Alignment has taken on increased importance given the current high-stakes nature of assessment. To make well-informed decisions about student learning on the basis of test results, assessment items need to be well aligned with standards. Project 2061 of the American Association for the Advancement of Science (AAAS) has developed a procedure for analyzing the content and quality of assessment items. The authors of this study used this alignment procedure to closely examine 2 mathematics assessment items. Student work on these 2 items was analyzed to determine whether the conclusions reached through the use of the alignment procedure could be validated. It was found that the Project 2061 alignment procedure was effective in providing a tool for in-depth analysis of the mathematical content of the item and a set of standards and in identifying 1 particular content standard that was most closely aligned with the standard. Through analyzing student work samples and student interviews, it was also found that students' thinking may not correspond to the standard identified as best aligned with the learning goals of the item. This finding highlights the potential usefulness of analyzing student work to clarify any additional deficiencies of an assessment item not revealed by an alignment procedure.
Two forces in American education currently have a political and practical impact on assessment, especially in K–12 mathematics. An ever increasing demand for accountability is fueling the use of greater numbers of tests that have high-stakes implications for students, teachers, schools, and public officials. At the same time, calls for higher standards and fears that American students may not be competitive are pressing teachers to address achievement on tests and also to develop mathematical skills, understanding, and problem-solving abilities for all students. These concurrent demands make it increasingly essential that assessments be well aligned with important learning goals (i.e., national standards) and that they have a high degree of validity, so they reflect student learning fairly and accurately.

Many states, districts, and schools are attempting to develop new curricula and new assessment instruments to determine the extent to which students are meeting new academic standards. The issue of alignment is becoming central to these developments. Policymakers, educators, and journalists have called for better alignment between standards and assessments (National Council of Teachers of Mathematics [NCTM], 1995; Quality Counts, 2001; Romberg & Wilson, 1995; Romberg, Wilson, Khaketla, & Chavarria, 1992). Attention to the idea of analyzing assessments for their alignment and quality is on the rise. Webb (1999), for example, analyzed state mathematics assessments to examine their coverage of standards content and their cognitive level. Blank, Porter, and Smithson (2001) used an alignment procedure like Webb’s to study the alignment of state mathematics and science assessments with classroom instructional practices. An in-depth analysis of the 1996 National Assessment of Educational Progress (NAEP) mathematics test (Silver & Kenney, 2000), while not applying a particular procedure, provides another example of the interest and value of careful study of assessment items.

Alignment in the context of assessment is usually taken to mean the degree to which a test (or test item) assesses the same learning goals as a given standard or set of standards (Wilson & Kenney, 2003). Due to the high-stakes nature of assessment as a tool for educational reform, issues of alignment have taken on increased immediacy. If test items are poorly aligned with standards, the high-stakes decisions made on the basis of the results of the tests may not be based on valid information (Rothman, Slattery, Vranek, & Resnick, 2002). Calls for alignment have been addressed in varied ways (e.g., LaMarca, Redfield, Winter, Bailey, & Despriet, 2000; Webb, 1999). Several projects have been initiated to evaluate the alignment of an entire test with a given set of standards. This study, in contrast, takes a more focused and in-depth approach to the alignment of individual assessment items with one or more learning standards.

The analysis of assessment items has the potential to aid in the reform of mathematics and science education. Better information about student learning can help improve instruction as well as testing practices. However, a useful and trustworthy procedure for analysis takes time and effort to develop and implement. Project
2061 of the American Association for the Advancement of Science (AAAS), building on experience with the analysis of curriculum materials in mathematics and science (Kulm, 2000b; Kulm, Roseman, & Treistman, 1999) has developed a procedure for analyzing the content and quality of assessment items. This work has produced information about a few items, as well as about the challenge of doing such an analysis (Kulm, 2000a).

The authors of this study have attempted to use this alignment procedure to closely examine two specific assessment items in mathematics. Our goal was to determine whether analyzing student work for these items could validate the conclusions reached by the alignment procedure. We used the following research questions to guide our work:

1. How well does each item align with content standards for mathematics, according to the Project 2061 procedure? Do student data support or refute these conclusions?
2. What is the likely effectiveness of the items in assessing student attainment of specific mathematics learning goals?
3. What can the student data tell us about the effectiveness of the items and does this support or refute the conclusions based on the alignment procedure?

This study provides insight into an intensive analysis procedure and raises some general questions about the interaction between alignment and validity of open-ended items in mathematics.

THE ALIGNMENT ANALYSIS PROCEDURE

AAAS Project 2061 researchers developed the alignment procedure over a period of 2 years, with input from a national advisory board of science and mathematics educators and assessment experts. The intent, unlike efforts of many other procedures that analyze an entire test, was to develop an alignment tool that would focus on individual items and provide in-depth information about the alignment of an item with a specific learning goal or objective. Similar but separate versions of the procedure were developed for science and mathematics. The initial version of the mathematics procedure was pilot tested by teams of trained analysts and the analysis was then applied to a variety of released items from state mathematics assessments and from the NAEP. The procedure was revised and pilot tested again by another group of trained analysts, then revised to obtain the version used in this study.

