

## Evaluating Assessments Using Standards and Quality Criteria



Presented at the annual meeting of the American Educational Research Association, New Orleans, April 28, 2000

Gerald Kulm

Curtis D. Robert Professor

Texas A&M University

The AAAS document *Blueprints for Reform: Science, Mathematics, and Technology Education* (1999) recommends that in addition to being aligned with standards, classroom assessment should include a variety of techniques, encourage students to go beyond simple recall of data or facts, close the gap between the classroom and the real world, and include opportunities for students to perform tasks and solve problems. Webb (1997) has proposed a model of vertical and horizontal alignment of assessment systems that includes the alignment of expectations (standards or benchmarks) with assessments. He suggests that alignment criteria include content focus, articulation across grades and ages, equity and fairness, pedagogical implications, and system applicability.

Other important characteristics of meaningful assessment are described in national standards documents produced by the NCTM and NRC and in state blueprints for assessment (e.g., Michigan, Massachusetts, Connecticut). Finally, other organizations and groups have proposed principles and standards for assessment and alignment (e.g., FairTest, 1998; Achieve, 1998).

But classroom teachers, administrators, and test developers who are required to choose, assemble, or develop assessments have little to guide them. There is no useful synthesis of the latest thinking on assessment or practical advice on how to align assessment with standards and benchmarks. Moreover, claims of alignment can be very superficial. It is critical that tests' claims of alignment with national and state benchmarks and standards be carefully checked. Educators need a clear set of criteria for alignment and a reliable means of applying those criteria to actual assessment tasks.

### Developing an Alignment Procedure

In order to determine whether an assessment task is aligned with mathematical ideas contained in a standard, two essential questions should be addressed. First, does the content of the task match the content of the standard? Second, does the format of the task reflect the intention and expectations of instruction? Project 2061 of the American Association for the Advancement of Science is developing a procedure that can be used to analyze K-12 mathematics and science assessments for (1) alignment with a set of benchmarks or standards, and (2) the quality of their presentation (AAAS, 1998).

\*This paper is based on the work of the AAAS/Project 2061's "Resources for Science and Mathematics Literacy: Assessment" project, funded by NSF Grant #ESI-9819018. The project is co-directed by Gerald Kulm and Andrew Ahlgren, assisted by Ryan Arndt, whose contributions are gratefully acknowledged. All opinions and conclusions contained in the paper, however, are solely the author's and not those of NSF, AAAS, or other persons.

The criteria used in the procedure should provide a detailed and comprehensive approach to determining the specific mathematics or science content that is assessed by tasks and instruments, producing a profile of the quality of format of a task. The term "standards" is used generically to denote the learning goals

used as content criteria for analysis. In practice, the standards could be statements from the NCTM *Standards* (1989, 1998), AAAS *Benchmarks* (1993), the *National Science Education Standards* (NRC, 1996), the National Assessment of Educational Progress Framework, or a state or local mathematics or science framework. The following sections describe the current progress of Project 2061's work on developing and refining the procedure. The procedure will be finalized in the coming months, then applied to a range of assessment tasks, producing a set of prototype descriptions, profiles, and cases of mathematics and science assessments from international, national, state, and standardized tests.

## **The Analysis Procedure**

The analysis consists of four stages. First, a Preliminary Analysis identifies the standards that are apparently addressed by each task. An important step in this analysis is to clarify the meaning and intent of the standards or benchmarks that the task addresses. Following this preliminary step, a "Triage" is performed, separating tasks into those that merit further analysis, those that are promising with some revisions, and those that are discarded because there is either no content match and/or their quality is too poor to attempt fixing. The core of the work is to perform a complete analysis of Content Alignment and Technical Quality. An essential task in the analysis is a careful clarification not only of the content standards but also clarification of the assessment task. As the work on a task proceeds, analysts return repeatedly to these clarifications, refining them so the task and the standard can be compared and aligned. The Content and Quality alignment is done for each assessment task, using a set of criteria and indicators to guide the analysis. Finally, a summary report and profile is constructed that describes the characteristics of the task. It should be noted that the procedure considers individual assessment tasks, or closely related sub-questions or items, not an entire test. Psychometric properties of tests are beyond the scope and intent of the procedure. Information, including difficulty indices, and sample responses on the individual tasks, however, is important and critical to doing the analysis.

