

With respect to content area representation across years, for the Physical Sciences and Earth and Space Sciences content areas, the proportion of the total test score associated with the content area was generally consistent across years (less than 2% difference).

Moderate differences across years (about 5%) occur for the Life Sciences and Inquiry content areas. The Technology content area had the largest change across years. This content area represented about 27% of the total test score in 1999, but only about 17% of the total test score in 2000.

Table 1
Content Specifications for 1999 and 2000 Science and Technology Tests

<table>
<thead>
<tr>
<th>Content Area</th>
<th>Item Format</th>
<th>Number of Items</th>
<th>1999 % of Total Score</th>
<th>Number of Items</th>
<th>2000 % of Total Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>Inquiry</td>
<td>MC</td>
<td>3</td>
<td>5.0%</td>
<td>2</td>
<td>10.00%</td>
</tr>
<tr>
<td></td>
<td>CR</td>
<td>0</td>
<td></td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>Physical Sciences</td>
<td>MC</td>
<td>8</td>
<td>26.67%</td>
<td>13</td>
<td>28.33%</td>
</tr>
<tr>
<td></td>
<td>CR</td>
<td>2</td>
<td>26.67%</td>
<td>1</td>
<td>26.67%</td>
</tr>
<tr>
<td>Life Sciences</td>
<td>MC</td>
<td>9</td>
<td>21.67%</td>
<td>8</td>
<td>26.67%</td>
</tr>
<tr>
<td></td>
<td>CR</td>
<td>1</td>
<td></td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>Earth &amp; Space Sciences</td>
<td>MC</td>
<td>8</td>
<td>20.00%</td>
<td>7</td>
<td>18.33%</td>
</tr>
<tr>
<td></td>
<td>CR</td>
<td>1</td>
<td>20.00%</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>Technology</td>
<td>MC</td>
<td>8</td>
<td>26.67%</td>
<td>6</td>
<td>16.67%</td>
</tr>
<tr>
<td></td>
<td>CR</td>
<td>2</td>
<td>26.67%</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>42</td>
<td>100.0%</td>
<td>42</td>
<td>100.0%</td>
<td></td>
</tr>
</tbody>
</table>

Note: MC items contribute 1 raw score point; CR items contribute 4 raw score points to overall test score.
The developmental literature supporting the tests described three cognitive areas that were measured by the items: Thinking, Procedural, and Application Skills (Massachusetts Department of Education, 1998). The literature supporting these tests implied that all items were targeted to one of these cognitive skill areas; however, no data were available from the state department of education regarding the cognitive skill classifications of the items. These classifications were not retained after the test development stage, since the student results were not reported for the cognitive skill areas. For this reason, and given that one of the content areas was labeled “inquiry,” a central focus of our analysis was getting an understanding of the cognitive skills measured by the items.

**Participants**

To evaluate what was being measured on each test, ten subject matter experts (SMEs) were recruited from the same state where the tests were administered. These experts were recruited from a list of twelve science teachers in the state who were certified by the National Board of Professional Teacher Standards (NBPTS), and from a list of science teachers who were recommended by science curriculum leaders within the state. Six of the ten SMEs were certified by the NBPTS. All participants had more than four years experience teaching science (median teaching experience was 12 years). Six of the SMEs were female and four were male.

**Procedure**

**SME tasks**

After agreeing to participate in the study, the SMEs were mailed the two test booklets and directions for providing traditional content validity ratings (Sireci, 1998a).
Specifically, they were asked to review the descriptions of the content and cognitive areas measured on the test and assign each test item to one content area and one cognitive area. The descriptions of the content and cognitive areas were taken verbatim from publications describing what the test measures that were published by the state’s department of education. Eight of the ten SMEs completed these “item congruence” ratings.

The SMEs were then convened for a one-day meeting to further evaluate the items. Only nine SMEs were able to participate in the meeting (seven of those who completed the earlier ratings and two new SMEs\(^5\)). At this meeting, the SMEs (a) provided ratings of item similarity, (b) discussed the best framework for classifying the items according the cognitive skills measured, (c) classified a subset of items into this framework using a group consensus procedure, and (d) completed a survey about their overall impressions of the tests and their opinions about the meeting. Brief descriptions of these tasks follow.

**Item similarity ratings**

Sireci and Geisinger (1992, 1995) argued that traditional ratings of content validity do not provide a rigorous appraisal of how well test items represent their content and cognitive domains because they inform SMEs of the knowledge and skill areas presumably measured by the items. They recommended asking SMEs to rate the *similarity* of test items to one another *with respect to the knowledge and skills being measured by the items*. This procedure does not require informing the SMEs of the

\(^5\) The numbers of SMEs performing item congruence and item similarity ratings was not consistent because not all SMEs performed both tasks. Seven SMEs performed both the item congruence and similarity rating tasks, one SME performed only the item congruence rating task, and two SMEs performed only the item similarity rating task.
content or cognitive areas defined in the test specifications. The logic underlying this item similarity rating task is that items designed to measure the same knowledge or skills will be rated more similar to one another than they are to items intended to measure different knowledge and skill areas (e.g., two Life Sciences items would be expected to be rated more similar to one another than a Life Sciences and a Physical Sciences item would be rated to one another). Sireci and Geisinger recommended analysis of the item similarity data using multidimensional scaling (MDS) so that any characteristics used by the SMEs for judging item similarity would emerge in the analysis. Using this procedure, the content validity of a test is supported when the items group together in the MDS space in a manner congruent with their intended content and cognitive designations.

