Measurement Issues in the Alignment of Standards and Assessments: A Case Study

CSE Report 653

Joan L. Herman, Noreen M. Webb, & Stephen A Zuniga
National Center for Research on Evaluation,
Standards, and Student Testing (CRESST)
Graduate School of Education and Information Studies
University of California, Los Angeles

May 2005

Center for the Study of Evaluation
National Center for Research on Evaluation,
Standards, and Student Testing
Graduate School of Education & Information Studies
University of California, Los Angeles
Los Angeles, CA 90095-1522
(310) 206-1532

Project 2.3 Indicators of Classroom Practice and Alignment—Strand 1: Methodological Issues
Project Directors: Joan L. Herman, Jia Wang, and Barbara Wells, CRESST/UCLA

The work reported herein was supported in part by WestEd grant No. ESI-01-0119790 (Center for Assessment and Evaluation of Student Learning) to the Center for the Study of Evaluation/CRESST and in part under the Educational Research and Development Centers Program, PR/Award Number R305B960002, as administered by the Institute of Education Sciences (IES), U.S. Department of Education.

The findings and opinions expressed in this report are those of the authors and do not reflect the positions or policies of the National Center for Education Research, the Institute of Education Sciences, or the U.S. Department of Education; nor do these findings and opinions necessarily reflect the position or policies of the Center for Assessment and Evaluation of Student Learning.

MEASUREMENT ISSUES IN THE ALIGNMENT OF STANDARDS AND ASSESSMENTS: A CASE STUDY

Joan L. Herman, Noreen M. Webb, & Stephen A. Zuniga
National Center for Research on Evaluation, Standards, and Student Testing (CRESST)
UCLA Graduate School of Education and Information Studies

This study examined the impact of rater agreement on decisions concerning the alignment between the Golden State Examination (GSE) in High School Mathematics and the University of California (UC) *Statement on Competencies in Mathematics*. UC faculty and high school mathematics teachers (n = 20) rated the mathematics items of the GSE relative to the expectations identified in the UC competency statement, identifying item features related to content and dimensionality. Raters assigned values for a primary topic, a secondary topic, item/topic centrality, and depth of knowledge. Agreement within these criteria was the basis of the assessment of alignment. Results showed considerable variability in judgments across raters and different pictures of alignment depending on the particular subset of raters providing the ratings. Results also varied by rater type and the method of determining rater agreement.

The alignment of standards and assessment is key to today's standards-based reform where assessment serves as both a lever and a measure for the reform effort. State assessments send strong signals to schools about what they should be teaching and what students should be learning, and schools respond by teaching what is assessed (Herman, 2004; Koretz, Barron, Mitchell, & Stecher, 1996; Koretz, Mitchell, Barron, & Keith, 1996; Lane, Stone, Parke, Hansen, & Cerillo, 2000; McDonnell & Choisser, 1997; Stecher, Barron, Chun, & Ross, 2000). At the same time, assessment results are expected to provide accurate information to the public, its policymakers, educators, parents, and students themselves about how students are doing and to provide stakeholders with important feedback on which to base their improvement efforts. Absent strong alignment between standards and assessment, schools may ignore desired standards and instead teach only what is tested. Moreover, if what is tested does not well reflect expectations for student performance, test results cannot provide accurate data about students' or schools' progress relative to those expectations, and improvement actions based on such results are unlikely to further intended goals. Recognizing these key validity concerns, federal Title I legislation

since 1994 has required the alignment of standards and state assessments, and current regulations under No Child Left Behind (2002) require states to conduct alignment studies to document the technical quality of their tests.

As with any area of inquiry, the methodology and rigor with which such studies are conducted have a significant influence on their results. On the one hand, most states have considered their standards and tests aligned on the basis of internal or publisher-conducted studies (Wixson, Fisk, Dutro, & McDaniel, 2002). On the other hand, independent analyses have shown uneven results (Rothman, Slattery, Vranek, & Resnick, 2002; Porter, 2002; Webb, 1999), raising questions about the relative quality of the alignment processes employed and making clear the importance of a quality process.

Pioneered by Andrew Porter and Norman L. Webb, systematic procedures for assessing alignment have been well developed (Ananda, 2003; Bhola, Impara, & Buckendahl, 2003; Herman, Webb, & Zuniga, 2003; Olson, 2003; Porter & Smithson, 2001; Rothman et al., 2002; Webb, 1997, 2002) and now are being applied in states across the country. In essence, these approaches convene panels of experts to analyze assessment items against a matrix defined by an exhaustive set of topics comprising a subject area domain and by levels of cognitive demand, reflecting a range from rote memory to procedures, applications, and complex problem solving. The matrices then become the basis for computing various indices of alignment to convey how well a test reflects intended standards. Yet, while the process rests firmly on expert or rater judgment, basic questions about the reliability of the process have not yet been fully addressed. When expert raters are used to assess student performance, questions about the reliability of ratings and the number of raters needed to achieve acceptable precision are routinely addressed by empirical study (Shavelson, Baxter, & Gao, 1993; Shavelson & Ruiz-Primo, 2000; Shavelson, Ruiz-Primo, & Wiley, 1999). Yet, in alignment studies, custom and feasibility considerations seem to have been a driving force. Typical studies have used panels of 3 to 10 content experts, including teachers and subject matter experts, to make their determinations, and the extent to which these experts' judgments are representative of a larger population and/or the extent to which disagreements among experts may influence alignment conclusions have remained unexamined (Buckendahl, Plake, Impara, & Irwin, 2000; Porter, 2002; Porter & Smithson, 2001; Webb, 1997, 1999, 2002). We believe that empirical evidence, rather than

convenience, needs to be brought to bear in assuring the reliability of the alignment process.

Our research provides a case study of core measurement issues in the context of a study of the alignment of California's Golden State Examination in High School Mathematics (GSE) with the mathematics competencies expected of entering freshman at the University of California (UC), which here serve as a proxy for state content standards. At the time of the study, the GSE was being considered as an alternative to the SAT-1 in the wake of then UC President Atkinson's decision to reconsider admissions and eligibility testing for the University and to find options that would have more productive impact on teaching and learning.

The study purposively used multiple raters and separate panels of UC faculty and high school educators to assess the alignment of the test with University expectations—essential if studying for the test was to help students become better prepared for the University, as Dr. Atkinson desired—and to explore the agreement between faculty and high schools educators about University expectations. The focus on agreement served both technical and socio-political purposes. From a psychometric standpoint, as noted above, rater agreement is essential to reliable measurement, and from a socio-political perspective, agreement represents the extent to which common understandings are shared—essential if standards and tests are to serve their intended communication and instructional purposes. The underlying assumption was that the UC statement of expectations would enable high school educators to understand what the University expected and thus to know what to teach to prepare their students for the University. In the absence of common interpretation of the expectations, however, high school educators could think they were preparing their students for UC expectations, but the specifics of their content teaching could be at variance with college expectations. The same logic is true in the case of state standards and tests: Educators must have consistent understandings of the content expectations resident in state standards and assessments if they are to prepare their students to do well.

The relatively large number of raters involved in our study provided a context for examining methods for addressing the reliability of the alignment process as commonly implemented and for considering the implications of such reliability for conclusions about the alignment between standards and an assessment. The study reported below thus addresses the following general issues:

1. How can the reliability of the alignment process be addressed? Because our alignment ratings included both categorical and discrete rating scales and those that could be considered either, we used rater agreement indices, generalizability theory, and decision studies in our exploration. We also used our sample to derive typically composed alignment panels of three educators and three college faculty and compared their results to our "gold standard" of the full complement of 20 raters.

2. What are the implications of the reliability of the process for inferences about the alignment of an assessment with standards? Here we compared results from two separate measures of alignment—comprehensiveness and balance—for our "gold standard" sample and our constituted panels of six.