## **Preliminary Analysis**

The goal of this phase is to select a set of standards that are credibly addressed by an assessment task. A list is made of standards that *at first glance* seem relevant to a task, and then this "suspected" list is narrowed to a smaller "central" list of standards that will be pursued through a full Content Analysis. The entire set of standards is searched for specific statements that seem to be targeted by the-task, as far as that can be inferred from what the student sees, the administration instructions, the answer key or scoring guide, and sample student responses. The product of this step is a list of "suspected standards" that the task might be intended to target.

Having completed the Preliminary Inspection, the reviewer now has at hand two sets of information: (a) a complete statement of the assessment task, along with all of the relevant supporting material (i.e., scoring guide, sample student responses, administration guide, etc.) and (b) a list of suspected standards, along with brief justifications of why the task addresses each standards on the list.

## **Content Analysis**

The next step is to do a Content Analysis which involves clarifying the meaning of both the standard and the task, then applying the Content Criteria to judge the extent and quality of the alignment of the two. The Content Analysis consists of the following steps:

### **1. Clarification of the Standard**

It is essential to (a) specify exactly what mathematical ideas are contained in the standard, (b) clarify the meaning of these ideas by referring to other standards statements and documents, and (c) identify student cognitive development, difficulties, or misconceptions about the ideas, using research findings or teacher experience as a guide. Figure 1 provides a brief sample clarification for a geometry standard taken from *Benchmarks for Science Literacy* (AAAS, 1993).

## Figure 1. Sample Standards Clarification

### **Standard:**

Students should know that:

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 (AAA Benchmark 9C#4, grades 3-5).

### **Clarification of Standard Ideas:**

Idea 1: The term "object" can mean both real, physical objects and mathematical objects such as geometric figures themselves. Ways to describe the objects can include verbal descriptions, drawings or sketches, or physical models.

Idea 2: Shapes can include both two and three dimensional shapes. The concepts following the phrase "such as" are not inclusive of all possible concepts for comparing shapes, but ought to be included. The terminology itself is less important than understanding the concepts and being able to use them in making comparisons.

Idea 3: The phrase "can be found" includes the expectation of being able to describe the nature of a symmetry with one of these three concepts and being able to identify and produce a symmetric figure that involves one of the three.

The corresponding NCTM Standards 2000 (draft) for the benchmark are statements under the standards for grades 3-5:

- Analyze characteristics and properties of two and three dimensional geometric objects
- Recognize the usefulness of transformations and symmetry in analyzing mathematical situations. According to NCTM Standards, describing involves "learning to use mathematical terminology by hearing it used repeatedly in context." In comparing, students tell what the characteristics are of each shape and how they are different.

### **Student Cognition and Performance:**

According to Van Hiele levels, students (1) identify shapes with concrete examples, the (2) become able to use properties to identify and describe shapes. These levels, progress in them, and what students understand about shapes is highly influenced by instruction, which is lacking in the early grades. Given appropriate instruction, students can understand abstract properties of geometric figures by 5th grade (Clements & Battista, 1992). Students expand their notion of symmetry by 5th grade using more than one line of symmetry and describing rotational symmetry more precisely with angle measures.

## **2. Task Triage**

Based on the clarification of the standards, the analyst sorts the task into one of three categories, deciding whether to: (1) pursue the task further, if it appears to address at least one standards idea, (2) retain the task as "fixable," if it appears to have good technical quality but slightly off the mark in addressing a standards idea, or (3) discard the task because it does not appear to address a standards idea and/or does not have good technical quality.

Standards that are addressed only in part by the task; that is, only one or two of several ideas from the standards, can still be included. It is also legitimate to include standards at a grade level that is different

from the grades specified for the task. There are no guidelines or expectations for the number of standards ideas addressed by a task. An extended or multiple-part assessment task might address as much as several standards, or as little as only part of one.

It is important to recognize that the decision about alignment still remains tentative at this stage. If the task is judged to be pursuable, the analysis continues to the next step, in which the list of standards that the task actually addresses is honed further. Still closer inspection or discussion of what the standards intend or what the task assesses often changes opinions about whether the substance of standards ideas are addressed by a task.

