A Study of STEM Assessments in Engineering, Science, and Mathematics for Elementary and Middle School Students

Michael Harwell  
University of Minnesota

Mario Moreno  
University of Minnesota

Alison Phillips  
University of Minnesota

S. Selcen Guzey  
Purdue University

Tamara J. Moore  
Purdue University

Gillian H. Roehrig  
University of Minnesota

The purpose of this study was to develop, scale, and validate assessments in engineering, science, and mathematics with grade appropriate items that were sensitive to the curriculum developed by teachers. The use of item response theory to assess item functioning was a focus of the study. The work is part of a larger project focused on increasing student learning in science, technology, engineering, and mathematics (STEM)-related areas in grades 4–8 through an engineering design-based, integrated approach to STEM instruction and assessment. The fact that the assessments are available to school districts at no cost, and represent psychometrically sound instruments that are sensitive to STEM-oriented curriculum, offers schools an important tool for gauging students’ understanding of engineering, science, and mathematics concepts.

Recent reports have emphasized the importance of improving K-12 mathematics and science education in the United States in order to motivate more students to pursue science, technology, engineering, and mathematics fields in college (National Academy of Engineering [NAE] & the National Research Council [NRC], 2009; National Research Council, 2011). The emergence of a national consensus on the need to introduce K-12 students to science, technology, engineering, and mathematics (STEM) and prepare them for STEM majors in college has already generated significant changes as noted in the recent report of the NAE and the NRC:

Today, several dozen different engineering programs and curricula are offered in school districts around the country, and thousands of teachers have attended professional development sessions to teach engineering-related coursework. In the past 15 years, several million K–12 students have experienced some formal engineering education. (NAE & NRC, 2009, p. 1)

A consistent theme of these reports is the importance of integrating STEM education in K-12 to respond to the growing numbers of twenty-first-century jobs requiring knowledge from multiple STEM fields (NRC, 2005, 2011). However, STEM disciplines are generally taught as separate subjects in schools (NAE & NRC, 2009).

Clearly, new models of STEM learning and assessment are needed that attend to the specific content as well as the overarching ideas that integrate the STEM disciplines. STEM integration is defined by the merging of the disciplines of science, technology, engineering, and mathematics in order to: (a) deepen student understanding by contextualizing concepts; (b) broaden student understanding through exposure to socially and culturally relevant STEM contexts; and (c) increase interest in STEM disciplines and expand the pathways for students to enter STEM fields (Roehrig, Moore, Wang, & Park, 2012).

To promote the development of integrated models of STEM education, the National Science Foundation has begun to fund studies engaged in work of this nature. The current study is connected to a large mathematics and science project whose purpose is to increase student learning of engineering, science, and mathematics concepts in grades 4–8 using an integrative engineering design-based approach to teacher professional
development and curricular development. The project is partnering with teachers in multiple school districts to develop integrated engineering, science, and mathematics curricular materials for each of the major science topic areas within the standards for a Midwestern state for grades 4–8. The development of curricular materials that incorporate the standards by integrating engineering, science, and mathematics concepts is a key component of the project. Integrated STEM education curricular materials emphasize the connections between STEM subjects while using the context of real-world problems or challenges.

Briefly, teachers participating in the project receive intensive professional development during which they develop curricular materials corresponding to their teaching area (e.g., seventh-grade science) and which are linked to State standards. These teachers comprise the “treatment” cohort and subsequently teach using the curricular materials developed within the project. The impact of the professional development and the curricular materials are assessed in multiple ways including testing students. Teachers who do not participate in the professional development (and hence do not develop curricular materials tied to State standards) serve as a “business as usual” control cohort. Because teachers in the treatment are volunteers, the project employs a cohort (nonequivalent, quasi-experimental) cluster design in which engineering, science, and mathematics achievement data for approximately 20,000 students will be collected over four years. The project is currently in its second year.

The Measures of Academic Progress® (MAP) tests (Northwest Evaluation Association, 2011) and the State-mandated Minnesota Comprehensive Assessments (MCA) (Minnesota Department of Education, 2013) are used to obtain student achievement data in science and mathematics. These tests are administered at the end of a school year, after the new curricular materials in engineering, science, and mathematics have been taught to students whose teachers were in the treatment group. However, all students (treatment + control) take all tests.