3. What factors influence the reliability of ratings? Here we were limited to variables endogenous to the study, including rater status as high school teacher or college faculty and dimensions of alignment.

4. What are the implications for future research and practice? We believe our findings have both methodological and substantive implications.

Methodology

In this section we describe the panels that were convened to judge alignment and the tools they used to make the judgment: the high school mathematics test, the UC *Statement on Competencies in Mathematics Expected of Entering College Students* (see www.universityofcalifornia.edu/senate/reports/mathcomp.html), and the alignment instrument used for making comparisons between the two. We then summarize training and rating procedures and present an overview of our analysis strategies.

The Raters

We convened panels of University of California mathematics faculty and high school mathematics educators who were subject matter experts and experienced in reform and assessment issues in the K-12 educational system. A total of 10 faculty members and 10 high school educators were recruited, and separate panels of each were convened in both northern and southern California.

The 10 teachers who rated the exam averaged 13 years' teaching experience and had helped develop district standards, written exit exams, high school math programs, and the Golden State Exam. They all had experience in grading statewide exams, including the GSE. The 10 UC faculty had extensive background in K-12 mathematics, including participation in the development of California's curriculum

framework for mathematics, review of the math content for the state's assessment, teacher training for the California Math Project, scoring of statewide performance assessments, and development of UC mathematics competencies.

High School Mathematics Golden State Examination

The study used the 2001 Golden State Exam in High School Mathematics, intended for students who had completed 2 years of high school algebra and geometry. Administered in two 45-minute sessions, the test consisted of 40 multiple-choice questions and 2 written-response items. The content was based on the *Mathematics Standards for California Public Schools, Kindergarten Through Grade 12* (California Department of Education, 1997; see also California Department of Education, 10/04/01, 10/16/01) and covered topics in algebra I, geometry, algebra II, and probability and statistics. Students designated at the three highest levels on the test received recognition as Golden State Scholars.

Statement on Competencies in Mathematics Expected of Entering College Students

The *Statement on Competencies in Mathematics Expected of Entering College Students* was developed by a joint task force of representatives from the University of California, California State University, and the California Community Colleges and was formally adopted by UC academic senates as the University's official position. The document is intended to provide a clear picture of what mathematics students need to know and be able to do to be successful in college. Section III, the core section for the current study, describes areas of mathematical content that are

1. essential for all entering college students;
2. desirable for all entering college students;
3. essential for college students to be adequately prepared for quantitative majors; and
4. desirable for college students who intend to declare quantitative majors.

The full statement can be found at www.universityofcalifornia.edu/senate/reports/mathcomp.html This study focuses on topics in category 1 essential for all entering college students (see Appendix A).

Alignment Rating Instrument

Adapted from procedures developed by Norman L. Webb (1997, 1999), the Alignment Rating Instrument asked reviewers to examine each item on the GSE in the following ways:

1. Identify the content topic(s), if any, from the *Statement on Competencies* to which each item corresponded. Raters could identify both a primary and secondary topic, as appropriate, and with these selections implicitly made judgments about each item's dimensionality. Items for which only a primary topic was identified were considered unidimensional, and those with a secondary topic were defined as multi-dimensional.

2. Rate the centrality of the item to the topic it addresses, using the following rating scale:

   - Within the topic area, but not essential for students
   - Within the topic area, of moderate importance
   - Within the topic area, of central importance

3. Judge the depth-of-knowledge level of each assessment item, using the following levels (see Appendix B for detailed descriptions):

   - Recall and Reproduction (Level 1)
   - Skills and Concepts (Level 2)
   - Problem Solving and Strategic Thinking (Level 3)
   - Extended Thinking (Level 4)

As described by Webb (1997, 1999), this 4-point hierarchy is based on two factors: the mathematical sophistication of an item and the likelihood that students were familiar with the problem type through prior instruction. The mathematical sophistication of the item depended on such things as the abstractness of the problem, the number of mathematical principles to be employed, problem novelty, and the need to extend or produce original findings. However, these characteristics could be difficult to judge, because assessment items may look challenging to a novice but in fact represent a low depth-of-knowledge level because the knowledge required to solve the item is commonly taught, and students are likely to have had the opportunity during normal instruction to routinely (habitually) solve such items. Anything that was considered routine or algorithmic in this sense was considered low (level 1) depth of knowledge.

Rating Procedures

Prior to the rating meetings, participants were informed of the general goals of the study and were sent the *Statement on Competencies* to review. At the meetings, participants were given additional orientation to the project and its goals and then introduced to the rating instrument and process. Participants reviewed the specific written guidance provided on each of the rating criteria, practiced using the coding scheme, shared answers, and discussed points of disagreement until reaching reasonable levels of agreement. Panelists then embarked on individual ratings of each item on the test. After all ratings had been completed, a debriefing session was held for participants to provide their general reactions to the rating process.

Analysis Methods and Decisions

The analysis plan considered appropriate ways to assess rater consistency as well as decisions about when to consider an item aligned with a particular topic or category. In addition to reporting descriptive information showing exact agreement among raters, we calculated kappa coefficients for categorical ratings (specific mathematics topic and category assignment, item dimensionality) and dependability coefficients for ratings that had inherent quantitative meaning (depth of knowledge and centrality of the item for measuring a particular topic). Where ratings could be considered either categorical or continuous, we considered multiple indices.

We report rater consistency and alignment results for the full panel of 20 raters and for all 6-rater subsets of three high school teachers and three college faculty that could be constituted from the full panel (described in detail in a later section), and we compare the results of each. In addition, we compare rater consistency and alignment results for high school teachers and college faculty.

Agreement Analyses

To examine the agreement among raters for the categorical ratings, we calculated kappa coefficients of agreement (Cohen, 1960; Fleiss, 1971). Because the kappa coefficient takes into account chance agreement among observers, it is preferred over other summary indices such as exact percent of agreement (see Watkins & Pacheco, 2001).

To examine the agreement among raters for the quantitative ratings, we conducted generalizability analyses with items crossed with raters. Each generalizability analysis produced an estimated index of dependability, a reliability-

like coefficient that showed the consistency of raters in coding attributes of the items. The index of dependability provides information about rater consistency on the absolute level of an item attribute (e.g., depth of knowledge), not the consistency of raters in their assessment of the relative standing of items on a particular attribute such as would be provided by intraclass correlations (see Brennan, 2001; Shavelson & Webb, 1991).

Rater Agreement Yardsticks for Determining When a Topic or Category Was Covered

To examine the alignment between the test items and the topics listed in the *Statement on Competencies,* we had to have a decision rule about the minimum rater agreement necessary for declaring that a test item measured a particular topic. We used a 65% agreement level (agreement of 13 out of 20 raters) for analyses using all 20 raters, a 67% agreement level for 6-rater subsets (4 out of 6 raters), and a 70% agreement level for analyses of rater types (7 out of 10 raters) because these levels represented a clear majority of raters.

Because our training process included only general rules for defining primary and secondary topics, we could not be confident about the relative weights that raters gave to their primary and secondary topic ratings. To determine the specific topic agreed upon by the raters, then, we decided to combine the primary and secondary ratings given by each rater to each item and use whichever topic rating, if either, agreed more strongly with the ratings given by other raters.

Results

In this section, we report rater agreement and alignment results for (a) the full panel of 20 raters and (b) the 6-rater subsets drawn from the full 20-rater panel. In interpreting the results from the 6-rater subsets, we use the decisions made by the full 20-rater panel as the "gold standard."

Results for the 20 Raters: The Gold Standard

This section examines the agreement among the 20 raters when assigning mathematics topic, mathematics content category, dimensionality, depth of knowledge, and centrality ratings to each item. Using the results for rater agreement, we summarize how the 20 raters characterized the test in terms of these five item features.