The alignment procedure provides data on the extent to which a mathematics or science assessment item (a) is aligned with the content of a learning goal and (b) is of sufficient quality to provide valid information about student learning of the goal.
It is important to note that an item cannot be of high quality unless it assesses an important content learning goal. Similarly, even if an item seems to be well aligned with a learning goal, it must have sufficient quality to be useful as an assessment tool (Kitchen & Wilson, 2004). Quality is, in one sense, a validity issue. Items of high quality can, when accumulated, yield valid inferences about student learning. However, if there are flaws in an item, there will likely be impediments to the validity of the inferences made.

The analysis itself consists of four stages as outlined in Table 1. Following the analysis, a final activity is to prepare a report that provides a profile of the assessment item. The profile consists of ratings and commentary on the Content and Likely Effectiveness criteria. Typically, two independent teams complete the analysis, working to reconcile their ratings. The report includes an account of this process, including a discussion of criteria on which the analysts disagree or have different interpretations.

BRIEF LITERATURE REVIEW OF ALIGNMENT CRITERIA

The criteria in the AAAS alignment procedure are intended to reflect research on effective assessments and to help test writers and teachers develop items that have a greater potential for making highly valid inferences about student learning. The AAAS alignment procedure reflects criteria that are used to create and evaluate the effectiveness of cognitively demanding problems such as performance-based assessments and constructed response items. Some researchers believe that the observed limits on the measurement quality of performance assessments due to the relatively small number of items that contribute to an assessment score may be altered significantly through attention to assessment design and related scoring processes (see Linn, 2000; Linn, Baker, & Dunbar, 1991; Linn, DeStefano, Burton, & Hanson, 1995). These researchers have suggested that impressive gains can be obtained by careful consideration of the manner in which information is communicated to examinees and assessors. Their data demonstrate that we can improve the technical quality of performance assessment by attending to the cognitive demands placed on those taking the assessments and those scoring them.

Content Alignment

We first consider the issue of content alignment. Traditional test development requires only that item developers write items intended to align, at least topically, with test content matrixes or specifications. It is difficult to determine which, if any, precise guidelines writers use to align an item with specific mathematics content.
<table>
<thead>
<tr>
<th>Analysis Component</th>
<th>Brief Description</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Groundwork</strong></td>
<td></td>
</tr>
<tr>
<td>Item completeness</td>
<td>Make sure the complete item, scoring guide, directions for administration, and sample student responses or p-values are available for review.</td>
</tr>
<tr>
<td>Item clarification</td>
<td>Identify prerequisite concepts, skills needed for response, possible misconceptions, multiple solution strategies.</td>
</tr>
<tr>
<td>Candidate goals</td>
<td>Tentatively identify all relevant learning goals, or parts of learning goals required for a response.</td>
</tr>
<tr>
<td>Goal clarification</td>
<td>Compare the identified learning goals with other standards, review research findings, and define terms and conditions.</td>
</tr>
<tr>
<td>Potential alignment</td>
<td>Eliminate marginally relevant learning goals, identify essential parts of learning goals for content alignment.</td>
</tr>
<tr>
<td><strong>Content analysis</strong></td>
<td></td>
</tr>
<tr>
<td>Necessity</td>
<td>Determine whether a correct response requires knowledge of everything in the learning goal, only a part of it, or a simpler or more complex form of the knowledge.</td>
</tr>
<tr>
<td>Sufficiency</td>
<td>Determine whether a correct response requires knowledge not in the learning goal, or if an incorrect response reflects not knowing the learning goal.</td>
</tr>
<tr>
<td><strong>Likely effectiveness</strong></td>
<td></td>
</tr>
<tr>
<td>Comprehensibility</td>
<td>Determine whether the item is clearly worded, the drawings or figures are accurate, and the vocabulary is appropriate.</td>
</tr>
<tr>
<td>Clarity of expectations</td>
<td>Determine whether students know what they are expected to do, how to make choices, how long or detailed their response is supposed to be.</td>
</tr>
<tr>
<td>Context</td>
<td>Determine whether the item is equally interesting to all students, and if the context is equally familiar or unfamiliar.</td>
</tr>
<tr>
<td>Test-wiseness</td>
<td>Determine how susceptible the item is to guessing through choices that are unlike others, or if responses use special vocabulary.</td>
</tr>
<tr>
<td>Scoring guide</td>
<td>Determine whether the scoring guide is accurate, clear, complete and specific.</td>
</tr>
<tr>
<td><strong>Descriptive characteristics</strong></td>
<td></td>
</tr>
<tr>
<td>Cost effectiveness</td>
<td>Determine whether what we learn about students’ knowledge is worth the “cost” of the item in terms of time and effort.</td>
</tr>
<tr>
<td>Other characteristics</td>
<td>Determine whether the item possesses any other noteworthy characteristics.</td>
</tr>
</tbody>
</table>
In the view of test developers, content alignment is accomplished through constructing a content-by-behavior matrix. This approach has been used for many years by the NAEP, which is one of the most widely referenced assessments in mathematics and is noted for its careful design and development. The framework for the 1996 and 2000 NAEP mathematics assessments used a content-by-ability matrix with five areas of content (NAEP Mathematics Consensus Project, 1996). Content alignment takes place as developers select items from a pool by matching them with a list of objectives under each broad topic area such as “number” or “geometry.” Items are selected to assure appropriate coverage by the necessarily limited number of items on a test. Romberg, Zarinnia, and Collis (1990) have outlined the disadvantages of this model, pointing to the difficulty of classifying mathematics content into neat and distinct categories and the more important danger that such a model reinforces the idea that mathematics is a static and fragmented discipline. Nevertheless, the matrix approach to aligning assessment items with mathematics content has continued.