The MAP and MCA tests in science and mathematics represent standardized instruments with considerable psychometric evidence supporting inferences based on test scores. However, there are two key concerns related to these tests within the project. First, the MAP test does not cover engineering and the State-mandated science test covers a few engineering topics but these are quite limited. Second, the MAP mathematics test and the MCA science and mathematics tests provide a general assessment of these topics but are not linked to the curricular materials being developed, and the MAP is not directly linked to the State standards. Because assessing student knowledge tied to State standards is crucial in evaluating the impact of the treatment on achievement, assessments in engineering, science, and mathematics were developed within the project to be used in conjunction with the MAP and MCA tests.

The purpose of this study was to develop, scale, and validate assessments in engineering, science, and mathematics with grade appropriate items drawn from national item banks that were sensitive to the curriculum developed by teachers that incorporated state-mandated standards in engineering, science, and mathematics. To our knowledge, no such assessments currently exist. The assessments may supplement existing measures used by schools or serve as a primary indicator of student understanding of STEM concepts.

**Student Learning and STEM Education**

While many recent reforms call for improving STEM education, there is little research on the impact of STEM education, particularly integrated STEM education, on student learning and achievement. In a recent report, entitled STEM integration in K-12 education: Status, prospects, and agenda for research, the NAE/NRC committee on integrated STEM education reviewed the research on the impact of integrated approaches on student learning and concluded that “the positive impact on learning appears to differ for science and mathematics, with less evidence of a positive impact on mathematics outcomes, based on current assessments for those subject areas, which might not fully capture integrated learning in STEM” (NAE & NRC, 2014, p. 6). To document potential benefits of integrated STEM education, robust research on student learning is necessary. However, there are two factors that limit integrated STEM education research: a variety of STEM integration approaches and a lack of assessment tools to measure student learning and achievement resulting from integrated STEM education.

The method and degree of integration of STEM subjects varies (Bybee, 2013). While one form of integration may focus on combining two STEM subjects, another approach may use a real-world related problem or challenge that requires students to apply all STEM subjects to understand and solve the challenge. Thus, STEM integration takes place in classrooms in a variety of forms. Ideally, STEM integration can be done when the integration and
interconnectedness of STEM subjects is made explicit to the students. However, not all students receive enough support to build knowledge across STEM subjects since each STEM integration approach is unique to grade levels, teachers, and schools. The integrated STEM education approaches and the context within which integrated STEM education vary, and consequently present challenges to researchers designing and conducting studies of student learning in STEM education settings (NAE & NRC, 2014).

Another major challenge for studies focusing on integrated STEM education is the lack of measurement instruments for student learning. “Existing assessments tend to focus on knowledge in a single discipline. Furthermore, they typically focus on content knowledge alone and give little attention to the practices in the disciplines and applications of knowledge” (NAE & NRC, 2014, p. 6). As addressed in engineering in K-12 education: Understanding the status and improving the prospects (NAE & NRC, 2009), A Framework for K-12 Science Education (NRC, 2012), and Next Generation Science Standards (NRC, 2013), scientific and engineering practices are key factors of integrated education. Measuring the learning required in the framework and Next Generation Science Standards (NGSS) requires the development of assessments that measure understanding of content knowledge, connections across STEM subjects, and practices of STEM subjects. While modification of existing assessment tools is useful, new assessments that are sensitive to the impact of integrated STEM education on student learning are also needed.

As recommended by the NAE/NRC committee on integrated STEM education, researchers should clearly document the development and use of appropriate assessment tools and STEM integration approaches (NAE & NRC, 2014). This would help in identifying existing STEM approaches and collecting more robust evidence on the impact on integrated STEM education on student learning. In our research study, the assessments were developed to measure student learning and achievement that resulted from integrated STEM education units developed and implemented by teachers. The STEM integration approach involves the use of engineering as a vehicle to teach science, mathematics, and technology. Students are given an engineering challenge or problem to solve by applying science and mathematics concepts that they learn through the STEM unit. The development of our assessment tools is presented in the next section.

Methodology

The development of assessments was largely guided by the standards in an upper Midwestern U.S. state in engineering, science, and mathematics for grades 4–8 and the process described in the Standards for Educational and Psychological Testing (American Educational Research Association, 1999). Because of the project’s focus on integrated STEM instruction and assessment, students take the three assessments (engineering, science, mathematics) as a single test at the beginning and at the end of the STEM unit(s) in which these topics are covered.