A benefit of the item similarity rating task is that SMEs are not required to rate the items according to any prespecified criteria, which avoids response biases such as social desirability (i.e., rating the items in a way they think the test developer would want them to respond). For seven of the nine SMEs who participated in this study, this removal of bias was not possible, since they previously completed the item congruence ratings and were familiar with the content and cognitive areas to which the items were written. However, we concluded it would still be worthwhile to gather and analyze these data to discover which dimensions would be most important for characterizing their perceptions, and if other dimensions would be necessary to understand the similarities among the items.6

---

6 In addition, we performed a preliminary analysis to see if the two SMEs who did not complete the item congruence ratings in advance rated the item similarities differently than the other SMEs. As discussed in the results section, no such differences were observed.
In this study, the item similarity ratings were gathered by presenting the SMEs with a pair of test items accompanied by a rating scale ranging from 1 (extremely similar) to 10 (not at all similar). Only the endpoints of the rating scale were defined. To reduce the amount of time required to gather the item similarity data, 15 items were selected from both the 1999 and 2000 tests. These two subsets were carefully selected so that the proportions of items in each content area were equivalent to the full-length tests. The proportions of MC and CR items were also equivalent to the full-length test, as were the proportions of items that involved accompanying graphics.

To train the SMEs to complete the item rating task, they were asked to respond to the 30-item subset (to familiarize them with the items) and then they individually rated the similarity of a subset of four items. They discussed these individual ratings as a group. This step was helpful for ensuring the SMEs understood the task and for allowing the SMEs to hear others’ perceptions. After this step, it was determined that all SMEs were on task. They then proceeded to rate the remaining pairings of these 30 items, which amounted to a total of 435 similarity ratings. To gauge the amount of error present in the similarity data, eight item pairs were repeated.

**Discussion and selection of cognitive skill taxonomy**

After completing their item similarity ratings (and after a brief lunch), the SMEs were presented with four taxonomies for classifying the cognitive skills measured by Science and Technology test items: (a) Bloom’s (1956) taxonomy, (b) the cognitive areas described in the state’s test documentation, (c) the 2000 NAEP Science “ways of knowing and doing science” (National Assessment Governing Board, 2001), and (d) descriptors provided by the National Science Teachers Association (2002). They were
also asked whether they were aware of any other classification systems that might be appropriate for classifying the cognitive skills being assessed by Science and Technology tests. After a brief discussion, they unanimously decided that Bloom’s taxonomy would be the best framework for classifying the items measured on the test. This hierarchical taxonomy specifies six cognitive skill areas: knowledge, comprehension, application, analysis, synthesis, and evaluation. The SMEs were then asked to assign each item from the 2000 test into one of Bloom’s six levels (time did not permit for rating the 1999 items). They first rated the items individually, then, they discussed their ratings as a group. The number of SMEs assigning an item to each cognitive level was recorded (see Gierl, 1997, for a similar use of Bloom’s taxonomy to help appraise the cognitive skills measured by test items).

Survey completion

Before the meeting adjourned, the SMEs completed a comprehensive questionnaire that asked them about the criteria they used to make their similarity ratings, the confidence they had that their similarity ratings “accurately reflect the content and cognitive similarities among the item pairs,” and several other questions regarding the quality of the state’s Science and Technology assessments and how they might be improved in the future.

Analysis of SME data

The SMEs’ ratings of the test items vis-à-vis the content areas specified by the test developers (i.e., traditional content validity ratings) were analyzed to determine the congruence of the items to each of the content areas specified in the test blueprints. The percentage of SMEs that classified the item in the same category as the test developers
was computed. Although there are no absolute rules for determining when such item congruence ratings support the conclusion of content validity, Popham (1992) and Sireci (1998a) suggested that an item should be considered matched to its content area if seven out of ten SMEs place it in the same area as the test developers. This “70%” criterion was computed for all items, as was the number of items that were unanimously matched to the test blueprint by all SMEs.

Multidimensional scaling analyses

The SME item similarity data were analyzed using MDS to discover the content, cognitive, and other qualitative dimensions that the SMEs perceived in making their similarity ratings. First, an “individual differences” (weighted) MDS model (Carroll & Chang, 1970) was used to determine if there were differences among the SMEs in their similarity ratings. Specifically, we investigated whether one or more SMEs were using dimensions that were different from the others. No such differences were found and so the results are not reported here. The SMEs item similarity ratings were subsequently averaged and analyzed using classical MDS (Davison & Sireci, 2000; Kruskal & Wish, 1978).