The reform movement in mathematics education (see NCTM, 1989, 2000; National Science Foundation, 1996) has brought about increased attention to the mathematics content that is communicated through national and state standards documents. The ensuing focus on testing and accountability has brought attention to the issue of content alignment, including the idea that test results should indicate what students know and are able to do, not just how they compare with other students or with a “national average.” Proponents of performance and constructed response assessments claim that these items broaden the kinds of information a test can represent. Wiggins (1993), for example, claimed that these items allow appropriate room to accommodate students’ learning styles, aptitudes, interests, and ultimately levels of conceptual understanding. Authentic assessment items designed to require particular knowledge or skill to obtain the solution can measure students’ achievement significantly better than a task that allows students to select a choice (Lu & Suen, 1995).

Performance and constructed response assessments also provide a means to pinpoint the areas in which students need additional academic instruction and inform policymakers and teachers about where they should direct their resources and energies (Rothman, 1995). If tests are to help teachers and others to improve instruction on specific mathematics concepts and skills, content alignment between learning goals and assessment items must be very specific and very precise.

**Likely Effectiveness**

Assessment is the primary tool that is used to gauge how students gain in academic achievement or how much value has been added to the youngsters by their schooling. Even well-aligned items may be less effective if they are unclear or do not engage students in their best efforts (Baxter & Glaser, 1997). Assessments can iden-
ify which content areas students are most familiar with and perform well in and which content areas students are least familiar with and do not perform well in (Resnick & Resnick, 1992). Familiarity with the context of an assessment item can affect performance, especially for low-performing students (Telese & Kulm, 1995).

RESEARCH METHODS

To illustrate application of the procedure and criteria and to provide examples of particular assessment items, we present analyses of two constructed response items. The procedure has parallel criteria for multiple-choice items and constructed response items, but, for constructed response items, there are extended criteria that take into account the complexity involved in judging possible student responses. The items we chose to analyze were (a) a fourth-grade released item of the 1996 NAEP and (b) a sixth-grade task from the Balanced Assessment (2000) materials. See Appendixes A and B for the items and scoring guides. It is worth noting that we did not consider the purposes for which these items were designed. Our analysis was not intended to provide information about consequential validity. Instead, for the sake of providing insight into the Project 2061 alignment procedure, we focus on how student data confirm or contradict what is learned about item alignment with a set of standards.

Pairs of analysts applied the Project 2061 alignment procedure to the items individually, then combined their results to reach consensus. The second and third authors, along with other analysts, participated in training sessions conducted by the first author on the Project 2061 alignment procedures. For standards we chose the Principles and Standards for School Mathematics (PSSM; NCTM, 2000) and the Benchmarks for Science Literacy (AAAS, 1993). After choosing potential standards that appeared to match the content of the assessment tasks, analysts applied the Clarification steps of the alignment procedure on each standard. This process resulted in selecting one standard that was most closely aligned with each item. Although we report briefly on the content alignment analysis for other closely aligned standards, the analysis of likely effectiveness for both items is based on the PSSM standard that proved to be most closely aligned with the item.

For the first item, entitled “Comparing Two Figures,” students were provided with two geometric figures (Appendix A). Students were asked to compare the figures to describe how the figures are alike and different. Not only did we have access to the performance results at the national level for this item but we also had a national random sample of 237 student papers. In addition, individual student interviews were conducted with 13 students from two geographically and demographically diverse areas. In the interviews, students completed the item, were asked to verbally explain their solution, and then were asked questions about how...
they knew when they were finished writing their solution (clarity of expectations) and whether they liked the item (appropriateness of context). We completed each interview by asking the students to describe the two figures on the phone to a friend who could not view the figures.

Once the alignment procedure was accomplished, we turned to the student data to determine whether the claims made within the procedure for that standard were supportable by the student data. Specifically, we examined the data from the written work and the student interviews to see whether the item actually elicited student thinking about the mathematics in the chosen content standard. The qualitative research was approached from an interpretivist qualitative perspective (Gall, Borg, & Gall, 1996; Glesne & Peshkin, 1992; Merriam, 1998). We then examined the technical qualities of the item to see whether the student data supported or refuted the claims made by the procedure.

A sample of 134 sixth-grade students who were using the Connected Mathematics curriculum (Lappan, Fey, Fitzgerald, Friel, & Phillips, 1998) were administered the second task, an open-ended item from the Balanced Assessment for the Mathematics Curriculum (2000) entitled the “Vet Club” (Appendix B). The item was selected as a tool to assess the pervasiveness and persistence of student misconceptions in statistical thinking (Kulm, Capraro, Capraro, Burghardt, & Ford, 2001). The item provided students with data about the type and number of pets owned by 14 students and asked them to construct a graph that would assist them in writing a headline for a news story about the typical number of pets owned. Students were then asked to explain their choice of the typical number.