Classification of Test Items by Specific Topic and General Content Category

Rater agreement. We first explored interrater agreement for raters' specific topic assignments (57 specific topics) and category assignments (10 categories) by calculating coefficient kappa for multiple raters (Fleiss, 1971). The constraints of computer programs for calculating kappa did not allow us to analyze the results of all 20 raters simultaneously; consequently we calculated kappa coefficients separately for faculty (10 raters) and teachers (10 raters). The kappa coefficients for assignment of items to specific topics were .55 and .58 for faculty and teacher raters, respectively; the kappa coefficients for assignment of items to categories were .71 and .74 for faculty and teacher raters, respectively. According to the guidelines suggested by Watkins and Pacheco (2001; see also Cicchetti, 1994; Fleiss, 1981), kappa coefficients greater than .75 indicate excellent agreement, values between .60 and .75 indicate good agreement, values between .40 and .60 indicate fair agreement, and values below .40 indicate poor agreement. Consequently, the agreement levels among the 20 raters for assignment of items to specific topics and general categories can be characterized as fair to good. Though the kappa coefficients suggested a moderate level of rater agreement across the 42 test items, inspection of the data showed considerable variability of rater agreement from item to item. For some items, all 20 raters assigned the same specific topic, whereas for other items, very few raters agreed on a specific topic. Table 1 gives the number of items for which at least 65% of the 20 raters agreed on the specific topic and the number of items which met this same agreement threshold for the general content category. As seen in Table 1, raters reached agreement about the specific topic assignment on 30 (71%) of the items on the test, and reached agreement about the general content category on 40 (95%) of the test items. Inspection of the ratings given by faculty and teachers showed that both rater groups generally assigned the same specific topics and the same general topic categories for each of the 30 items.

No item features predicted topic agreement. First, average agreement level for an item did not relate to its depth of knowledge, dimensionality, or centrality. Second, items on which raters agreed at this study's agreement threshold did not differ on these item features from items on which raters did not agree.

Alignment of the test to the UC competencies for entering freshmen. We examined alignment in two ways. First we looked at the comprehensiveness of content coverage, defined as the proportion of topics addressed by at least one item on the test. Using the topic decisions made by the 20 raters on the 30 items for which

Table 1

Level of Agreement Among the 20 Raters on Item Features

Item feature	Number of items	Percent of items
Assignment of items to specific topics		
65% agreement or higher[a]	30	71
Lower than 65% agreement	12	29
Assignment of items to general content categories		
65% agreement or higher	40	95
Lower than 65% agreement	2	5
Item dimensionality: whether items require a secondary topic assignment		
65% agreement or higher	18	43
Lower than 65% agreement	24	57
Depth of knowledge of items		
65% agreement or higher	27	64
Lower than 65% agreement	15	36
Centrality of items		
65% agreement or higher	31	74
Lower than 65% agreement	11	26

[a]Agreement among at least 13 of 20 raters.

they reached agreement, the raters judged that only a third of the topics considered essential for entering UC freshmen (14 of 41 topics, 34%) were represented on the test. A more specific picture of content coverage appears in Figure 1: the proportion of topics in each general category that were addressed by at least one item on the test. Figure 1 shows that the 20 raters agreed that the test represented roughly 40% of the topics in each of four categories: 38% (3 out of 8 topics) in the category Variables, Equations, and Algebraic Expressions, and 40% for Families of Functions and Their Graphs (4 of 10 topics), Geometric Concepts (4 of 10 topics), and Probability (2 of 5 topics). For Data Analysis and Statistics, comprehensiveness of content coverage dropped to 25% (1 of 4 topics); and for Argumentation and Proof, the 20 raters perceived none of the four topics to be represented by any item on the test. We consider Figure 1 to represent an estimate of the "benchmark" or "gold standard" for comprehensiveness of content coverage, to be compared with pictures of alignment produced by subsets of the 20 raters, considered in a later section.

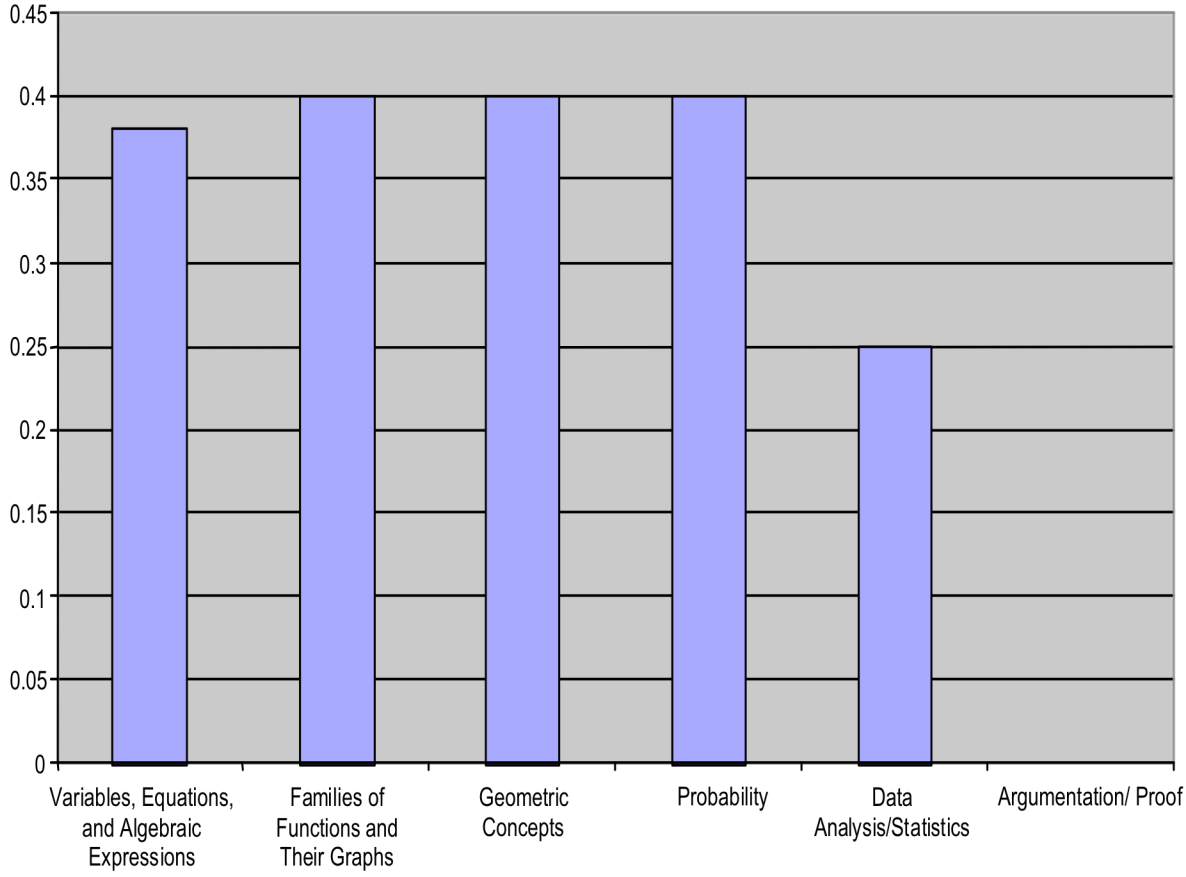

Figure 1. 20-Rater picture of comprehensiveness of content coverage: Proportion of topics in each content category addressed by at least one item on the test. (Figure includes the 30 items for which at least 65% of the raters agreed on the specific topic.)