All MDS analyses were implemented using the ALSCAL program in SPSS, version 10.0 (Young & Harris, 1993). The nonmetric option was used, to maximize the fit of the data to the MDS model. One- through six-dimensional models were fit to the

---

7 Given that only 8 SMEs actually completed the item classifications, 6 of 8 SMEs (i.e., 75%) was used as the criterion for concluding that an item was matched to its designated content area.
data (six is the maximum number of dimensions that can be fit in ALSCAL), and the STRESS and R² fit indices were used to evaluate the fit of each MDS model.⁸

There are no absolute guidelines for determining adequate fit in MDS, but simulation research reported by Kruskal and Wish (1978) indicate that if a one-dimensional model has a STRESS value of .15 or less, the data can be considered unidimensional. However, the amount of error present in the data can obscure dimensionality and so the typical procedure is to fit several dimensional solutions to a data set and evaluate when improvement in fit begins to taper off (since adding dimensions to the model will always improve fit due to capitalization on chance). The second criterion for determining dimensionality was interpretability of the solution.

The content, individual cognitive ratings, and group consensus cognitive ratings were all related to the MDS solutions via correlational analysis (Schiffman, Reynolds, & Young, 1981; Sireci, 1998a), to help interpret the dimensions that were obtained in the analysis of the SME item similarity data.

⁸ STRESS represents the square root of the normalized residual variance of the monotonic regression of the MDS distances on the transformed item dissimilarity data. Thus, lower values of STRESS indicate better fit. The R² index reflects proportion of variance of the transformed dissimilarity data accounted for by the MDS distances. Thus, higher values of R² indicate better fit.
Results

Item Congruence Ratings

Before the meeting where the item similarity ratings were gathered, eight of the ten SMEs evaluated the items and assigned each item to one of the content areas listed in the test blueprint, as well as to one of the cognitive areas described in the test documentation. The degree to which the SMEs classified the content designations of the items in a manner consistent with the test blueprint was ascertained by computing the percentage of SMEs who classified the item into the same content category as the test developer. These findings are summarized in Tables 2 and 3. Three criteria are reported in each table: (a) the number of test items unanimously matched to its test blueprint content area, (b) the number of test items matched to its blueprint content area by at least six of the eight SMEs (i.e., 75% of the SMEs), and (c) the number of items that not even one SME matched to its blueprint content area. This last criterion was added to highlight items that the test developers should review for possible reclassification.

Table 2 presents an overall summary for both the 1999 and 2000 tests stratified by item format. In each year, the SMEs unanimously classified about half the MC items in the same content area specified in the test blueprint. In 2000, half of the CR items were also unanimously matched, but only one-third of the CR items were unanimously matched in 1999. Using the six-of-eight criterion, 33 of the 42 items (78.8%) from the 2000 test were matched to the content areas specified in the test blueprint by at least six SMEs, while 30 of the 42 items (71%) from the 1999 test met this criterion. Each test also had one item that none of the SMEs matched to its designated content area. The 2000 test had better congruence ratings for the CR items, relative to 1999, but the small
number of these items (n=6) limits the meaningfulness of this finding. Therefore, in subsequent tables, we do not break out the results by item type.

Table 2
Summary of Content Area Congruence Ratings By Item Format

<table>
<thead>
<tr>
<th>Year</th>
<th>Item Format</th>
<th># Items</th>
<th>% Items Unanimously Matched</th>
<th>% Items Matched by 6 of 8 SMEs</th>
<th>% Items Not Matched by Any SMES</th>
</tr>
</thead>
<tbody>
<tr>
<td>1999</td>
<td>MC</td>
<td>36</td>
<td>50.0%</td>
<td>75.0%</td>
<td>2.8%</td>
</tr>
<tr>
<td></td>
<td>OR</td>
<td>6</td>
<td>33.3%</td>
<td>50.0%</td>
<td>0.0%</td>
</tr>
<tr>
<td>2000</td>
<td>MC</td>
<td>36</td>
<td>55.6%</td>
<td>77.8%</td>
<td>2.8%</td>
</tr>
<tr>
<td></td>
<td>OR</td>
<td>6</td>
<td>50.0%</td>
<td>83.3%</td>
<td>0.0%</td>
</tr>
</tbody>
</table>