To ensure accuracy in the application of the rubric for the Vet Club (see Appendix B), 10 independent scorers were trained to apply the scoring rubric. Each of the analysts functioned as the primary evaluator for 14 papers. Each paper was scored by the primary analyst and a record of each analyst’s scoring decisions was noted on a separate sheet that was matched to the paper by a preassigned identification number. After the primary analyst had scored a paper, it was placed into a single stack. After all primary analysts were finished scoring, the secondary scoring began. When all the scored papers were returned to a single stack, they were randomized. Each analyst took a turn selecting one paper that started the process over. If an analyst had scored the paper previously it was replaced into the stack. The process was repeated until each paper was scored three times. The purpose of this procedure was to ensure consistent application of the rubric across analysts. When analysts disagreed on a score, the trio entered a reconciliation phase where each member explained the reasons for assigning a particular point value. The discussion continued until consensus was achieved.

Interviews were conducted with 7 sixth graders to understand their reasoning and to explore difficulties they had with constructing graphs. The question protocol (Appendix C) was used for a quasi-scripted interview to determine how students solved and interpreted the item. The interviews were audio taped, allowing the investigator to comment as the students answered the guided questions and ex-
plained their solution strategies expressed on their papers (Glesne & Peshkin, 1992). The use of the tape recorder provided a more detailed record of the interview while providing a more relaxed environment for the participant and investigator (Merriam, 1998). All of the audio taped interviews were later transcribed. Student-generated work was also collected for analysis. As with the first item, the interview data was evaluated from an interpretivist qualitative perspective (Gall et al., 1996; Glesne & Peshkin, 1992; Merriam, 1998).

ANALYSIS OF THE ASSESSMENT ITEMS

Alignment

The first step in the alignment process is to select learning standards or benchmarks that appear to align most closely with the learning objectives of the item. With regards to Comparing Two Figures, the following four standards were selected because each appeared to contain elements of the learning goals inherent in the fourth-grade item:

2. Standard from Standards 2000 (PSSM 2). Geometry Standard for Grades 3 to 5 (NCTM, 2000, page 164): Classify two- and three-dimensional shapes according to their properties and develop definitions of classes of shapes such as triangles and pyramids.
4. Benchmark from Benchmarks for Science Literacy (Benchmarks). (AAAS, 1993, page 223): Many objects can be described in terms of simple plane figures and solids. Shapes can be compared in terms of concepts such as parallel and perpendicular, congruence and similarity, and symmetry. Symmetry can be found by reflection, turns, or slides.

For the Vet Club task, the following three standards were selected because each appeared to contain elements of the learning goals inherent in the sixth-grade Balanced Assessment item:

use appropriate graphical representations of data, including histograms, box plots, and scatterplots.


For each item, these standards were then clarified, which illuminated how they differ from each other and how they are connected to learning standards that extend at higher or lower levels in the PSSM and Benchmarks documents. For an analysis of the alignment between each of the four standards and Comparing Two Figures, using criteria specified in the procedure, see Wilson and Kitchen (2002).

To determine the potential alignment of the three standards selected with the Vet Club, analysts read essays in the PSSM (NCTM, 2000) to clarify the intent of each standard, as well as research on student learning of the targeted mathematical skills and concepts. They answered the following two questions:

1. Is the targeted knowledge and skill in the standard involved in the response strategy?
2. Do any specific misconceptions contribute to an unsatisfactory response?

The answer to both questions for the Vet Club was “yes” for each of the three standards. For the purposes of this discussion, we decided to provide an analysis of the item’s alignment and effectiveness in relation only to the learning goal PSSM3: Select, create, and use appropriate graphical representations of data, including histograms, box plots, and scatterplots.

The analysis clearly demonstrated that PSSM1 is the most closely aligned to Comparing Two Figures. In our analysis of the substance of the item, we determined that knowledge of PSSM 1 is necessary to accomplish the item because students must identify, analyze, and compare attributes of the two shapes to determine likenesses and differences. With regards to the other three standards analyzed, we found that knowledge of PSSM 2 is not necessary, because students do not need to classify shapes according to their properties or develop definitions of classes of shapes to correctly respond to the item. Knowledge of PSSM Comm and the standard from the Benchmarks (AAAS, 1993) could contribute to a correct response to Comparing Two Figures but is not necessary for a correct response.

While PSSM 1 was determined to be most closely aligned with Comparing Two Figures, we also determined that the alignment was partial. Although students
must be able to identify, analyze and compare attributes of the two shapes, it is not necessary for students to use mathematical vocabulary to make a satisfactory response. Therefore we determined that knowledge of PSSM 1 was sufficient, but not totally necessary, for a satisfactory response to the item. The first portion of the standard was deemed to be necessary and sufficient but the second part (about using mathematical vocabulary) was not necessary. This analysis led us to design our review of student papers and the student interviews to examine each of these aspects of the standard.