### **3. Content Alignment Analysis**

A set of five criteria has been developed to guide the Content Analysis. In practice a detailed clarification, along with examples of applying the criteria to assessment tasks is provided for the persons doing the analysis. These criteria are summarized in Figure 2, providing the key question to be answered by analysts.

Figure 2. Criteria for Mathematics and Science Assessment Content Analysis

#### **CONTENT ALIGNMENT CRITERIA (Draft)**

Criterion CA-1 Substance-. Does the task address the specific substance of the standard or is there only a “topic” match?

Criterion CA-2 Sophistication. Does the task reflect the level of sophistication of the standard or does it target a standard at an earlier or later grade level?

Criterion CA-3 Cognitive Demand. Does the knowledge type of the task match that intended or implied by the standard?

Criterion CA-4 Beyond Standards. Does the task assess content that is not required for achieving the standards?

Criterion CA-5 Content and Context. Is the assessment context of the task consistent with the content intended by the standard?

For each of the criteria, a set of 3 to 5 indicators is used to guide the analyst in making judgements about how well a task meets the criterion. For example,

Criterion CA-1 Substance-. Does the task address the specific substance of the standard or is there only a “topic” match?

#### ***Indicators:***

1. The standard idea essential (necessary) to respond to the task satisfactorily.
2. The task can't be responded to satisfactorily through general intelligence or test-wiseness, without specific knowledge of the standard idea.
3. A student cannot use some other way to respond to the task, for example, by using knowledge of another standard.

4. The task can't be answered by mere memorization of an idea or operation that doesn't require understanding the standard idea.

The analyst decides whether or not each of the indicators is met by the task, making notes and justifications on each one. Later, these notes and evidence for meeting the indicators become the source for writing a report of the analysis, summarizing and constructing a profile of content alignment and technical quality for the assessment task.

### **Technical Quality Analysis**

A set of five criteria has also been developed to guide the Technical Quality Analysis. Detailed clarification statements and examples of applying the criteria to assessment tasks is also provided for these criteria. The criteria are summarized in Figure 3, providing the key question to be answered by analysts.

Figure 3. Criteria for Mathematics and Science Assessment Technical Quality

#### **TECHNICAL QUALITY CRITERIA (Draft)**

**Criterion TQ-1 Comprehensibility.** Is the task (including diagrams and symbols it uses) likely to be familiar and comprehensible to the intended students?

**Criterion TQ-2 Engagement.** Is the task likely to be motivating and engaging for the intended students?

**Criterion TQ-3 Clarity.** Does the task and/or directions make clear to the students what they are expected to do and what constitutes success?

**Criterion TQ-4 Commonly Held Ideas.** Does the task anticipate students' commonly held ideas so that incorrect answers have implication for interpreting misconceptions or other important errors?

**Criterion TQ-5 Alternative Responses.** Do students have the opportunity to demonstrate their knowledge or skill in alternative ways?

The following example illustrates the indicators that are used to guide the judgment of how well an assessment task meets criterion TQ-4.

**Criterion TQ-4 Commonly Held Ideas.** Does the task anticipate students' commonly held ideas so that incorrect answers have implication for interpreting misconceptions or other important errors?

#### ***Indicators:***

1. The task takes advantage of opportunities to elicit or probe for students' commonly held ideas.
2. The commonly held ideas elicited are specifically relevant to the standard (rather than just to related difficult or complex ideas).
3. The task's use of commonly held ideas accurately reflects research.