Table 3 presents the congruence data stratified by content area for 1999 and 2000, respectively. Using the six-of-eight SME criterion, the rank ordering of the content areas with respect to SME/blueprint congruence is the same. The Life Sciences items were generally matched into their blueprint categories by the SMEs (all items in 1999 and 9 of 10 items in 2000), followed by the Physical Sciences items (8 of 10 in 1999 and 13 of 14 in 2000), and the Earth and Space Sciences items (7 of 9 in 1999 and 6 of 8 in 2000). For the most part, the content classifications for the items in these three areas were supported by the SME ratings (however, one of the 2000 Earth & Space Sciences items was classified as a Physical Sciences item by all 8 SMEs). For the Technology content area, five of the seven items from 2000 were designated as Technology items by at least six SMEs, but only five of the ten 1999 items met this criterion. Items from the Inquiry content area had the lowest congruence ratings. None of these items met the six-of-eight SME criterion. These results indicate that the salience of the test content varied by content area. Furthermore, the differences in content congruence between 1999 and 2000 for the technology area may be a cause of concern.
Table 3  
Summary of Congruence Ratings By Content Area

<table>
<thead>
<tr>
<th>Content Area</th>
<th># Items</th>
<th>% Items Matched by 6 of 8 SMEs</th>
<th>% Items Unanimously Matched</th>
<th>% Items Not Matched by Any SMEs</th>
</tr>
</thead>
<tbody>
<tr>
<td>Life Sciences</td>
<td>10</td>
<td>10</td>
<td>100.0%</td>
<td>90.0%</td>
</tr>
<tr>
<td>Physical Sciences</td>
<td>10</td>
<td>14</td>
<td>80.0%</td>
<td>92.9%</td>
</tr>
<tr>
<td>Earth &amp; Space Sciences</td>
<td>9</td>
<td>8</td>
<td>77.8%</td>
<td>75.0%</td>
</tr>
<tr>
<td>Technology</td>
<td>10</td>
<td>7</td>
<td>50.0%</td>
<td>71.4%</td>
</tr>
<tr>
<td>Inquiry</td>
<td>3</td>
<td>3</td>
<td>0.0%</td>
<td>0.0%</td>
</tr>
<tr>
<td>Total</td>
<td>42</td>
<td>42</td>
<td>71.4%</td>
<td>78.6%</td>
</tr>
</tbody>
</table>

An analysis of variance (ANOVA) was conducted on the item congruence data to determine whether statistically significant differences in content representation were present across years and content area. For each item, the dependent variable was the number of SMEs classifying it into its prescribed content area. These results are summarized in Table 4. The main effect due to content area was statistically significant ($F_{(4,74)}=10.54$, $p < .001$), but the year main effect and year-by content interaction were not statistically significant. These results suggest that the content representation of the test was consistent across years, but that there were differences in content representation across content areas. Post hoc analyses indicated that the item congruence ratings for the Inquiry content area were significantly lower than those of the other content areas, and the ratings for the Technology content area were significantly lower than those for the Life and Physical Sciences content areas. Although these differences across content area were statistically significant, it should be noted that the effect size associated with content was small, accounting for only 4% of the variance in the item congruence data (i.e., $\eta^2=.042$)
Table 4  
Item Congruence ANOVA Summary Table

<table>
<thead>
<tr>
<th>Source</th>
<th>SS</th>
<th>df</th>
<th>MS</th>
<th>F</th>
<th>p</th>
<th>$\eta^2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Year</td>
<td>2.43</td>
<td>1</td>
<td>2.43</td>
<td>0.64</td>
<td>.426</td>
<td>.001</td>
</tr>
<tr>
<td>Content</td>
<td>159.66</td>
<td>4</td>
<td>39.91</td>
<td>10.54</td>
<td>.000</td>
<td>.042</td>
</tr>
<tr>
<td>Year*Content</td>
<td>10.71</td>
<td>4</td>
<td>2.68</td>
<td>0.71</td>
<td>.590</td>
<td>.003</td>
</tr>
<tr>
<td>Error</td>
<td>280.30</td>
<td>74</td>
<td>3.79</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>3841.00</td>
<td>84</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Item Cognitive Classifications

Although we did not have the test developer’s classifications of the cognitive areas to which each item belonged, we were interested in the SMEs’ perceptions of the how well the cognitive areas were measured, as well as the consistency of this representation across years.\(^9\) Figure 1 illustrates the number of items within each cognitive skill area that were matched to that area by at least 6 of the 8 SMEs. The results were generally consistent across years. Most items were linked to the Thinking skill area (19 items in 1999 and 17 in 2000). Using the 6 of 8 criterion, only 7 items in 1999 and 6 items in 2000 would be classified as measuring higher-order skills.