We investigated PSSM1 in two parts: (a) whether students identified, compared, and analyzed the attributes of the shapes in the item and (b) whether students used geometric vocabulary to describe the attributes of the shapes. The data from the student papers and interviews (see Wilson & Kitchen, 2002, for the examination of the student papers, the student interview protocol, and the results from the interviews) showed that the majority of students provided evidence of thinking about attributes of the figures (64% of the student papers) but most (70% of the student papers) did not use geometric vocabulary correctly in their responses. We ascribe students’ lack of use of geometric vocabulary to the item’s failure to require the use of geometric vocabulary and to students’ poor understanding of measurement concepts (particularly, perimeter and congruence).

When students did incorporate correct geometric vocabulary in their responses, it was not uncommon that the only geometric word used was “squares.” For example, in the student interviews, one student wrote that “they [the figures] are both five squares across and three squares down,” another wrote “the two shapes both have 15 complete squares [sic],” and a third wrote “they both have 3 squares [sic] in each column.” Two students misunderstood the shapes to be squares. One of these students discussed how the shapes looked like squares and the other wrote that the “squares are the same but 1 is tilted a little.”

Rarely in the student interviews or in the student work did students identify the first figure as a rectangle and the second as a parallelogram. In addition, students infrequently made any references to the opposite sides in the two figures being parallel nor did they make many references to the adjacent sides in the rectangle being perpendicular while in the parallelogram adjacent sides are not perpendicular. Thus the examination of student papers and the results from the interviews confirmed the results of the alignment procedure. The part that was aligned was evident in student work, while the part that was not aligned was not evident in student work.

The analysis of the alignment of PSSM 3 with the Vet Club found that the alignment was partial. Although a correct answer to the three parts of the item requires knowledge of selecting, creating and using an appropriate graphical representation of the data provided, knowledge of and the ability to apply two other standards are needed for a complete response: finding and using measures of central tendency and discussing and understanding the relationship between data and graphs. All
satisfactory responses to the item require the use of PSSM3 but students would still need to apply additional learning goals. Therefore we determined that knowledge of PSSM3 was necessary, but not sufficient, for a satisfactory response to the item. This analysis led us to design our review of student papers to examine the depth of students’ knowledge and ability to apply PSSM3. In particular, a review is provided of student work as a means to investigate students’ ability to select, create, and use appropriate graphical representations of the data provided in the Vet Club.

Sixty-seven percent of the sixth-grade students who used a bar graph actually drew a histogram or called the bar graph a histogram. Students were unable to differentiate between bar graphs and histograms even though being able to make such distinctions was part of the state curriculum. Students believed that because histograms were pushed close together it was easier to make comparisons. Students believed that “the bars should always be pushed together and that leaving a space was incorrect or done to confuse students.” The students who drew a bar graph and called it a histogram believed that it was the “correct mathematical term.” They believed that it was like learning that “a square was square and rectangle was a rectangle but then learned that not really, a square was also a rectangle but not all rectangles were squares.”

Two of the students credited their teacher last year with telling them that “bar graph” was the correct name for their graphs and one thought she saw it in a book. Only one of the students who chose a histogram to represent the data also graphed zero. In general, students seemed to possess an unrefined knowledge of bar graphs and histograms. Yet, they were able to convey a graphical message and interpret the data. It is this aspect of the learning process that we term “naïve conception.” This is a conception that is necessary at some particular developmental level but that must be expanded on to enhance students’ depth of understanding of a mathematical idea. That is, deep understanding of mathematical ideas requires that the learner have opportunities to continually enhance the depth of understanding of concepts. In this example, students held a naïve conception about the difference between graphical representations of histograms and bar graphs.

The data summarized in Table 2 indicate that only 46% of the sixth-grade students who represented their data using a bar graph believed they needed to represent zero as a data point (see Appendix D for an example of student work). For students, the implication here was that zero appears above the X-axis and on the Y-axis. If students were representing zero as a categorical variable then it would and should be located below the X-axis where the quantity (Y-axis) should still represent zero at the origin. Students explained that it did not seem correct to leave a categorical variable without some representation. One student commented, “Four kids don’t have pets and it doesn’t seem right to leave them blank.” When prompted to explain what it would mean if there was a student without anything represented above the X-axis; the student responded, “It would mean that the kid did not answer the question.”
Of the 21 sixth-grade students who represented the data with a circle graph only 35%, respectively, believed that zero should be represented on the graph. Forty-seven percent of the sixth-grade students who drew a bar graph placed the quantity on the X-axis and the categories on the Y-axis. Students who made this error also believed that each type of pet or person should be represented as a categorical variable.

Another student also demonstrated a misconception about the meaning of the origin. When prompted to explain the meaning of the intersection of the Y and X axes, the student indicated that it “was the place where the two lines meet.” The student was unaware of anything special about the intersection and was not able to identify the intersection when it was referred to as the origin. The student’s lack of knowledge did not seem to interfere with the ability to graph. However, the lack of sophistication supported the idea that a naïve conception existed when dealing with the origin and in the representation of data graphically. The predominate interpretation of the X-axis, common to most students who represented the data incorrectly was that “the X-axis is where you begin and all the data goes above that line.”