Second, we examined the balance of content coverage, defined as the proportion of the total test (42 items) that addressed each general content category (Figure 2). As was already seen in Table 1, raters did not meet agreement on content category classification for two of the 42 items (5%), which is represented by the right-hand bar in Figure 2. For the remaining 40 items for which raters agreed on content category classification, raters saw that the three categories Variables, Equations, and Algebraic Expressions, Families of Functions and Their Graphs, and Geometric Concepts received the most (and equal) attention on the test, with 24% of the items fitting into each of these categories. Less of the test addressed Probability (14%). Very little of the test addressed Data Analysis and Statistics (2%) and no item addressed Argumentation and Proof. Finally a small portion of the test addressed

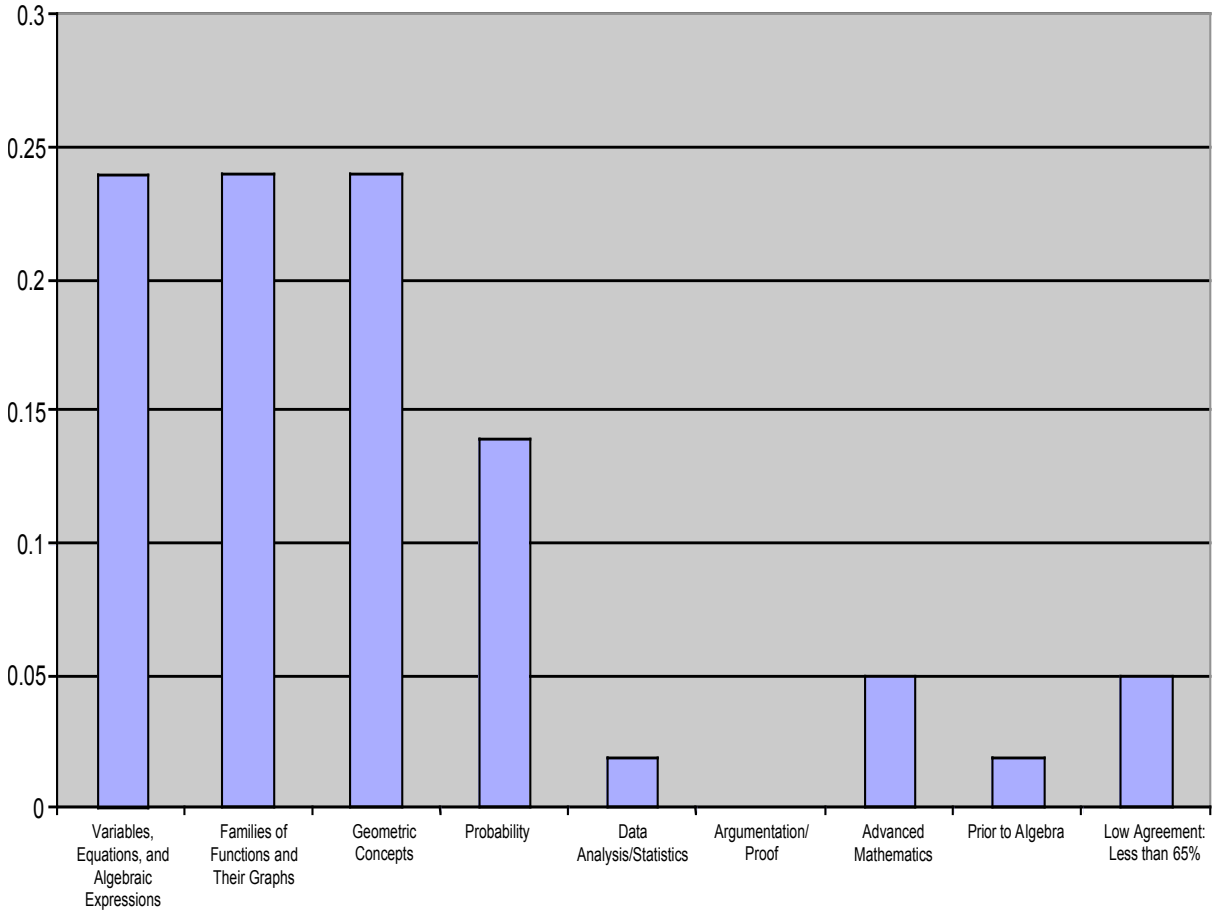

Figure 2. 20-Rater picture of balance of content coverage: Proportion of the total test addressing each content category. (Figure includes the 30 items for which at least 65% of the raters agreed on the general content category.)

content that was perceived to be more advanced than that considered to be essential for entering freshmen (e.g., inverse functions and their graphs) or content covered prior to algebra I. Figure 2 will be considered an estimate of the benchmark for balance of content coverage (the "gold standard"), to be compared with pictures of balance of content coverage produced by subsets of the 20 raters, to be described in a later section.

Although raters perceived that most of the test addressed algebra and geometry, this result should be tempered by the results in Figure 1, showing that relatively few topics in any general content category were represented on the test. For example, about a quarter of the test (10 items, 24%) was devoted to Families of

Functions and Their Graphs (Figure 2) but these items only addressed 4 out of 10 topics judged to be essential for entering college freshmen (Figure 1). Multiple items measured the same topic; for example, three items dealt with exponential functions. These results show that it is important to describe alignment in terms of both comprehensiveness and balance of content coverage.

Assessment of Item Dimensionality: Assignment of One vs. Two Topics

Kappa coefficients for dimensionality ratings were low (.16 for faculty, .13 for teachers), showing that raters barely agreed above a chance level about whether an item should be assigned only a primary topic or both primary and secondary topics; that is, whether an item should be considered uni- or multidimensional. Applying the 65% agreement rule to determine the number of items for which raters agreed on item dimensionality yielded similar results. On fewer than half of the items (45%) did a minimum of 13 out of 20 raters agree on item dimensionality. Of the 18 items that the raters agreed upon, they classified 9 as multidimensional (addressing multiple topics) and 9 as unidimensional (addressing only one topic). The decisions made on these 18 items will serve as the benchmark for item dimensionality to be compared with decisions produced by 6-rater subsets of the 20 raters, considered in a later section. Dimensionality was one of the few dimensions on which differences between faculty and high school teacher raters emerged. Teachers rated more items as multidimensional than faculty did. Faculty agreed on item dimensionality for 27 items. Faculty classified 41% of these 27 items as multidimensional. Teachers also agreed on item dimensionality for 27 items (although not all of the same items as faculty), and classified 70% of them multidimensional. The difference between these proportions is statistically significant, $\chi^2(1) = 4.80$, $p = .028$.

Interestingly, on six items, faculty and teachers came to opposite conclusions about item dimensionality, with faculty judging them to be unidimensional and teachers judging them as multidimensional. For example, for items presenting specific functions (e.g., $f(x) = 2x^3 - 3$), teachers often assigned the secondary topic "function notation" in addition to their primary topic (e.g., linear functions, exponential functions, inverse functions and their graphs), whereas faculty rarely did so. In contrast to college faculty, then, high school teachers conceptualized interpretation of function notation as a separate skill, perhaps reflecting their approach to teaching students how to handle such problems.

Assessment of Items' Depth of Knowledge

As another indicator of the degree of intellectual challenge of an item, raters were asked to assign depth-of-knowledge ratings using a quantitative scale with 4 ordered levels (although raters in this study used only levels 1 to 3 in their ratings). Using these ratings, it is possible to conceive of depth of knowledge as a continuous scale (for example, with an item scored 2.2 as having higher depth of knowledge than an item scored 1.6), or, alternatively, as a scale with discrete levels (that is, an item requires skills and concepts, level 2, or it does not; requires only recall and reproduction, level 1, or not). Treating depth of knowledge as a continuous scale makes it possible to define an item's mean depth of knowledge independent of rater agreement, whereas treating depth of knowledge as a set of discrete values requires a certain level of rater agreement to define a value for an item. For example, if equal numbers of raters assign values of 1, 2, and 3 to an item, the item can be assigned a value of 2 on the continuous scale, but cannot be assigned a value if the ratings are assumed to be discrete levels because there is insufficient agreement among raters to do so.