Assessment Development

The test construction process began with an assessment development team of classroom teachers, school curriculum specialists, and academic researchers with collective expertise in engineering, science, and mathematics. This team clarified the purpose of each test (e.g., assessing earth science knowledge) and carefully described the knowledge domain to be sampled (e.g., plate tectonics, consequences of erosion). A curricular map of topics consistent with State standards (e.g., earth science) embedded in the teacher-constructed STEM curricula were then developed. This produced a content domain of topics and knowledge which test items should reflect dependent on a student’s grade.

A pool of multiple choice items was then generated using the above test specifications for each content test for upper elementary (grades 4–5) and middle school (grades 6–8) students in the following areas:

1. earth science (plate tectonics, grades 5 and 8).
2. earth science (erosion, grades 5 and 7).
3. engineering design (grades 4–8).
4. life science (ecosystems and environments, grades 5 and 7).
5. mathematics (data analysis and measurement, grades 4–8).
6. physical science (states of matter, grades 4 and 6).
7. physical science (heat transfer, grades 4 and 6).

The intent was that items on each content test reflect a single underlying factor (e.g., earth science knowledge).

All items were initially obtained from public item banks linked to the Trends in International Mathematics and Science Study (TIMMS) (International Association for the Evaluation of Educational Achievement, 2007), National Assessment of Educational Progress (NAEP) (National Center for Education Statistics, 2010), and the American Association for the Advancement of Science (AAAS, 2014). Central to selecting items was the
requirement that they assess topics and knowledge within a content domain consistent with the curricular map developed in step 1. Items were in some cases modified slightly to be consistent with State standards and with the kinds of tests students in the partner school districts were familiar with.

The items were also multiple choice items scored correct/incorrect. This feature was important because it imposed less of a burden on the class time of teachers participating in the project and their students compared with open-ended items. Another advantage of multiple choice items is that they are relatively inexpensive to score which is particularly important when large numbers of students are involved. Relatedly, because the content tests will be available to the broader educational community, a multiple choice format likely makes the assessments more attractive to school districts.

The result was a pool of items that were quite similar across grades for a given content area with modest changes in presentation in a few items to reflect differences in students’ reading, comprehension, and reasoning skills. For example, the earth science assessment contained an item about a characteristic of erosion that would be presented somewhat differently to fifth-grade students compared with seventh-grade students. On the other hand, an engineering design item that fourth- and fifth-grade students complete would be very similar, meaning that students in these grades would respond to essentially the same item. This should facilitate combining the test scores on each assessment of students in grades 4 and 5, and, likewise, combining the scores of students in grades 6, 7, and 8. Once drafted, the items were carefully vetted by the assessment development team and revised or omitted based on a group consensus.

Our goal was to construct assessments that in their final form had 10 items each for engineering, science, and mathematics for grades 4 and 5, and 15 items each for assessments for grades 6–8. Thus fourth- and fifth-grade students participating in the project complete a 30-item test in one hour at the beginning and at the end of the instructional unit(s) in which these topics are covered; middle school students complete a 45-item assessment in one hour at the same time points. The difference in the length of the assessments between upper elementary students and middle school students was a function of their expected ability to comprehend and complete the assessments in one hour. Naturally, we would prefer more items per assessment but had to balance this against the burden longer tests impose on students and on class time. These concerns were exacerbated by the fact that several items require substantial reasoning and involve pictures, charts, or diagrams.

Results

Item Analyses and Scaling

Next, the assessments were piloted in two waves with the goal of identifying final versions of each assessment via item analyses and scaling students’ responses to produce an estimate of their proficiency. In the first wave, the assessments were administered to a small group of about 10–20 students in each grade. The purpose was to obtain and analyze preliminary item response data as well as obtain feedback on characteristics of the assessments and the testing environment that affected performance (e.g., readability, clarity of items, were students able to finish the assessment in the allotted time, were students able to get questions they had during the test answered).

Item analyses consisting of the proportion of students selecting each option for an item and point biserial correlations were examined. The latter represents the correlation between whether a student responded correctly to an item and their total correct score. Traditionally a large positive point biserial correlation is taken as evidence that an item is functioning as desired, i.e., students who answer an item correctly are likely to have an above average total correct score. A negative point biserial correlation suggests an item is not functioning as desired (Mehrens & Lehmann, 1991). Feedback about characteristics of the tests and testing environment was provided by classroom teachers who administered the assessments and used a project-developed protocol to record features of the tests and testing they thought relevant, such as students asking several questions about particular items. Based on information provided by wave 1 piloting the assessments were modified by the assessment development team.