Technical Qualities

The alignment procedure showed us that Comparing Two Figures seemed to have three areas of strength and three areas of weakness. Specifically, the scores were high on comprehensibility, appropriateness of context, and resistance to test-wiseness, but low on expectations, adequacy of scoring guide, and cost effectiveness (see Wilson & Kitchen, 2002). That is, we would expect that students would not have difficulty comprehending the item and that they would have a difficult time “guessing” the right answer. We would also expect that students would find the item relatively interesting. In addition, the analysis showed that there were
potential difficulties with expectations. For instance, it might not be clear to students what would constitute a satisfactory response to Comparing Two Figures or even how to know when they were finished with the item. In addition, the scoring guide for the item is flawed, in that it does not assist the scorer to interpret some potential student responses, because students may identify attributes of the shapes using nonmathematical vocabulary. Last, it was determined that Comparing Two Figures lacked cost effectiveness, in that it does not provide sufficient information about students’ knowledge of geometry that is worth the time and effort the item demands of students and scorers.

To determine whether these technical qualities were borne out in the experiences of students, we examined the data from the student interviews. In our protocol for the interviews, we looked for evidence of each of these four technical qualities. Our results showed that the conclusions reached from the alignment procedure were supported by the data from the student interviews.

Of the 13 students who participated in the interviews, only one seemed to have some difficulty comprehending the item. This particular student had a difficult time perceiving that there were actually two figures on the grid. He seemed to focus his attention on the trapezoidal-shaped “space” between the rectangle and the parallelogram, describing that shape as “the keel of a boat.” Although he was directed toward the two shapes, he was unsuccessful at attending to particular attributes, likenesses or differences, of those shapes. This student’s difficulty was not necessarily with the item prompts but with the visual presentation of the figures on the page. The other 12 students, however, had no difficulty understanding what the prompt meant and all proceeded without any further assistance from the researcher. It is important to note that two of the participating students were English Language Learners. Neither of these two students had difficulties comprehending the item, either.

All of the students were enthusiastic about solving Comparing Two Figures and expressed this when asked. Most said that they thought other students would also find it interesting. When asked why they found it interesting, some students said it was because the item was “easy,” while others said because it was “hard,” but that that made it fun.

In the interviews we asked the students when they knew to stop working on the item. Nearly all of them responded that they stopped when they could no longer think of anything to write. One researcher observed that the students also seemed to be influenced by the amount of space provided on the page for the answer and were satisfied with stopping when the space for each question was filled up. One boy seemed intent on writing the same number of likenesses as differences. Clearly, students were unsure about how much was expected of them to score high on the item.

In summary, the information from the student interviews supported the data obtained from the alignment procedure regarding the technical qualities of the item.
What we did learn was more detailed information about students’ responses to those qualities. Interestingly, when we gave students the opportunity to respond to the item in an alternative way (verbally) some of them demonstrated knowledge of likenesses and differences between the figures that they had not presented in their written solution.

The alignment procedure showed us that the Vet Club had two areas of strength and two areas of weakness. The item is engaging and is not susceptible to guessing or bluffing knowledge. However, the item seems to lack clarity in the way questions connect the graph with the headline and with communicating expectations. Thus, we would expect that students might have difficulty comprehending the item but that they would have a difficult time “guessing” the right answer. We would also expect that students would find the item relatively interesting.

As with the first item, to determine whether the technical qualities identified were confirmed by how students actually solved the item, we examined the data from the student work and interviews. Also similar to work done with the initial item, in our protocol for the interviews, we looked for evidence of each of these four technical qualities. Our results showed that the conclusions reached from the alignment procedure were supported by the data.

The Vet Club appeared to lack clarity particularly in that students did not make the connection between analyzing their graph and writing the headline. When asked to use their graph to answer a question about the number of pets in the “typical” family, many sixth graders failed to interpret the graph. Laura was the only student who made an explicit connection between her headline, what she said, and her drawn graph: “They want you to figure out how … many pets the majority of her friends have and I put one because on the graph, one pet is the highest, the bar is highest.” Four of the seven students who were interviewed accurately displayed data on their graph; however, they did not make connections and construct a relationship between their graphing and the newspaper headline. Even those who constructed the graph correctly often did not interpret the typical number of pets from the graph. Hank expressed his ideas about displaying data by constructing a line graph and a bar graph with accurate information. However, he resorted to using an algorithm to find the average but did not specifically describe which numbers he was adding or dividing.

Some students used their own experiences, informal methods, or calculated the mean. For example, Jack used the information on his graph but needed to think about his ideas by reflecting on real world experiences. This was how he interpreted his answer: “The reason I chose this answer was because most kids had one pet, um, they, um, I think a lot of people have pets so there is a bigger chance of someone having one pet than none.”

Those who did not categorize the data and instead drew a separate bar for each person’s number of pets had even more difficulty in using the graph to help answer the question about typical pet owners. Finally, none of the students, including
those who did graph zero as a number of pets, used zero as typical, even though there were the same number of families with zero pets as with one pet. In summary, the majority of students made little connection between the task of constructing the graph and using it to answer a question.