Treating depth of knowledge as a continuous scale makes it possible to examine overall summary statistics for the test. Averaging over all 20 raters' judgments about the depth of knowledge yielded a mean of 1.59 across the 42 items on the test and a standard deviation of .35. On average, raters judged the items to require low to medium depth of knowledge, although there was considerable variability across items. To obtain a more detailed picture of the range of depth of knowledge across items, we divided the continuous distribution into intervals: item values from 1.0 to 1.49 corresponded to low depth of knowledge, item values from 1.50 to 2.49 corresponded to medium depth of knowledge, and item values 2.50 and above corresponded to high depth of knowledge. Using this approach yielded the following distribution of item depth of knowledge as judged by the 20 raters: 16 items were low (1), 25 were medium (2), and one was considered relatively high (3).

The summary just given does not provide any information about the variability in depth-of-knowledge ratings across raters. To analyze rater agreement about depth of knowledge, we carried out a generalizability study using a design with items crossed with raters (i x r design) in which both sources of variation (items, raters) were treated as random. In this design, because the intent was to measure the depth of knowledge of the items, the item was the object of measurement. The rater constituted the source of error (called a facet in generalizability theory). The

generalizability study provides information about the magnitude of variation across raters that in turn can be used to determine the number of raters that should be used in a decision study about alignment. In the current case, we use the results of the generalizability study to estimate reliability for a decision study in which alignment decisions are to be made using 6 (instead of 20) raters.

The estimated variance components from the generalizability study appear in Table 2. The estimated variance component for items ($\hat{\sigma}_i^2 = .1124$) provides information on item-to-item differences; it is called universe-score variance and is analogous to true-score variance in classical test theory. The main effect for raters ($\hat{\sigma}_r^2 = .0246$) is small compared to the variability across items, showing that, averaging over items, raters did not differ greatly in their ratings of depth of knowledge. The rater means corroborate this result. Averaging over the 42 items on the test, rater means ranged from 1.26 to 1.90 on the 4-point depth-of-knowledge scale, showing that all raters perceived the items, on average, to go beyond recall and reproduction (level 1) and involve some skills and concepts (level 2).

The very large estimated variance component for the residual ($\hat{\sigma}_{ir,e}^2 = .2326$, 63% of the total variance) relative to the estimated variance component for items suggests a large item x rater interaction (raters rank-ordered items differently on depth of knowledge), and/or other sources of error variability not captured with this design.

Table 2

Estimated Variance Components From Generalizability Studies of Depth-of-Knowledge and Centrality Ratings

	Item feature							
	Depth of knowledge				Centrality			
Source of variation	Generalizability (G) study		Decision (D) study with 6 raters		Generalizability (G) study		Decision (D) study with 6 raters	
Items (i)	$\hat{\sigma}_i^2$	.1124	$\hat{\sigma}_i^2$	.1124	$\hat{\sigma}_i^2$	.0165	$\hat{\sigma}_i^2$	.0165
Raters (r)	$\hat{\sigma}_r^2$	.0246	$\frac{\hat{\sigma}_r^2}{n_r'}$	.0041	$\hat{\sigma}_r^2$	.0988	$\frac{\hat{\sigma}_r^2}{n_r'}$	.0165
ir,e	$\hat{\sigma}_{ir,e}^2$	.2326	$\frac{\hat{\sigma}_{ir,e}^2}{n_r'}$	.0388	$\hat{\sigma}_{ir,e}^2$	.2153	$\frac{\hat{\sigma}_{ir,e}^2}{n_r'}$	.0359

Note. n'_r is the number of raters in the decision (D) study. Here, $n'_r = 6$.

Table 2 also gives the estimated variance components for a decision study using the mean of six raters' depth-of-knowledge ratings. This number of raters ($n_i' = 6$) was selected because it represents the number of raters typically used in alignment studies (e.g., Buckendahl et al., 2000, 10 raters; Porter, 2002, 3 raters; Webb, 1997, 1999, 2002, 4-6 raters). Because interest in this study lies in identifying the absolute level of depth of knowledge of an item (an absolute decision) rather than in rank ordering items in terms of depth of knowledge (a relative decision; Shavelson & Webb, 1991), both rater variation ($\hat{\sigma}_r^2$) and the residual variation ($\hat{\sigma}_{ir,e}^2$) contribute to error variation here (see Appendix C). Following the procedures given in Appendix C, the estimated error variance for absolute decisions is .0429 (.0041 + .0388).

The absolute error variance can be used to gauge the consistency of item depth-of-knowledge scores across different randomly sampled sets of six raters. In particular, we can use the square root of estimated absolute error variance ($\sqrt{.0429} = .2071$) as the standard error of measurement (SEM; see Brennan, 2001) to construct a confidence interval that shows the likely range of an item's depth-of-knowledge rating that would be produced across randomly sampled sets of six raters. A 95% confidence interval is ± 1.96 * .2071, which gives an interval of width .8118. This width is quite large, nearly a whole point on the 4-point scale for depth of knowledge, suggesting considerable variability (inconsistency) in an item's depth-of-knowledge rating across 6-rater samples.

Absolute error variance can also be used to calculate a reliability-like coefficient called an index of dependability for absolute decisions ($\hat{\Phi}$, Brennan, 2001; Brennan & Kane, 1977; see Appendix C). For a decision study with six raters, $\hat{\Phi}$ is .72, suggesting that using six raters would produce a moderate level of dependability for estimating items' level of depth of knowledge. The preceding description of results concerning the absolute variance, however, suggests that "moderate" may be an overly optimistic characterization of dependability of raters' ratings of depth of knowledge.

How well raters agree about whether an item meets a certain threshold of depth of knowledge (analogous to making mastery/non-mastery decisions about examinees in a criterion-referenced measurement context) may be more relevant to alignment decisions than information about agreement among raters in coding the absolute level of an item's depth of knowledge. In the current study, a relevant threshold may be whether an item's depth of knowledge is above or below a value

of 2.0 (requires skills and concepts). The dependability coefficient for making this decision is denoted as $\hat{\Phi}(\lambda)$ where λ is the threshold (Brennan, 2001; see Appendix C). For $\lambda = 2.0$, the dependability coefficient is .87, suggesting that using six raters would produce a fairly high level of dependability for deciding whether item depth of knowledge meets this threshold.

An alternative analytic approach is to treat depth of knowledge as a categorical variable with discrete levels, which requires raters to reach a threshold of agreement to declare an item's depth of knowledge. Consistent with the approach to examining rater agreement for topic assignment and item dimensionality, we examined rater agreement for depth of knowledge using the 65% agreement rule (a minimum of 13 out of 20 raters making the same decision) to determine the number of items for which raters agreed on depth of knowledge. The results suggested moderate (but not high) agreement among raters in their assessment of items' depth of knowledge. Table 1 shows that, at the 65% agreement level, the 20 raters agreed on the depth of knowledge for about two thirds of the items: 27 (64%) items. Of these 27 items, raters classified 15 items as requiring low depth of knowledge, 11 items as requiring medium depth of knowledge, and one item as requiring high depth of knowledge.

Interestingly, the two approaches to analyzing depth of knowledge, using all raters' ratings on the 1-to-4 scale versus using categorical ratings only for items reaching the 65% agreement threshold, produced somewhat different pictures about the distributions of items' depth of knowledge. As seen above, using all raters' ratings on all items yielded a test with the majority of items (25 of 42 items, 59%) characterized as medium depth of knowledge. Focusing on only those items on which raters reached sufficient agreement yields a test in which only a minority of items (11 of 42 items, 26%) was characterized as medium depth of knowledge.