In the second wave of piloting the revised assessments were administered to approximately 100 students in grades 4 and 5 (total of about 200) and approximately 75 in grades 6–8 (total of about 225). For various reasons including classroom size and students who were absent, the piloted assessments were based on different numbers of students.

Data from wave 2 were used to conduct extensive item analyses of the assessments. We initially performed a factor analysis of the data for each assessment for elementary school students, and, separately, for middle school students. The goal of these analyses was to assess
the likelihood that items represented a single underlying construct (e.g., earth science erosion). The results suggested that a single (major) factor underlies each assessment, a finding that was robust to different methods of factoring (principal axis, maximum likelihood) and factor rotation (varimax, oblique). These results suggest that the goal of constructing items for each assessment that reflected a single factor was generally met, and also helped to justify the use of item response theory (IRT) (discussed below) to generate proficiency scores for students.

Next the Rasch IRT model was fitted to the data for each item on each assessment to estimate that item’s ability to contribute to estimates of student proficiency (e.g., earth science erosion) (Embretson & Reise, 2000). The Rasch model characterizes the relationship between a student’s response to an item and their estimated proficiency on an assessment \( \hat{\theta} \) using a difficulty parameter. The difficulty parameter for an item is the location of the item on the proficiency scale in logits (Winsteps & Rasch Measurement Software, 2010) for which the probability of a correct response to the item is .5. Larger positive values of a difficulty parameter imply a more difficult item and signal that substantial proficiency is needed to have a probability of responding correctly greater than .5. Put another way, a more difficult item requires that a student have a large positive value of \( \hat{\theta} \) to have a high probability of answering the item correctly.

IRT analyses were done separately for elementary and middle school students. Item fit statistics reported by the Winsteps software (Winsteps & Rasch Measurement Software, 2010) were used to assess how well the model fitted the data for each item. These measures reflect information in residuals that reflect what the Rasch model predicted for an item and the actual student responses for that item; residuals near zero suggest that the model adequately fits the data for an item and can be used to scale responses in the form of \( \hat{\theta} \). On the other hand, residuals far from zero for an item suggest the Rasch model does not adequately fit the responses (shows poor fit) and that the item cannot be used to scale responses (i.e., produce \( \hat{\theta} \) estimates) (Wright & Masters, 1982). Because item fit information provided by Winsteps was a critical part of identifying items for the final assessments, this procedure is briefly described below.

Information about item fit began by using Winsteps to fit the Rasch model to student responses and estimate each student’s proficiency \( \hat{\theta} \). The software then constructs a \( 2 \times J \) contingency table for each item in which rows indicate whether the item was answered correctly, columns reflect a partitioning of the estimated proficiency \( \hat{\theta} \) variable into \( J \) intervals, usually 10, and row and column marginal frequencies that are assumed to be fixed, meaning that if the study was replicated the same marginal values would be observed (Marascuilo & McSweeney, 1977). Each cell in this table contains frequency information about discrepancies (residuals) between what is expected and observed for the model. For example, a simple table with only three \( \hat{\theta} \) intervals for hypothetical data might look like that in Table 1. In Table 1, \( O_{ij} \) is the observed number of students in the ith row (i = 1, 2, or 3) column, and \( E_{ij} \) represents the number of students (in parentheses) expected in a cell if the statistical null hypothesis of no relationship between proficiency and item response is true.

The proportion of students responding correctly to the item in Table 1 is \( \frac{70}{100} = .7 \) which, under the null hypothesis, should be distributed equally across the cells in the first row i.e., regardless of proficiency we expect 70% of the students in each \( \hat{\theta} \) interval to answer correctly. Thus \( E_{11} = (.70)(14) = 9.8, E_{12} = (.70)(30) = 21 \) and the remaining \( E_{ij} \) can be found by subtraction using the marginal frequencies.