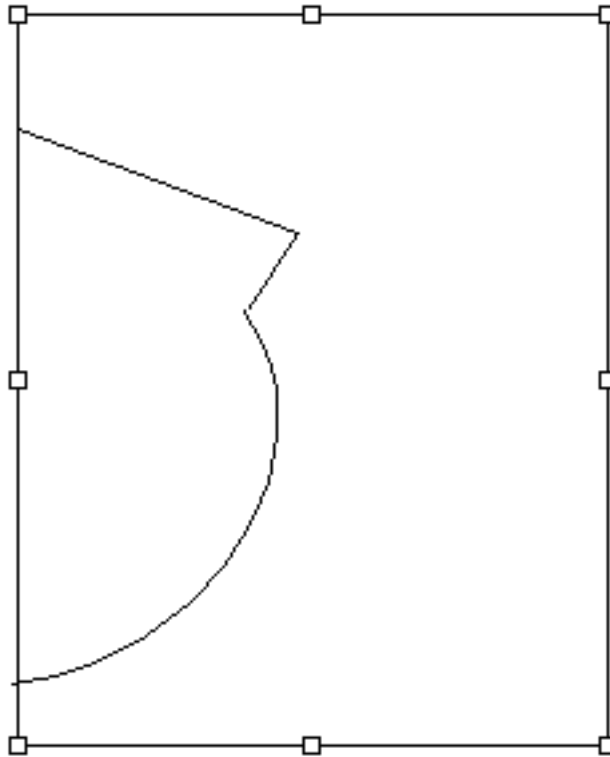
In addition to using these indicators as a guide to preparing a summary of the quality of the task, they can be used to help revise or improve the quality of a task, if that seems warranted or necessary.

## Sample Assessment Task

The following task addresses one or more of the ideas in the benchmark that was clarified on page 3. In order to do an analysis, complete information about the task is required, including the conditions under which it is administered, the scoring rubric, and sample student responses. For this illustration, we use only the item and the scoring guide. The scoring guide does provide some information about types of student responses but not the distribution of these. Internationally, 59 percent of 4<sup>th</sup> graders and 45 percent of 3<sup>rd</sup> graders answered the task correctly. [See the Appendix for a full statement of the item and the scoring guide.]

Example: Folding Symmetry (TIMSS, grade 4)

Craig folded a piece of paper in half and cut out a shape.



Draw a picture to show what the cut-out shape will look like when it is opened up and flattened out.

## Content Analysis Example

The following analysis summaries provide brief examples of the results of content alignment work. In practice, these summaries would be accompanied by analysts' ratings of each indicator and detailed evidence for whether these indicators are met. Summary statements such as those below are written for a quick reading of the alignment profile.

CA-1: Substance- The task most closely addresses only the part of benchmark related to symmetry through reflection. Drawing the shape requires the student to understand properties of reflection across a line, even though the term 'reflection' isn't used by the task. On the other hand, an inaccurate drawing, which is scored as correct may not reveal a full understanding of the properties of reflection. Also, without a requirement for an explanation, students may be able to produce a satisfactory drawing without a full understanding of symmetry or reflection.

CA-2: Sophistication The requirement to draw the complete figure does address the level of sophistication expected by the standard. The task seems to fit squarely within the grade 3-5 expectations. However, the difficulty as indicated by the percent responding correctly may indicate that the task is at the high end of sophistication for this grade level.

CA-3: Beyond Standards Most of the content required for answering the question is contained in Benchmarks. Drawing does require some skills that are not a part of the benchmark, but the scoring guide indicates that accuracy is not important. There is also some requirement for spatial visualization that may be outside of Benchmarks.

CA-4: Content and Context At this grade level making a drawing is appropriate, and sketching one that is reasonably accurate is an appropriate context. Some students might do better if concrete material (a paper with the figure drawn on it) were available.

CA-5: Cognitive Demand The standard is at the conceptual level, and the task is also at this level of cognitive demand. It might be possible that students who have a great deal of experience with similar tasks could produce the drawing without conceptual understanding but not likely, since the figure is not familiar, such as a heart or tree, for example.

### *Technical Analysis Example*

TQ-1: Comprehensibility The reading level and terminology of the task is clear and appropriate. The diagram, however, is unclear in that it does not indicate where the fold is, producing the possibility of an incorrect response for students who do understand the benchmark.

TQ-2: Engagement The drawing in the task is somewhat abstract which might reduce students' interest. On the other hand, it is neutral in not being more familiar to particular students or demographic group. The task does require hands-on work which engages students and helps to produce their best efforts.

TQ-3: Clarity The question clearly indicates to students that they are expected to draw a picture of the flattened shape. On the other hand, there is no indication of how accurate the picture must be or whether there are optional possible answers.

TQ-4: Commonly Held Ideas At grades 3 or 4, students can identify shapes according to their properties (for example, number of sides, angle, curves, straight lines), which is necessary for this task. The task does not appear to provide an opportunity to identify misconceptions about these properties since "other incorrect" responses are lumped together by the scoring guide.