\(^9\) Actually, we thank Joan Herman for encouraging us to further explore the cognitive classification data in response to an earlier version of this article.
We also recorded the SMEs’ specific cognitive skills classification for each item. These results are summarized in Table 5. The cell entries indicate the number of times any SME classified an item into a cognitive skill area. There were only 3 missing classifications across the 672 ratings (42 items X 2 years X 8 SMEs). These results indicate slightly higher emphasis on the Procedural skill area in 2000 and slightly lower emphasis on the Thinking skill area. However, a chi-square test on these data was not statistically significant ($\chi^2(2)=3.94, p=.24$), which suggests that the distribution of items with respect to cognitive skill areas was consistent across years.
Table 5
Summary of Cognitive Classifications for Items

<table>
<thead>
<tr>
<th>Cognitive Area</th>
<th>1999</th>
<th>2000</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Thinking</td>
<td>205</td>
<td>187</td>
<td>392</td>
</tr>
<tr>
<td>Procedural</td>
<td>43</td>
<td>61</td>
<td>104</td>
</tr>
<tr>
<td>Application</td>
<td>86</td>
<td>87</td>
<td>173</td>
</tr>
<tr>
<td>Total</td>
<td>334</td>
<td>335</td>
<td>669</td>
</tr>
</tbody>
</table>

Note: Cell entries represent the number of times an SME classified an item into a cognitive area. $\chi^2(2) = 3.94$ (n.s.).

MDS Results

As mentioned earlier, preliminary analysis of the data indicated a high degree of congruence among the SMEs regarding the way in which they rated the item similarities. We initially fit an individual differences MDS model to the nine separate SME item similarity matrices to determine if there were differences among the SMEs with respect to the dimensions that best accounted for their item similarity ratings. We compared the male and female SMEs as well as the two SMEs who did not complete the item congruence ratings in advance of the item similarity ratings with the other seven SMEs. Across all SMEs, the personal dimension weights were very similar, which indicated the more complex weighted MDS model was unnecessary. Therefore, we averaged the similarity ratings across the nine SMEs and conducted a classical (unweighted) MDS analysis.

We obtained a simple estimate of the reliability of the individual SMEs’ similarity ratings by including 8 replicated item pairs toward the end of the similarity rating booklet for each SME. Across these 72 replications (9 SMEs times 8 replicated pairs) the Spearman rank correlation between the first and second rating was .86 (p < .0001). The average difference across ratings was 0.76 on the ten-point rating scale, which was not statistically significant ($t(71)=1.27$, p=.21). Separate dependent t-tests were also
conducted for the ratings of each SME and none were statistically significant at \( p \leq .05 \). These findings suggest that there was no systematic tendency for the raters to provide higher or lower similarity ratings as they completed their task, and that their ratings displayed adequate reliability. In addition, given that their similarity ratings were averaged and analyzed using classical MDS, the reliability of the averaged rating data is probably high.

The fit statistics for the one- through six-dimensional MDS solutions fit to the SME item similarity data are presented in Table 6. The one-dimensional solution did not show good overall fit (STRESS = .31), but the fit improved substantially for the two-dimensional solution, and then began to taper off. These results suggest that the two- and three-dimensional solutions exhibit reasonable data-model fit.

Table 6
STRESS and \( R^2 \) Results from MDS Analysis SME Similarity Ratings

<table>
<thead>
<tr>
<th>Dimensional Solution</th>
<th>Stress</th>
<th>R-Squared</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>.31</td>
<td>.74</td>
</tr>
<tr>
<td>2</td>
<td>.17</td>
<td>.87</td>
</tr>
<tr>
<td>3</td>
<td>.12</td>
<td>.92</td>
</tr>
<tr>
<td>4</td>
<td>.09</td>
<td>.94</td>
</tr>
<tr>
<td>5</td>
<td>.07</td>
<td>.95</td>
</tr>
<tr>
<td>6</td>
<td>.06</td>
<td>.96</td>
</tr>
</tbody>
</table>

To interpret the MDS solutions of the SME data, all external data that described characteristics about the items were correlated with the MDS item coordinates. The logic motivating this analysis was that characteristics of the items that correlated with their MDS coordinates could help explain the item features perceived by the SMEs and would facilitate interpretation of the MDS solutions. Only correlations that were statistically significant at \( p < .01 \) were considered meaningful (note that, \( n=42 \), the number of items, for all correlation analyses). The external data available on the items included: (a) the \( a \)
(discrimination), and $b$ (difficulty) item response theory (IRT) parameter estimates, (b) classical (proportion correct) item difficulty statistics, (c) the numbers of SMEs classifying each item into the test blueprint content areas, (c) the numbers of SMEs classifying each item into each level of Blooms’ taxonomy, and (d) the numbers of SMEs classifying each item into each level of the state’s cognitive areas.

In addition to these variables, a cognitive variable, called the Bloom’s index was calculated by assigning ranks of 1 through 6 to each level of the taxonomy (with “knowledge” given a 1 and “evaluation” given a 6), and then multiplying the number of SMEs who classified each item into a level by the rank assigned to the level. This index provided a more informative (i.e., non-dichotomous) index of the cognitive level measured by an item ranging from 9 (i.e., all SMEs classified the item in the knowledge level) to 54 (i.e., all SMEs classified the item as evaluation). Lastly, dichotomized variables based on item format (MC versus CR), year of administration, and items that contained a graphical element were also created.