The Winsteps software draws on the traditional chi-square test for contingency tables

\[
\chi^2 = \sum_{i=1}^{I} \sum_{j=1}^{J} \frac{(O_{ij} - E_{ij})^2}{E_{ij}}
\]

(Marascuilo & McSweeney, 1977) by exploiting the fact that each cell in Table 1 provides information about fit through \( \frac{(O_{ij} - E_{ij})^2}{E_{ij}} \); the larger this value the larger the discrepancy and the poorer the fit. The purpose of this approach is to identify discrepancies. For example, we would intuitively expect that most students who answered the item correctly would fall in the \( \hat{\theta} > +1 \) interval in

<table>
<thead>
<tr>
<th>Proficiency (( \hat{\theta} ))</th>
<th>Item Response</th>
<th>Row Total</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>( \hat{\theta} &lt; -1 )</td>
<td>( -1 \leq \hat{\theta} \leq +1 )</td>
</tr>
<tr>
<td>Correct</td>
<td>5 (9.8)</td>
<td>10 (21)</td>
</tr>
<tr>
<td>Incorrect</td>
<td>25 (4.2)</td>
<td>4 (9)</td>
</tr>
<tr>
<td>Column Total</td>
<td>14</td>
<td>30</td>
</tr>
</tbody>
</table>

Table 1. Sample Table for an Item Illustrating How Fit Measures in Winsteps Are Constructed
Table 1 and far fewer students in the $\hat{\theta} < -1$ interval. If this pattern appeared then $\sum_{j=1}^{J} \sum_{i=1}^{I} \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$ would be close to its expected value of $J-1$. On the other hand, large discrepancies across cells imply that (conditional on marginal frequencies) more students in the lowest ability interval answered the item correctly than expected, producing a $\sum_{j=1}^{J} \sum_{i=1}^{I} \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$ much larger than its expected value of $J-1$ and evidence of poor fit. The evidence in Table 1 is that the model fits the data for this item adequately.

Equation (1) could also be used to test the statistical null hypothesis of no relationship between item responses and $\hat{\theta}$ intervals but Winsteps instead provides various fit indices to guide decisions about whether an item is functioning as intended. The rationale for avoiding significance tests is presumably to discourage binary decision making about item fit as a result of whether a statistical test is significant.

Building on these ideas, Table 2 provides an example of Winsteps output based on wave 2 data that were used to make decisions about items for the grades 6–8 engineering assessment. Shaded items were eliminated to reduce the test to the desired length of 15 items.

In Table 2, ENTRY represents items and MEASURE represents each item’s estimated difficulty. For item 1, the estimated difficulty parameter of 3.104 means that the probability a student whose estimated proficiency $\hat{\theta} = 3.104$ will answer this item correctly is .5; a student with $\hat{\theta} > 3.104$ has a probability of answering this item correctly greater than .5 whereas one with $\hat{\theta} < 3.104$ has a probability of answering correctly of less than .5. Because difficulty (MEASURE) is in standard deviations $\hat{\theta} = 3.104$ for item 1 means that this item was very difficult for students.

The COUNT column in Table 2 represents the total number of students, IN.MSQ is the “infit mean square” which is the chi-square in equation (1) divided by its degrees of freedom and weighted by a value that gives greater importance to values near the center of the proficiency distribution; OUT.MS is the same as IN.MSQ but without the weighting. IN.ZSTD and OUT.ZSTD are simply standardized versions of IN.MSQ and OUT.MS, respectively, and are approximately normally distributed with a mean of zero. We focused on IN.MSQ and OUT.MS. Winsteps recommends that items with IN.MSQ or OUT.MS values above 1.5 be treated as potentially problematic (Winsteps & Rasch Measurement Software, 2010). PTME represents the point-to-measure correlation (i.e., correlation between total test scores and responses for an item).

A careful look at the IN.MSQ and OUT.MS values in Table 2 suggests that the Rasch model adequately fits the data of all 21 items. This result is not surprising because