Finally, as was the case for item dimensionality, the results for all 20 raters combined mask differences between faculty and teacher raters in depth-of-knowledge ratings. First, considering depth of knowledge as a continuous scale and including ratings on all items, teachers rated the items as requiring more depth of knowledge ($M = 1.67$, $SD = .36$), on the average, than did faculty ($M = 1.52$, $SD = .38$), a statistically significant difference, $t(41) = 4.69$, $p < .001$. A similar, although not statistically significant, trend appeared when we considered depth of knowledge as discrete categories and looked at items meeting the 70% agreement threshold (7 out of 10 faculty raters; 7 out of 10 teacher raters) for depth of knowledge. Teachers rated a majority of items (57%) as requiring at least medium depth of knowledge

whereas faculty rated a minority of items (33%) as requiring at least medium depth of knowledge. This difference between faculty and teacher raters did not reach statistical significance, however, $\chi^2(1) = 2.56$, $p = .11$.

The fact that teachers saw greater item dimensionality and depth of knowledge in the test than did faculty may be due in part to the raters' failing to distinguish the two item features. As an indicator of the overlap between raters' perceptions of these two indicators of intellectual challenge, we looked at the correspondence between an item's depth of knowledge and its dimensionality for those items reaching the minimum threshold of rater agreement (65% of raters agreeing on each dimension). On the 15 items for which raters reached minimum agreement levels for both depth of knowledge and dimensionality, the ratings of the two indicators matched almost perfectly. Items that were rated as requiring low depth of knowledge were rated as addressing only one topic area (undimensional); items that were rated as requiring medium or high depth of knowledge were rated as addressing multiple topic areas (multidimensional). Only one item deviated from this pattern. When we analyzed faculty and teacher raters separately, we found the same results. Faculty reached agreement (a minimum of 7 out of 10 raters making the same judgment) on the two indicators for 18 items. Faculty ratings of depth of knowledge and dimensionality corresponded for 16 out of the 18 items. Similarly, teacher ratings of depth of knowledge and dimensionality corresponded for 16 out of 17 items.

The overlap in depth of knowledge and dimensionality ratings suggests that raters often did not distinguish between these two indicators of intellectual challenge. If an item was perceived to assess multiple topic areas, raters tended to assign it a medium or high depth-of-knowledge rating. Items that addressed only one topic area were rated as requiring low depth of knowledge. Whether this is the true character of the items on this form of the test (for example, there were no items confined to a single topic domain that required high depth of knowledge), or whether raters were not able to distinguish between these indicators is not known. It is of interest to note that depth of knowledge ($r = .37$, $p < .05$), but not dimensionality ($r = .12$), was significantly related to student performance—that is, students found items to be most difficult when they were judged to have high depth of knowledge but not when they spanned two topics. In any event, the findings do suggest the need in future studies either to better disentangle these two dimensions or to combine them.

Assessment of Items' Centrality

Lastly, raters were asked to indicate the centrality of items to the topic area by indicating whether an item fit a topic area but was not essential in assessing student understanding of that topic (1), was moderately important (2), or was of central importance (3). As was the case for depth of knowledge, item centrality could be treated as a continuous scale using all raters' ratings or as a scale with discrete levels in which an item's centrality can only be defined when raters reach a threshold of agreement. Averaging over all raters' ratings on the continuous scale yielded a high mean rating for the test overall (M = 2.67, SD = 0.16), indicating that the 20 raters judged the test's items to be highly central to the topics being measured, on average. Dividing the continuous distribution into intervals showed a similar result: Raters judged 38 of 42 items (90%) to be of central importance (ratings of 2.50 or above) and judged 4 items (10%) to be of moderate importance (ratings between 1.50 and 2.49).

As before, we carried out a generalizability analysis to examine rater agreement. The results of the generalizability study and estimated dependability for a decision study using the average of six raters' ratings appear in Table 2. In Table 2, the estimated variance component for items ($\hat{\sigma}_i^2$ = .0165, universe-score variance) was very small compared to the other estimated variance components in the generalizability study ($\hat{\sigma}_r^2$ = .0988 and $\hat{\sigma}_{ir,e}^2$ = .2153), suggesting that item centrality differed little from item to item, on average. Indeed, the mean item centrality ratings (averaged across the 20 raters) showed a restricted range, from 2.30 to 3.00 on the 3-point scale. The large main effect for raters ($\hat{\sigma}_r^2$ = .0988, 30% of the total variance) shows that raters saw different centrality, on the average across items, with some raters perceiving items to be central and other raters perceiving items to be less central. The rater means were, in fact, quite disparate, ranging from 1.62 to 3.00. The very large estimated variance component for the residual ($\hat{\sigma}_{ir,e}^2$ = .2153, 65% of the total variance) suggests a large item x rater interaction (raters rank-ordered items differently on centrality), and/or other sources of error variability not captured with this design.

For a decision study with six raters, the estimated absolute error variance is .0524 (.0165 + .0359; see Appendix C). Using the square root of absolute error variance ($\sqrt{.0524}$ = .2289) as the standard error of measurement (SEM) to construct a 95% confidence interval for items' centrality (± 1.96 * .2289) gives an interval of width of .8973. This width is quite large, nearly a whole point on the 3-point scale

for item centrality, suggesting considerable variability (inconsistency) in an item's centrality rating across 6-rater samples.

The estimated level of dependability of centrality ratings for a decision study with six raters is quite low, $\hat{\Phi}$ =.24, due in part to the small estimated variance component for items ($\hat{\sigma}_i^2$ = .0165). The dependability of decisions about whether items were highly central ($\hat{\Phi}(\lambda)$ with λ = 3.0) was considerably higher, .67. This result suggests that the dependability of a 6-rater panel deciding whether items were central would be higher than their dependability in deciding the absolute level of item centrality.

Analyzing item centrality as a categorical variable revealed moderate agreement among the 20 raters. Applying the 65% agreement rule (a minimum of 13 out of 20 raters agreeing), raters agreed on the level of item centrality on the majority of test items (31 of 42 items, 74%; see Table 1). Raters assessed all 31 items as being highly central for measuring student understanding of the designated topic area. Faculty and teacher raters did not differ in their judgments about item centrality.

Rater Agreement on Combinations of Item Features

As was seen in Table 1, raters showed moderate to low agreement when assigning topics to items or assessing dimensions of intellectual challenge, such as item dimensionality, depth of knowledge, and item centrality. Agreement among raters dropped considerably when multiple item features were considered simultaneously. For example, the 20 raters agreed on both an item's topic and its centrality for 24 items; an item's topic and its depth of knowledge for 22 items; an item's topic and its dimensionality for 14 items; an item's topic, depth of knowledge, and centrality for 19 items; and an item's topic, its depth of knowledge, its dimensionality, and its centrality for only 11 items. These results suggest that it may be difficult to obtain sufficient rater agreement to judge multiple dimensions of alignment simultaneously.

Variability in Ratings Among 6-Rater Subsets

As described above, we recognize that studies of alignment are likely to be carried out by far fewer than the 20 raters used in this study, typically only 6. To more directly explore the dependability of raters' judgments for small panels (6 raters each), we examined the variation in ratings across the 6-rater subsets that could be formed from the 20 raters used in this study. We considered only rater

subsets composed of three faculty and three teachers because rater panels in alignment studies are likely to represent perspectives from multiple rater populations (Webb, 1997, 2002). This decision rule yielded 14,400 different 6-rater subsets that could be composed from the full set of 20 raters.

Classification of Test Items by Specific Topic and General Content Category (6-Rater Subsets)

We examined the variability of rater judgments about specific topic and general content category assignments in two ways: the success of rater subsets in matching the decisions made by the 20 raters (the "gold standard"), and variability in agreement across rater subsets without regard to whether they agreed with the decisions made by the 20 raters.