Correlations among the SMEs’ content, and cognitive classifications, and the MDS coordinates, are presented in Table 7. The correlations among the MDS coordinates, IRT parameters, and dummy coded variables for item format, use of graphs, and year of administration, are reported in Table 8. In all three MDS solutions, the SME ratings appear to be related to the cognitive distinctions among the items. The Bloom’s index was highly correlated with the first dimension from each solution, which suggests that the SMEs were primarily considering the cognitive skills measured by the items when making their similarity judgments (Table 7). Also, item format (i.e., multiple-choice versus constructed-response) was moderately correlated with dimensions from
each solution and the IRT discrimination parameter was correlated with a dimension from the three- and four-dimensional solutions (Table 7). Year of administration had near-zero correlations with all dimensions in all solutions. The three-dimensional solution exhibited moderate correlations with all five content areas and captured all patterns that were represented in the four dimensional solution. Since it also displayed adequate data-model fit, it was taken as the most appropriate representation of the SMEs’ similarity ratings.

Table 7
Correlations of SME Content and Cognitive Classifications, and Item Format with MDS Coordinates

<table>
<thead>
<tr>
<th>Dimensiona</th>
<th>Bloom Index</th>
<th>Inquiry</th>
<th>Physical Sciences</th>
<th>Life Sciences</th>
<th>Earth &amp; Space Sciences</th>
<th>Technology</th>
<th>Think</th>
<th>Proc.</th>
<th>Apply</th>
</tr>
</thead>
<tbody>
<tr>
<td>2_1</td>
<td>0.85*</td>
<td>0.64*</td>
<td>0.05</td>
<td>-0.14</td>
<td>-0.13</td>
<td>-0.11</td>
<td>-0.77*</td>
<td>0.19</td>
<td>0.71*</td>
</tr>
<tr>
<td>2_2</td>
<td>-0.11</td>
<td>0.01</td>
<td>-0.58*</td>
<td>0.19</td>
<td>-0.05</td>
<td>0.48*</td>
<td>-0.01</td>
<td>0.47</td>
<td>-0.32</td>
</tr>
<tr>
<td>3_1</td>
<td>0.87*</td>
<td>0.65*</td>
<td>0.06</td>
<td>-0.17</td>
<td>-0.14</td>
<td>-0.07</td>
<td>-0.79*</td>
<td>0.21</td>
<td>0.73*</td>
</tr>
<tr>
<td>3_2</td>
<td>0.16</td>
<td>0.03</td>
<td>0.56*</td>
<td>-0.21</td>
<td>0.06</td>
<td>-0.45</td>
<td>-0.06</td>
<td>-0.46*</td>
<td>0.39</td>
</tr>
<tr>
<td>3_3</td>
<td>-0.15</td>
<td>-0.09</td>
<td>0.41</td>
<td>-0.48*</td>
<td>-0.40</td>
<td>0.56*</td>
<td>-0.30</td>
<td>0.52*</td>
<td>-0.02</td>
</tr>
<tr>
<td>4_1</td>
<td>-0.88*</td>
<td>-0.65*</td>
<td>-0.07</td>
<td>0.16</td>
<td>0.15</td>
<td>0.08</td>
<td>0.79*</td>
<td>-0.21</td>
<td>-0.73*</td>
</tr>
<tr>
<td>4_2</td>
<td>0.18</td>
<td>0.05</td>
<td>0.54*</td>
<td>-0.22</td>
<td>0.08</td>
<td>-0.44</td>
<td>-0.07</td>
<td>-0.47*</td>
<td>0.42</td>
</tr>
<tr>
<td>4_3</td>
<td>0.12</td>
<td>0.10</td>
<td>-0.38</td>
<td>0.46</td>
<td>0.41</td>
<td>-0.58*</td>
<td>0.31</td>
<td>-0.54*</td>
<td>0.02</td>
</tr>
<tr>
<td>4_4</td>
<td>-0.30</td>
<td>0.31</td>
<td>0.09</td>
<td>-0.54*</td>
<td>0.28</td>
<td>0.16</td>
<td>-0.36</td>
<td>0.16</td>
<td>0.29</td>
</tr>
</tbody>
</table>