<table>
<thead>
<tr>
<th>Entry</th>
<th>Measure</th>
<th>Count</th>
<th>IN.MSQ</th>
<th>IN.ZSTD</th>
<th>OUT.MS</th>
<th>OUT.ZSTD</th>
<th>PTME</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>3.104</td>
<td>168</td>
<td>1.23</td>
<td>1.1</td>
<td>7.68</td>
<td>7.07</td>
<td>-.12</td>
</tr>
<tr>
<td>2</td>
<td>-.561</td>
<td>168</td>
<td>1.08</td>
<td>.84</td>
<td>1.01</td>
<td>.14</td>
<td>.41</td>
</tr>
<tr>
<td>3</td>
<td>-.257</td>
<td>168</td>
<td>1.02</td>
<td>.25</td>
<td>1.11</td>
<td>.97</td>
<td>.43</td>
</tr>
<tr>
<td>4</td>
<td>-1.321</td>
<td>168</td>
<td>1.03</td>
<td>.26</td>
<td>1.01</td>
<td>.11</td>
<td>.42</td>
</tr>
<tr>
<td>5</td>
<td>.659</td>
<td>168</td>
<td>1.17</td>
<td>2.63</td>
<td>1.35</td>
<td>2.77</td>
<td>.28</td>
</tr>
<tr>
<td>6</td>
<td>-1.23</td>
<td>168</td>
<td>.82</td>
<td>-1.49</td>
<td>.83</td>
<td>-.85</td>
<td>.56</td>
</tr>
<tr>
<td>7</td>
<td>.778</td>
<td>168</td>
<td>.93</td>
<td>-1.04</td>
<td>1</td>
<td>.05</td>
<td>.47</td>
</tr>
<tr>
<td>8</td>
<td>-.975</td>
<td>168</td>
<td>.94</td>
<td>-.5</td>
<td>.96</td>
<td>-.16</td>
<td>.48</td>
</tr>
<tr>
<td>9</td>
<td>.242</td>
<td>168</td>
<td>1.01</td>
<td>.22</td>
<td>1.03</td>
<td>.36</td>
<td>.44</td>
</tr>
<tr>
<td>10</td>
<td>.391</td>
<td>168</td>
<td>1.08</td>
<td>1.21</td>
<td>1.14</td>
<td>1.29</td>
<td>.38</td>
</tr>
<tr>
<td>11</td>
<td>-1.23</td>
<td>168</td>
<td>.76</td>
<td>-2.06</td>
<td>.62</td>
<td>-2.18</td>
<td>.61</td>
</tr>
<tr>
<td>12</td>
<td>-.935</td>
<td>168</td>
<td>.74</td>
<td>-2.61</td>
<td>.63</td>
<td>-2.54</td>
<td>.64</td>
</tr>
<tr>
<td>13</td>
<td>-.895</td>
<td>168</td>
<td>.78</td>
<td>-2.15</td>
<td>.68</td>
<td>-2.14</td>
<td>.61</td>
</tr>
<tr>
<td>14</td>
<td>1.111</td>
<td>168</td>
<td>1.14</td>
<td>2.06</td>
<td>1.3</td>
<td>1.95</td>
<td>.28</td>
</tr>
<tr>
<td>15</td>
<td>.51</td>
<td>168</td>
<td>1.04</td>
<td>.7</td>
<td>1.08</td>
<td>.79</td>
<td>.4</td>
</tr>
<tr>
<td>16</td>
<td>.778</td>
<td>168</td>
<td>1.01</td>
<td>.23</td>
<td>1.08</td>
<td>.67</td>
<td>.4</td>
</tr>
<tr>
<td>17</td>
<td>.302</td>
<td>168</td>
<td>1.17</td>
<td>2.39</td>
<td>1.15</td>
<td>1.39</td>
<td>.33</td>
</tr>
<tr>
<td>18</td>
<td>-.526</td>
<td>168</td>
<td>.8</td>
<td>-2.33</td>
<td>.69</td>
<td>-2.6</td>
<td>.61</td>
</tr>
<tr>
<td>19</td>
<td>-.596</td>
<td>168</td>
<td>.8</td>
<td>-2.24</td>
<td>.72</td>
<td>-2.26</td>
<td>.6</td>
</tr>
<tr>
<td>20</td>
<td>-.78</td>
<td>168</td>
<td>.83</td>
<td>-1.75</td>
<td>.76</td>
<td>-1.71</td>
<td>.58</td>
</tr>
<tr>
<td>21</td>
<td>1.432</td>
<td>168</td>
<td>1.2</td>
<td>2.46</td>
<td>1.52</td>
<td>2.62</td>
<td>.2</td>
</tr>
</tbody>
</table>
items were taken from large national assessments like TIMSS, NAEP, and AAAS with documented evidence of their psychometric prowess. Thus we had to look elsewhere for evidence to remove items to produce the desired 15-item test. The largest IN.ZSTD standardized residuals were for items five, 12, 17, 18, and 21 and these items were removed. To get to a 15-item test, we used the rather inelegant but functional point biserial correlation; because this value was negative for item 1, it was removed leaving us with the final version of this assessment.