TQ-5: Alternative Responses There is little flexibility in the mode of response; it must be a drawing. No tools or manipulatives are provided, nor are students asked or given an opportunity to show beyond what can be inferred from their drawing how they arrived at a solution. Although different solution strategies might be used, the task does not provide an opportunity to see these.

We have held training sessions for potential assessment analysts, who have also help to clarify and refine the criteria. The analysts and Project 2061 staff have applied the analysis to a few tasks in mathematics and science at several grade levels, and produced sample reports and profiles of these tasks. Based on this preliminary work and some earlier explorations before the project was funded, we can summarize some tentative findings.

### **Findings from Assessment Analysis Research.**

Project 2061's previous efforts to analyze assessment tasks in science and mathematics have identified factors that appear to affect the alignment of specific kinds of tasks with benchmarks and standards. These factors will need to be explored further in the proposed work. For example we have found:

- Few available mathematics and science assessment tasks appear to have good content alignment with national standards and standards.
- Many tasks require students to memorize the details of specific examples or vocabulary used in instruction, rather than understanding standard ideas.
- Most tasks have one or more technical quality deficiencies, especially in their anticipation of commonly held ideas and, for open-ended tasks, clarity of expectations.
- Many tasks require knowledge in addition to specific standard concepts, making it difficult to describe alignment and, therefore, difficult to diagnose the reason for unsuccessful responses.
- For many tasks, students can use less sophisticated ideas or their general intelligence alone to respond successfully without understanding the concepts in a targeted standard.
- Some open-ended tasks allow a number of acceptable responses, each of them aligned with a different standard, making it difficult to determine which standards the task aligns with.
- Tasks or scoring guides are seldom helpful in eliciting specific student difficulties or providing information that would guide teachers on how to modify instruction.

## **Future Work**

The next phase of the project will be to apply the procedure to a variety of assessment tasks in mathematics and science, producing reports on alignment and quality for different audiences. For test developers, we will provide prototypes and examples of "closing the loop" between test specifications and assessment tasks so that there is fidelity between standards and the task that are intended to address them. For teachers, the reports will provide information and guidance about the criteria that should be expected of assessments that are designed to diagnose, monitor, and measure student learning and achievement. For policy makers, parents, and the public the reports will provide a profile of the alignment and quality of the items and tasks that are contained in current tests. In an era of American education in which testing is assuming ever higher stakes, this type of information and analysis is critical to decision makers at every level.

## **References**

Achieve, Inc. (1997). *About Achieve*. Cambridge, MA: Author.

American Association for the Advancement of Science. (1993). *Benchmarks for Science Literacy*. New York: Oxford University Press.

American Association for the Advancement of Science (1998). A proposal to develop resources for science and mathematics literacy: Assessment. Washington, DC: Author.

American Association for the Advancement of Science. (1999). *Blueprints for reform*. New York: Oxford University Press.

- Clements, D. H. & Battista, M. T. (1992). Geometry and spatial reasoning. In D.Grouws (Ed.), *Handbook of research on mathematics teaching and learning*. New York: Macmillan.
- FairTest. (1995). *Principles and indicators for student assessment*. Cambridge, MA: Author.
- National Assessment of Educational Progress (1996). 1996 Assessment mathematics public release grade 4. [www.ed.gov/nces/naep](http://www.ed.gov/nces/naep)
- National Council of Teachers of Mathematics. (1989). *Curriculum and evaluation standards for school mathematics*. Reston, VA: Author.
- National Council of Teachers of Mathematics. (1998). *Principles and standards for school mathematics: Standards 2000* (Draft). Reston, VA: Author.
- National Research Council. (1996). *National science education standards*. Washington, DC: National Academy of Sciences Press.
- Third International Mathematics and Science Study (1996). TIMSS mathematics items Released set for population 1 (3<sup>rd</sup> and 4<sup>th</sup> grade). [www.timss.org](http://www.timss.org).
- Webb, N. L. (1997). *Determining alignment of expectations and assessments in mathematics and science education*. Madison, WI: University of Wisconsin, National Center for Improving Science Education.