a2_1 denotes the first dimension from the two-dimensional solution.
* denotes significance at p < 0.01.
<table>
<thead>
<tr>
<th>Dimension a</th>
<th>Format</th>
<th>Year of Admin</th>
<th>a parameter</th>
<th>b parameter</th>
<th>c parameter</th>
</tr>
</thead>
<tbody>
<tr>
<td>2_1</td>
<td>.59*</td>
<td>.029</td>
<td>.36</td>
<td>.10</td>
<td>-.04</td>
</tr>
<tr>
<td>2_2</td>
<td>.70*</td>
<td>-.033</td>
<td>.11</td>
<td>.05</td>
<td>.41</td>
</tr>
<tr>
<td>3_1</td>
<td>.58*</td>
<td>.022</td>
<td>.32</td>
<td>.10</td>
<td>-.06</td>
</tr>
<tr>
<td>3_2</td>
<td>-.70*</td>
<td>.036</td>
<td>-.09</td>
<td>-.08</td>
<td>-.37</td>
</tr>
<tr>
<td>3_3</td>
<td>.19</td>
<td>.094</td>
<td>-.48*</td>
<td>.30</td>
<td>-.14</td>
</tr>
<tr>
<td>4_1</td>
<td>.57*</td>
<td>-.017</td>
<td>-.31</td>
<td>-.10</td>
<td>.07</td>
</tr>
<tr>
<td>4_2</td>
<td>-.72*</td>
<td>.026</td>
<td>-.10</td>
<td>-.09</td>
<td>-.34</td>
</tr>
<tr>
<td>4_3</td>
<td>-.19</td>
<td>-.109</td>
<td>.50*</td>
<td>-.30</td>
<td>.14</td>
</tr>
<tr>
<td>4_4</td>
<td>-.07</td>
<td>-.041</td>
<td>.06</td>
<td>-.17</td>
<td>.12</td>
</tr>
</tbody>
</table>

a2_1 denotes the first dimension from the two-dimensional solution. * denotes significance at p < 01.

Visual inspection of the three-dimensional solution was also used to help interpret the dimensions. Based on the patterns of correlations noted earlier and the visual interpretations, we labeled the first dimension “cognitive complexity,” the second dimension “Life/Earth Sciences versus Physical Sciences/Technology,” and the third dimension “Item Format.”

The two-dimensional scatter plots from the three-dimensional MDS solution are presented in Figures 2 to 4. Figure 2 illustrates the Bloom’s Taxonomy classifications for the items as determined by the SMEs. All but one of the items that was considered to be measuring lower-level skills (i.e., Knowledge and Comprehension items) appear on the lower end of Dimension 1 (horizontal), while the Application and Analysis items appear at the upper end of the dimension. Figure 3 highlights the second (vertical) dimension that distinguishes among the item content designations. Although there is some overlap noted among the content areas, the Physical Sciences and Technology items tend to group
toward the top, where the Life and Earth sciences items group toward the bottom. Figure 4 depicts the item format and graphical aspects of test questions. Here the upper circle is drawn around items with some graphical component and the lower circle is drawn around CR items.

In general, the similarity data suggest that the primary characteristic the SMEs used to rate the item similarities was cognitive complexity of the item. Secondary characteristics included item content, item format, and graphical components. These findings are similar to those found for the 1996 Grade 8 NAEP Science exam (Sireci, Robin et al., 2000).
It is interesting to note that the MDS results from the similarity data were not related to year of administration. That is, the items from 1999 and 2000 did not load on different MDS dimensions. This finding supports the conclusion that the knowledge and skills measured on each test were consistent across years.

**Subject Matter Experts’ Survey Data**

Upon completion of all rating tasks the SMEs were asked to complete a short questionnaire. Of particular interest to interpretation of the MDS results were two questions, one about criteria used to complete the similarity ratings and a rating of how confident they were that their ratings reflected the content and cognitive similarities of the item pairs. As to criteria used, all nine SMEs said they evaluated the items based on cognitive level measured by the items. Eight used science discipline measured by the item, 7 used item difficulty, 5 used item format, and 5 used item context. With respect to their level of confidence in their similarity ratings, the average rating was 4 (slightly confident) on a six-point scale where 6 was “very confident.”
Figure 3. 3D CMDS Solution from SME Similarity Ratings

Dim 3: Item Format

(I = Inquiry, P = Physical, L = Life, E = Earth, T = Technology)
Discussion

In this paper, we recruited highly respected science experts to gauge the degree to which the content specifications of the test held up to external scrutiny, and the degree to which the structure of the test was consistent across years. In general, the results indicate that three of the five content areas were well represented on the test, a fourth was inconsistent with respect to content congruence across years, and the fifth was totally unsupported. However, the item similarity data did not uncover a “year effect,” which suggests that the content structure of the test is fairly comparable across years.

Equally important to the results that pertain to content area (or perhaps more important) were the results that spoke to the cognitive skills measured by the items. The results suggest that although the cognitive skills were consistently measured across years, predominantly lower-level skills were measured. It is also interesting to note that the
MDS results suggest that the cognitive skills measured by the items better characterize what the items are measuring than the content areas. This finding has important implications because content area subscores were reported for this test, but subscores were not reported for cognitive skill areas. Given that many of the curriculum frameworks associated with NCLB stipulate cognitive dimensions and areas, the degree to which such domains are adequately represented by state-mandated tests directly affects the educational utility of the assessments.