**DISCUSSION**

The Project 2061 alignment procedure was effective in providing a tool for in-depth analysis of the mathematical content of the item and a set of standards, and in identifying one particular content standard that was most closely aligned with the standard. During the process of clarifying Comparing Two Figures and the standards selected, both analysts found that a standard from the PSSM (NCTM, 2000) most closely aligned with the item. This was borne out in the analysis process. The procedure was also effective at evaluating the technical qualities of the item. A flaw of Comparing Two Figures that was brought to light by the “clarity of expectations” criterion in the procedure is that it is not clear to the students what constitutes success on the item, because they are not given access to the scoring rubric. In addition to other technical strengths and the weaknesses, the students substantiated this during the student interviews. We were also able to identify which portions of the standard were necessary and sufficient for completion of the item, and which ones were not. This was confirmed by our analysis of the student papers, in which 64% of the students demonstrated evidence of thinking about the attributes, although a majority (70%) did not correctly use geometric vocabulary in their responses.

Through analyzing student work samples and student interviews, we found that students’ thinking may not correspond to the standard identified as best aligned with the larger goals of the item. The quality of students’ thinking demonstrated by their responses may not match the quality of the learning goal intended by the item. In this case, when responding to the fourth-grade item, Comparing Two Figures, the majority of students did not use geometric vocabulary correctly. Furthermore, the scoring rubric for the item did not discriminate based on students’ use of vocabulary. Students could use nongeometric vocabulary when responding to the item and not be penalized during the process. This finding highlights the importance of analyzing student work when using the AAAS Project 2061 alignment procedure to clarify any additional deficiencies of the item not revealed by the procedure.

The “cost effectiveness” criterion of the alignment procedure proved especially valuable given that the flaws of the item are sufficiently serious that major consideration would have to be given to the use of this item for the assessment of this standard. According to the procedure, Comparing Two Figures is at least partially aligned to the standard but these serious flaws affect the essential knowledge gained about student learning to such an extent that the results of the alignment
procedure might look skewed to the casual observer. Cost effectiveness is an essential issue and deserves a prominent role in the alignment procedure.

In responding correctly to all parts of the Vet Club item, students must have the knowledge indicated by the selected learning goal, PSSM3. However, this knowledge is clearly not sufficient. This item requires knowledge of two other standards. Furthermore, the single question used in the study elicited several difficulties and misconceptions in statistical thinking. The notion that a particular type of graph is needed to best answer a question about typical data was poorly understood. Most students seemed to construct the graph that was most familiar to them or their favorite, with little notion of its use. This naïve conception probably was a result of previous experience in which constructing a graph was simply an exercise in which the graph itself was the end product. Moreover, students are often provided with a graph and asked a low-level interpretation question that simply requires reading data from the graph. Students were able to answer questions concerning a graph when the information was directly displayed in the graph. However, when asked to describe the typical number of pets, some sixth-grade students did not possess the concepts necessary to interpret their constructed graphs. Some students did not even have the ability to construct the graph correctly. On the positive side, when the sixth-grade students were interviewed later, many of them noted that they had made mistakes in their graphs or reasoning. They were able, sometimes without prompting, to recognize their mistakes or misconceptions.

**FINAL REMARKS**

The AAAS Project 2061 analysis procedure involves the reviewer in a careful study of the intended learning goals and illustrates how much or little a single item might align with the learning goal. Typically, a multiple-choice item assesses only a part of a standard. A constructed response item, especially one that has several parts, often requires a combination of parts of several learning goals. For this reason, although it is important to analyze individual assessment items, it is also important to understand how limited most items are in meeting the necessity and sufficiency criteria that we usually infer from items.

While clarifying the standards selected and the items themselves, the authors found that a standard from the PSSM (NCTM, 2000) most closely aligned with each item. This was borne out in the analysis process. Our analyses showed three strengths in the procedure. First, it is thorough, thus avoiding the errors or omissions that might result from more superficial alignment procedures. Second, the cost-effectiveness criterion can serve as a “bottom line” measure when decisions are made about whether a given constructed response item is worth the time and money required to administer and score it. Third, the alignment criteria of “necessary and sufficient” provide a complete picture of the match or mismatch between
any given item and standard, information that is vital when considering revisions either to items or to standards, as well as when making inferences based on that item and others.

The AAAS Project 2061 alignment procedure is a labor-intensive process but it is a powerful tool for analyzing the alignment of standards with singleton items. Taking the time to do such an analysis for multiple items on large-scale assessments may prove overwhelming to the faint of heart. Yet for small-scale assessments, or for single extended constructed response items, this procedure may prove to be quite valuable. Taking the extra time to analyze multiple pieces of student work and conducting individual interviews with students may also reveal further information about the qualities of an item that the procedure by itself may not reveal. Our analyses revealed flaws in the items that the procedure alone did not show. Such flaws may become visible by examining student work (beyond what is available in a scoring packet) or conducting student interviews.

REFERENCES


APPENDIX A
FOURTH-GRADE NAEP TASK ENTITLED “COMPARING TWO FIGURES”

Think carefully about the following question. Write a complete answer. You may use drawings, words, and numbers to explain your answer. Be sure to show all of your work.

In what ways are the figures above alike? List as many ways as you can.