Table 3 gives information about how well the 6-rater subsets matched the ratings produced by the 20 raters. We used the decision rule that a minimum of 4 out of 6 raters, 67%, had to agree with the decision made by the 20 raters, approximating the proportional standard used for the full panel. As reported earlier, and reported in the first row of Table 3, the 20 raters agreed (at a minimum 65% of raters agreement) on the specific topic assignment for 30 items. The number of items for which rater subsets matched the topic chosen by the 20 raters ranged from 21 to 30. Further inspection of the distribution of the judgments made by the 6-rater subsets showed that 7% (1,037) agreed with the 20 raters' judgments on all 30 items.

Table 3

Variability Among 6-Rater Subsets[a] in Agreement on Item Features—With Matching (Number of items on which raters in a subset showed agreement *and* matched the standard set by the 20 raters)

Item feature	Number of items agreed upon by 20 raters	Number of items on which raters in a subset showed agreement[b]		
		Range	*M*	*SD*
Specific topic rating	30	21-30 items	27.29	1.59
Content category rating	40	32-40 items	37.52	1.41
Item dimensionality	18	10-18 items	16.13	1.37
Depth of knowledge	27	14-26 items	22.27	1.91
Item centrality	31	11-31 items	26.51	3.74

[a]14,4000 6-rater subsets.

[b]Agreement of at least 67% (4 out of 6 raters).

Furthermore, about half of the rater subsets (47%) agreed with the 20 raters' judgments on 27 or more items. These results show moderate success of many rater subsets in matching the judgments made by the 20 raters (the "gold standard").

A similar picture emerges for success of the rater subsets in matching the 20-rater judgments about the general content category for an item. Rater subsets agreed with the 20-rater standard on a range of 32 to 40 items. More than half (53%) of the 6-rater subsets agreed with the 20-rater judgments on at least 38 items.

Table 4 presents information about the variability among the 6-rater subsets without regard to the judgments made by the 20 raters. On average, 6-rater subsets reached agreement about specific topic assignment on 32 items. Some 6-rater subsets agreed about topic assignment on as many as 39 items (nearly the whole test), whereas other 6-rater subsets agreed about topic assignment on as few as 23 items (only about half of the items on the test). Rater subsets also showed variability in their agreement about general content category. Some rater subsets agreed on the content category for all 42 items while others agreed on the content category for 33 items.

Rater subsets produced quite different pictures of alignment. Whereas the 20 raters saw that 34% of the 41 topics considered essential for entering UC freshmen were represented on the test, the percentage of topics that 6-rater subsets saw as

Table 4

Variability Among 6-Rater Subsets[a] in Agreement on Item Features—Without Matching (Number of items on which raters in a subset showed agreement without regard to whether rater subsets matched the standard set by the 20 raters)

	Number of items on which raters in a subset showed agreement[b]		
Item feature	Range	*M*	*SD*
Specific topic rating	23-39 items	32.03	2.29
Content category rating	33-42 items	38.92	1.22
Item dimensionality	20-41 items	31.66	3.05
Depth of knowledge	19-42 items	32.09	2.73
Item centrality	16-42 items	33.07	4.80

[a]14,4000 6-rater subsets.

[b]Agreement of at least 67% (4 out of 6 raters).

being represented on the test ranged from 32% to 56% (M = 44, SD = 3). Moreover, the specific profiles of comprehensiveness and balance of content coverage varied considerably across rater subsets and often differed from the profiles produced by the 20 raters, as can be seen in the example profiles presented in Figures 3 and 4. Concerning comprehensiveness (Figure 3), unlike the 20 raters (Figure 1), this rater subset perceived that the test did a fairly good job measuring topics in Variables, Equations, and Algebraic Expressions and in Probability (Figure 3). Concerning balance (Figure 4), this rater subset perceived the test to be heavily weighted by items measuring Variables, Equations, and Algebraic Expressions, with fewer items measuring Families of Functions and Their Graphs and Geometric Concepts than did the 20 raters (Figure 2).

Assessment of Item Dimensionality (6-Rater Subsets)

Table 3 also shows that rater subsets varied in the success with which they matched the 20 raters' decisions about item dimensionality. Some rater subsets agreed with the 20 raters' decisions on the same 18 items, whereas other rater subsets matched the 20 raters' decisions on only 10 items.

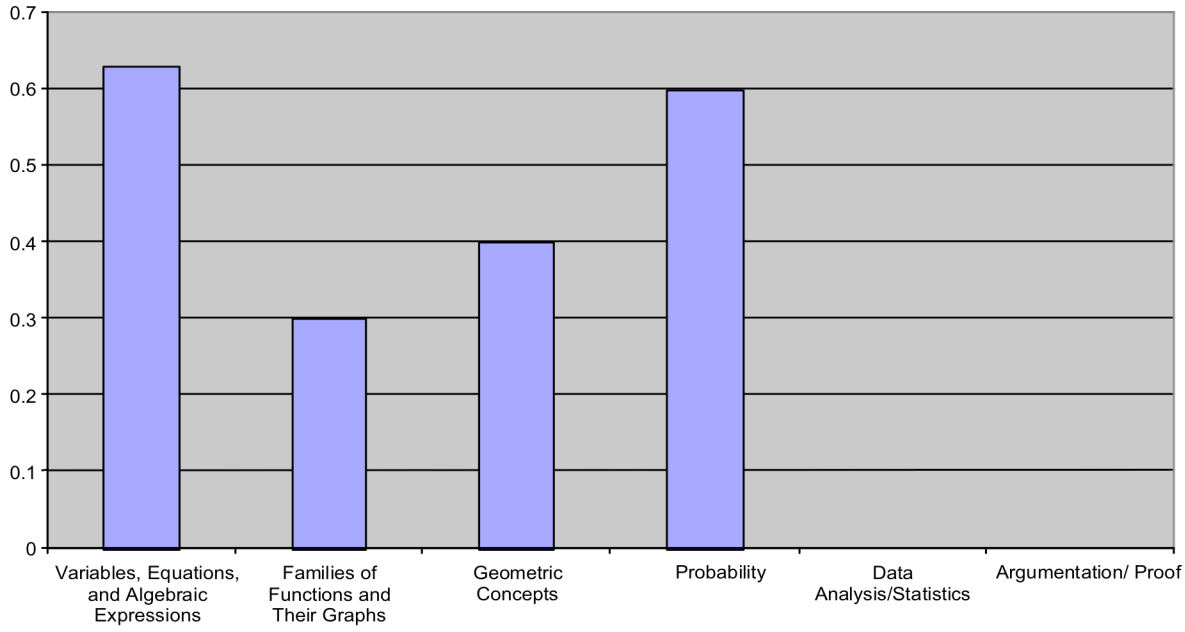

Figure 3. Picture of comprehensiveness of content coverage for one 6-rater subset: Proportion of topics in each content category addressed by at least one item on the test. (Figure includes the 34 items for which at least 67% of the raters agreed on the specific topic.)