A proficiency score for students completing the engineering assessment for grades 6–8 was then estimated using the final 15-item version of this test. To ensure that scores are reliable indicators of proficiency, we used WINSTEPS “person reliability” because of its focus on classifying students via their proficiency estimates (WINSTEPS & Rasch Measurement Software, 2010). IRT-based reliability is increasingly recommended rather than the traditional Cronbach’s alpha (Sijtsma, 2009) but these estimates tend to be smaller than alpha. Person reliabilities for the assessments are reported in Table 3. On the whole, the reliabilities for middle school students are higher due to the longer tests but are somewhat smaller than desired.

Once test scores were available for elementary and middle school students, we combined them for use in subsequent analyses. The assessments completed by these two groups of students were similar but not exactly the same, and to ensure that proficiency scores share a common meaning, an equating study was performed for each assessment across elementary and middle school students following the methods outlined in Kolen and Brennan (2004). This produced values on an assessment (e.g., engineering knowledge) that shared a common metric and meaning, allowing the elementary and middle school samples to be combined when the assessment served as an outcome in subsequent data analyses. Equating was applied separately to assessments in engineering, science, and mathematics for elementary and middle school students.

**Related Analysis Issues**

An important statistical consequence of lower reliability is that analyses of proficiency scores would contain more random error variation than would be the case if reliabilities were higher. However, our plan to use large numbers of students in the project data analyses suggest this extra variation will not significantly impact precision of parameter estimation or statistical power. Still, these findings suggest it would be appropriate to add a few items to each assessment especially those for upper elementary school students to increase reliability.

Reliability evidence must be accompanied by validity evidence, which for the assessments in engineering, science, and mathematics for the second wave of pilot data had two complementary sources. The fact that all of the items were taken from national assessments like TIMSS and possessed substantial evidence of their validity enhanced this process. However, it was important to attend to two types of validity in constructing the final versions of the assessments.

The first type was content-related evidence of validity that reflects the extent to which test items capture relevant facets of a specified knowledge domain such as engineering knowledge of fourth and fifth-grade students. Content validity specifically requires evidence that a thorough examination of the knowledge (subject) domain was conducted and that items were carefully written and included because they complied with what a test is intended to do and how it will accomplish this goal (i.e., test specifications) (Anastasi & Urbina, 1997).

In arguing for content validity, the use of a group of experts in the content area to develop and assess the extent to which the items reflect relevant facets of knowledge is crucial. As noted earlier, we employed an assessment development team of classroom teachers, school curriculum specialists, and academic researchers with collective expertise in engineering, science, and mathematics, and this group provided critically needed expertise in test development.

The content expertise of the assessment development team specifically helped to ensure that the: (a) knowledge domain was clearly specified and took a student’s grade and the STEM-oriented curriculum they experienced into account; (b) knowledge measured by an item was clearly defined and led to an unambiguous definition for evidence to remove items to produce the desired 15-item test. The largest IN.ZSTD standardized residuals were for items five, 12, 17, 18, and 21 and these items were removed. To get to a 15-item test, we used the rather inelegant but functional point biserial correlation; because this value was negative for item 1, it was removed leaving us with the final version of this assessment.

A proficiency score for students completing the engineering assessment for grades 6–8 was then estimated using the final 15-item version of this test. To ensure that scores are reliable indicators of proficiency, we used WINSTEPS “person reliability” because of its focus on classifying students via their proficiency estimates (WINSTEPS & Rasch Measurement Software, 2010). IRT-based reliability is increasingly recommended rather than the traditional Cronbach’s alpha (Sijtsma, 2009) but these estimates tend to be smaller than alpha. Person reliabilities for the assessments are reported in Table 3. On the whole, the reliabilities for middle school students are higher due to the longer tests but are somewhat smaller than desired.

Once test scores were available for elementary and middle school students, we combined them for use in subsequent analyses. The assessments completed by these two groups of students were similar but not exactly the same, and to ensure that proficiency scores share a common meaning, an equating study was performed for each assessment across elementary and middle school students following the methods outlined in Kolen and Brennan (2004). This produced values on an assessment (e.g., engineering knowledge) that shared a common metric and meaning, allowing the elementary and middle school samples to be combined when the assessment served as an outcome in subsequent data analyses. Equating was applied separately to assessments in engineering, science, and mathematics for elementary and middle school students.

**Related Analysis Issues**

An important statistical consequence of lower reliability is that analyses of proficiency scores would contain more random error variation than would be the case if reliabilities were higher. However, our plan to use large numbers of students in the project data analyses suggest this extra variation will not significantly impact precision of parameter estimation or statistical power. Still, these findings suggest it would be appropriate to add a few items to each assessment especially those for upper elementary school students to increase reliability.