In what ways are the figures above different? List as many ways as you can.
“COMPARING TWO FIGURES” SCORING GUIDE

Solution:

The figures are alike because:

   a. They both have 4 sides (or 4 corners or 4 angles).
   b. They both have parallel sides.
   c. They both have two sets of sides that are the same length.
   d. They have the same area.
   e. They have the same length (Base).
   f. They have the same height.
   g. They both have little squares.

The figures are different because: (Do not accept “They’re not both the same shape.”)

   h. One has 4 equal angles and the other does not.
   i. One has right angles or perpendicular lines and the other does not. (Students don’t need to make the comparison, i.e., they can just say, “one has 4 equal angles”.)
   j. One is “slantier” than the other.
   k. They have different perimeters.

Scoring Guide:

1. Incorrect response
2. A nonspecific response, i.e., “the one on the right is skinnier”  
   OR  
   Only one correct reason (alike or different).
3. Student gives one correct reason alike and one correct reason different  
   OR  
   Two reasons alike  
   OR  
   Two reasons different.
4. Student gives two reasons why the figures are alike and one reason why they are different  
   OR  
   One reason why they are alike and two reasons why they are different.
5. Student gives at least two reasons about why they are alike and at least two reasons about why they are different. (Two alike reasons are not both “a”.)
APPENDIX B
SIXTH-GRADE BALANCED ASSESSMENT TASK
ENTITLED “VET CLUB”

Jenny is writing a newsletter article about the members of the Future Veterinarians Club. She asked each of the members: “What pets live in your house?” Her notes look like this:

<p>| | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>R. J.</td>
<td>none</td>
<td></td>
</tr>
<tr>
<td>A. P.</td>
<td>1 dog and 1 cat</td>
<td>1 dog</td>
</tr>
<tr>
<td>K. K.</td>
<td>2 dogs</td>
<td></td>
</tr>
<tr>
<td>R. B.</td>
<td>1 bird, 1 dog, and 4 goldfish</td>
<td>2 dogs, 1 bird, and 2 fish</td>
</tr>
<tr>
<td>G. L.</td>
<td>1 dog</td>
<td></td>
</tr>
<tr>
<td>W. R.</td>
<td>none</td>
<td>1 bird</td>
</tr>
<tr>
<td>J. M.</td>
<td>1 dog</td>
<td></td>
</tr>
<tr>
<td>J. Z.</td>
<td>none</td>
<td>3 cats</td>
</tr>
<tr>
<td>S. R.</td>
<td>3 cats</td>
<td></td>
</tr>
<tr>
<td>E. R.</td>
<td>none</td>
<td></td>
</tr>
<tr>
<td>K. B.</td>
<td>1 bird</td>
<td></td>
</tr>
<tr>
<td>S. L.</td>
<td>2 dogs, 1 bird, and 2 fish</td>
<td></td>
</tr>
<tr>
<td>D. L.</td>
<td>8 fish</td>
<td></td>
</tr>
</tbody>
</table>

From Balanced Assessment for the Mathematics Curriculum Middle Grades Assessment Package, ©2000 by the Regents of the University of California. Published by Pearson Education, Inc. publishing as Dale Seymour Publications, an imprint of Peason Learning Group. Used by permission.
“Vet Club” Scoring Rubric

1. Your job is to prepare a graph to go with Jenny’s article. Organize the information from her notes into a graph that will show how many of her friends have no pets, one pet, two pets, and so on.

| 1 point | Draws a bar or line graph (What if the answer can not be read from the Chart?) |
| 1 point | The graph shows number on the vertical axis; categories on the horizontal (may not be labeled, but the data is shown this way) (Title of the Graph?) |
| 1 point | The axes are labeled correctly |
| 1 point | The number of pets (0, 1, 2, 3, etc.) is used as the category for the horizontal axis |
| 1 point | All of the values are correctly shown in the graph [i.e., (0,4), (1,4), (2,2), (3,1), (4,0), (5,1), (6,2)] |

Total possible: 5 points

2. Jenny plans to title the article: “Typical Future Veterinarians Club Member Has __ House Pets”. What number should Jenny put in the blank? _____________

| 2 points | 0, 1 or 2 (This number should be related to the graph or substantially justified and explained) |

Total possible: 2 points

3. Explain why the number you chose is the best number to complete the headline.

| 1 point | The explanation is connected to the data in some reasonable way (not just an opinion or life experience, for example) |
| 1 point | The explanation and the answer to Question 2 are connected |
| 1 point | The explanation refers to a measure of central tendency |

Total possible: 3 points

APPENDIX C
STUDENT INTERVIEW PROTOCOL FOR SIXTH-GRADE BALANCED ASSESSMENT

1. Explain the steps you took in solving the problem.
2. Is there a way to get the correct answer without working out the problem?
3. How did you decide to use this graph?
4. How did the graph help you to decide what number should go in the blank?
5. What mathematics did you use to get your answer?
6. What steps did you use to come up with your answer?
7. How did you know what to write down for your answer if you did not get it from the graph?
8. What did you learn (or do you know) in math that helped you solve this problem?
9. What part of the problem was hard for you?
10. What would another student doing this problem need to know to solve it?
11. What do you think is meant by typical?
12. Can you think of another way to solve the problem that would give another answer?

APPENDIX D
VET CLUB—EXAMPLE OF STUDENT WORK