With respect to the consistency of content quality of the test across years, the results were favorable for only three of the five content areas. For Physical Sciences, Life Sciences, and Earth and Space Sciences content areas, the content characteristics of the items were generally supported by the data. At least three-fourths of these items were matched to their content area by the SMEs, and these areas were conspicuous in the MDS results. The results were not as good for the Technology area, due to the fact that only half of the 1999 Technology items were matched to this area. However, it was encouraging to see improved congruence for this area in 2000. The legitimacy of Inquiry as a content area was not supported by the results, but it should be noted that there were only three such items on each test.

Independent of the present study (note that we were not working for the state), the state department of education (DOE) responsible for developing this assessment revised the content specifications for this test, apparently due to concerns raised by science educators. It is interesting to note that the changes made by the state are consistent with the results of this study. The DOE removed the Inquiry content area and revised the Technology content area, renaming it Technology/Engineering. The three other content
areas, Life Sciences, Physical Sciences, and Earth/Space Sciences, which exhibited high congruence and stability in this study, were retained. Thus, this study serendipitously supports the recent content changes instituted by the DOE.

It is hard to make an absolute conclusion whether the 1999 or 2000 tests possess “adequate” content domain representation due to the lack of uniform standards for certifying content validity. However, these results suggest that at least three of the content areas held up to external scrutiny, and that two others either (a) need increased representation in the form of better test items, (b) need to be reconceptualized, or (c) need to be removed from the test specifications. The DOE’s recent revision of the content specifications is congruent with these suggestions. However, the results of this study also suggest that cognitive distinctions among the items deserve more attention in the test specifications.

SMEs’ Opinions and Recommendations Regarding Science Assessment

Although it was not central to the initial purpose of the study, we were also interested in the reactions of the SMES to Science and Technology assessment in general. Given the august group we gathered, we did not want to miss an opportunity to gather suggestions for improving Science and Technology assessment. Upon completion of the meeting we asked several questions about their personal observations and beliefs with respect to how the skills measured on these tests might be better communicated, how the statewide Science and Technology tests impacted the classroom, and how future large-scale Science and Technology tests should be improved. Perhaps not surprisingly, they made several comments about incorporating a cognitive taxonomy like Bloom’s into the test specification to better communicate the skills measured on the test. There was also
some feeling that this particular test is not measuring complex science skill levels as it purports to do.

As for improving the test, most felt that incorporation of more performance-based items was necessary, not only from the standpoint of assessing complex skills, but also of involving the science teachers as a component (e.g., where portfolios might be incorporated). A few comments were also made about how the use of language within test questions should be scrutinized more thoroughly. It was felt that sometimes a student may know the answer to a question but is confused by what is specifically being asked.

When asked about how the test corresponds to what is being taught in the classroom, two points stood out. One was that in some schools, physics courses were not offered until after the 10th grade (when the tests are administered). The other point was that there were still cases in which the curricular reforms that the test was based on had not been fully implemented in some schools. In those instances, the tests obviously did not match up well with what was being taught in class. Lastly, we asked about how this test had impacted classroom instruction. On the negative side were comments about increased pressure on teachers and students, test anxiety, and the observation that this system has not improved teaching. There was also some concern that important content may not be taught because it falls outside of the common curriculum and because of more teaching-to-the-test. On the positive side, comments were made about how a unified curriculum was desirable, how communication between teachers across districts had increased, and how there was more interest in professional development of teachers as a result. One interesting comment illustrated the case where classroom instruction may be driven by individual teacher strengths. With a unified curriculum, some teachers were
less comfortable teaching subjects that were not their strong suit. This was seen as both negative and positive in that it would drive teachers to seek out more professional development resources.

**Limitations of the Study**

Although the MDS results suggest the SMEs perceived cognitive features of the items first and content features second, reviewers of this study pointed out that the MDS results were probably contaminated by the fact that 7 of the 9 SMEs who rated the item similarities previously performed the item congruence ratings and so they were already familiar with the content and cognitive areas the tests intended to measure. Having prior knowledge of the test specifications is not desired when making item similarity ratings because this knowledge may implicitly lead to supporting the test specifications (Sireci, 1998a; Sireci & Geisinger, 1995). The extent to which the similarity ratings in this study were biased by this a priori knowledge is unknown, but clearly, future studies employing this technique should gather item similarity ratings before item congruence ratings.

Other limitations of this study include the use of a small sample of SMEs and analyzing the data across only two successive administrations. The degree to which the results would generalize to a larger group of SMEs and to assessments administered in other years is also unknown.

**Conclusion**

In this study, we illustrated how SMEs can be used to evaluate the content quality of a test, evaluate the consistency of this quality across years, and provide feedback to the test developers for improving test content and test score reporting. Messick (1989) and many other validity theorists underscore the notion that evaluating the validity of
inferences derived from test scores is an ongoing process. The information we discover through such evaluations should be fed back to the test developers, because test development is an ongoing process, too. It is our hope that this study and others like it will lead to better measurement of students’ knowledge, skills, and abilities, which should also lead to improved instruction and student learning.
References