Reliability evidence must be accompanied by validity evidence, which for the assessments in engineering, science, and mathematics for the second wave of pilot data had two complementary sources. The fact that all of the items were taken from national assessments like TIMSS and possessed substantial evidence of their validity enhanced this process. However, it was important to attend to two types of validity in constructing the final versions of the assessments.

The first type was content-related evidence of validity that reflects the extent to which test items capture relevant facets of a specified knowledge domain such as engineering knowledge of fourth and fifth-grade students. Content validity specifically requires evidence that a thorough examination of the knowledge (subject) domain was conducted and that items were carefully written and included because they complied with what a test is intended to do and how it will accomplish this goal (i.e., test specifications) (Anastasi & Urbina, 1997).

In arguing for content validity, the use of a group of experts in the content area to develop and assess the extent to which the items reflect relevant facets of knowledge is crucial. As noted earlier, we employed an assessment development team of classroom teachers, school curriculum specialists, and academic researchers with collective expertise in engineering, science, and mathematics, and this group provided critically needed expertise in test development.

The content expertise of the assessment development team specifically helped to ensure that the: (a) knowledge domain was clearly specified and took a student’s grade and the STEM-oriented curriculum they experienced into account; (b) knowledge measured by an item was clearly defined and led to an unambiguous definition
of the knowledge element tapped by the item and a clear representation of the knowledge (e.g., pictures designed to reflect the concept of erosion were clear); and (c) settings in which the knowledge is used or needed were clearly explained.

A second and complementary source of validity evidence was curricular validity which is the extent to which the content of a test matches the objectives of a specific curriculum. The process whereby teachers in the treatment group worked with project personnel to develop a curriculum that reflected State standards in a topic area (e.g., earth science erosion) was marked by teamwork, rigor, and dedication to this laborious but crucial task. More formal evidence of curricular validity came from an assessment of the alignment of the curricula to a curricular map of topics and objectives generated by the assessment development team. Among the strategies used to assess this alignment was a curriculum assessment tool developed as part of the project. This tool was used by at least three project staff to assess each curricular unit (e.g., earth science erosion) on criteria related to State standards and key features of STEM education such as integration of content, engineering design, instructional strategies, and communication. Evidence from these ratings indicated that the curricula were generally consistent with the curricular map, which was used to select items from national item banks like those for TIMSS.

Collectively, evidence of validity of the items is strong, in part because the items were taken from national item banks and had considerable validity evidence, and in part because of evidence that content and curricular validity is quite strong. Collectively this evidence supports our assertion that test scores (e.g., earth science erosion) permit valid inferences about student proficiency.

Implications

There is wide agreement of the need to improve K-12 mathematics and science education in the United States and evidence that integrating STEM education in K-12 is a promising mechanism for doing so (NAE & NRC, 2009; NRC, 2011). Key to this process is the availability to school districts of psychometrically sound assessment tools to assess the impact of STEM-oriented instruction that incorporate State-mandated standards in engineering, science, and mathematics (NAE & NRC, 2014). To our knowledge, no such assessments currently exist.

This study reported on the process of developing, scaling, and validating assessments sensitive to STEM-oriented instruction reflecting engineering, science, and mathematics standards in one upper Midwestern state for upper elementary and middle school students. The results provide evidence that the assessments represent psychometrically sound instruments sensitive to STEM-oriented curriculum, and thus offer school districts an important tool for gauging the impact of an engineering design-based approach to teacher professional development and curricular development on students’ understanding of STEM concepts. The relatively short length of the grade-specific tests (10 or 15 items for each content area) suggests they will not overly burden teachers or students, and the fact the items are multiple choice implies they can be scored cheaply.

Of course, each state’s standards will differ somewhat but the use of items from national items banks to assess key facets of engineering, science, and mathematics suggests the assessments are likely to be sufficiently flexible to be used in different settings, for example, to supplement other assessments used in a school district or to serve as a primary indicator of a student’s STEM understanding. The fact that these assessments are available to school districts at no cost enhances the importance of the study.

References


Authors’ Notes

Keywords: STEM, measurement, engineering.

Correspondence concerning this article should be addressed to Michael Harwell, Department of Educational Psychology, University of Minnesota, 176 EdSciB 56 East River Road, Minneapolis, MN 55455, USA. Tel: 612-625-0196; E-mail: harwe001@umn.edu