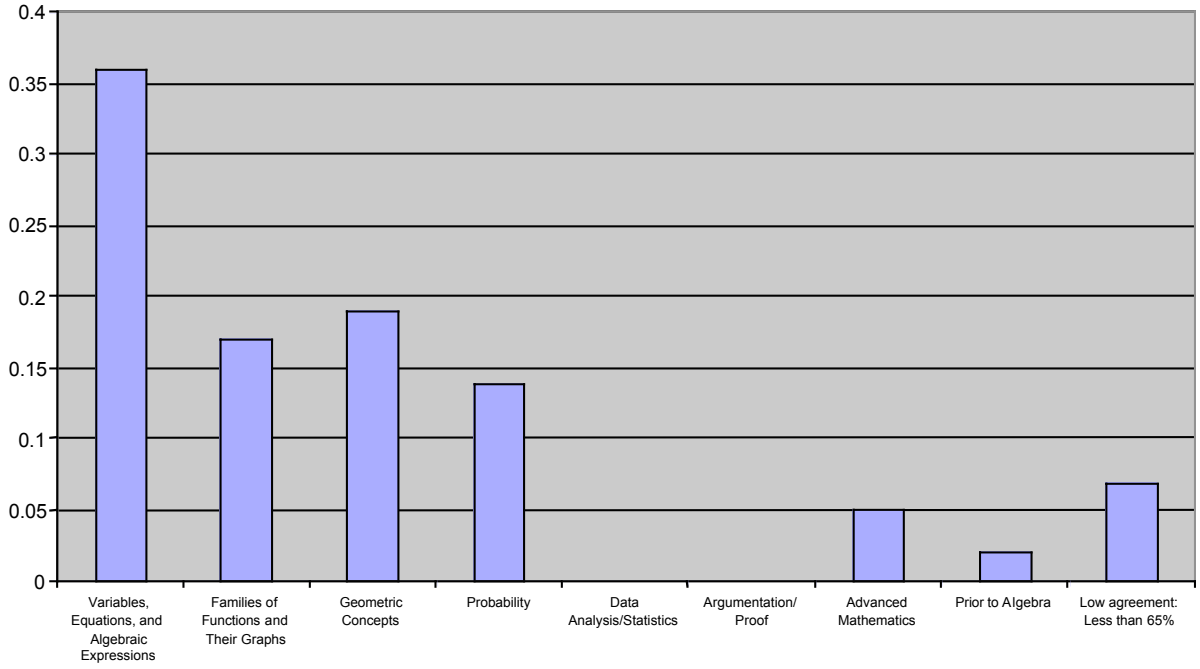

Figure 4. Picture of balance of content coverage for one 6-rater subset: Proportion of the total test addressing each content category. (Figure includes the 34 items for which at least 67% of the raters agreed on the general content category.)

Table 4 shows considerable variability among rater subsets in their level of agreement about item dimensionality without regard to whether they agreed with the decisions made by the 20 raters. Some rater subsets showed very high agreement: They agreed on the item dimensionality for 41 items. Other rater subsets showed much lower agreement: They agreed on item dimensionality for only 20 items. The overall picture of dimensionality for the test varied greatly across rater subsets. Some rater subsets perceived as many as 97% of items to be unidimensional whereas other rater subsets perceived 85% of the items to be multidimensional (on average, rater subsets perceived 48% of items to be unidimensional and 52% of items to be multidimensional).

Assessment of Items' Depth of Knowledge (6-Rater Subsets)

We analyzed variability across rater subsets in multiple ways. First, consistent with the analyses of topic assignment and item dimensionality just presented, we treat depth of knowledge as having discrete levels and require that 67% of raters in a subset agree on the depth-of-knowledge rating to classify an item. Table 3 shows substantial variation in rater subset success in matching the depth-of-knowledge

decisions made by the 20 raters. Some rater subsets matched the 20 raters' decisions on 26 items whereas other rater subsets matched the 20 raters' decisions on only 14 items, only a third of the test.

Table 4 shows considerable variability among rater subsets in their level of agreement about depth of knowledge without regard to whether they agreed with the decisions made by the 20 raters. Some rater subsets showed very high agreement: they agreed on depth of knowledge for all 42 items. Other rater subsets showed less agreement: they agreed on depth of knowledge for 19 items, less than half of the test. The overall picture of depth of knowledge for the test also varied considerably across rater subsets. Some rater subsets judged that over half of the test (52% of the items on which they reached agreement) had the lowest level of depth of knowledge, whereas other rater subsets judged that very few of the items (13%) were at a low level (on average, rater subsets judged that a third of the items, 33%, were at a low level).

Treating depth of knowledge as a continuous variable on a scale from 1 to 4, we computed an average depth-of-knowledge rating for the entire test for a rater subset using all six raters' ratings on all items. Across the 14,400 rater subsets, the mean depth-of-knowledge rating for the test was 1.62 (SD = .07), and the range was 1.43 to 1.87. This approach suggested some variation among rater subsets in terms of the average depth of knowledge of the test. We also carried out analyses to gauge the variability among rater subsets in how they perceived the distribution of depth of knowledge across items on the test. For each rater subset, we calculated the proportion of the 42 items that they declared as having low depth of knowledge (mean rating for an item between 1.0 and 1.49) and the proportion of items that they declared as having medium or high depth of knowledge (mean rating 1.50 or above; because very few items were ever classified as having high depth of knowledge, we did not distinguish between medium and high in these analyses). Across the 14,400 rater subsets, the mean percent of the test designated as low depth of knowledge was 37% (SD = 7), with a range of 10% to 60%. These results show large variation across rater subsets, with some rater subsets judging much of the test to be low level and others judging most of the test to be at a medium level or higher. Different selections of rater subsets, then, could lead to quite different conclusions about the depth of knowledge of the test.

Assessment of Items' Centrality (6-Rater Subsets)

As with depth of knowledge, we analyzed variability across rater subsets in multiple ways. First, consistent with the analyses of topic assignment and item dimensionality just presented, we treat item centrality as having discrete levels and consider only items for which 67% of raters agreed. Table 3 shows substantial variation in rater subset success in matching the item centrality decisions made by the 20 raters. Some rater subsets matched the 20 raters' decisions on 31 items whereas other rater subsets matched the 20 raters' decisions on only 11 items, less than a third of the test.

Table 4 shows great variability among rater subsets in their level of agreement about item centrality regardless of whether they agreed with the decisions made by the 20 raters. Some rater subsets agreed about item centrality on all 42 items; other rater subsets reached agreement on a fraction of the items (38%). The overall picture of item centrality for the test varied considerably across rater subsets as well. Some rater subsets classified all items as highly central, whereas others classified only about half (48%) as highly central (M = 95%, SD = 6).

Second, we consider item centrality as a continuous variable on a scale from 1 to 3 and include all raters' ratings, regardless of agreement. For each rater subset, we computed an average centrality rating for the entire test using all six raters' ratings. Across the 14,400 rater subsets, the mean centrality rating for the test was 2.69 (SD = .12), and the range was 2.30 to 2.98, showing some variability across rater subsets in their assessment of average item centrality. We also carried out analyses to gauge the variability among rater subsets in how they perceived the distribution of centrality across items on the test. For each rater subset, we calculated the proportion of the 42 items that they declared as having high centrality (mean rating for an item between 2.5 and 3.0) and the proportion of items that they declared as having medium or low centrality (mean rating below 2.50; because very few raters ever classified an item as having low centrality, we did not distinguish between medium and low in these analyses). Across the 14,400 rater subsets, the mean percent of test items designated as high centrality was 84% (SD = 5) and the range was 31% to 100%. These results show large variation across rater subsets, with some rater subsets judging most or all of the items being highly central and others judging only a small fraction of the test as highly central. Different selections of rater subsets, then, could lead to quite different conclusions about item centrality.

Summary and Conclusions

Our study convened panels of high school mathematics educators and University faculty to examine the alignment between California's Golden State Examination in High School Mathematics relative to the University of California's *Statement on Competencies in Mathematics Expected of Entering College Students*. Ten educators and 10 faculty members each rated the topic, dimensionality, depth of knowledge, and content centrality of each item on the examination relative to the *Statement of Competencies*. Results provided the context for an in-depth case study of rater agreement and carry implications for the reliability of alignment measures and the factors that may influence it. Study findings also carry practical relevance for assumptions underlying current standards-based reform.

Reliability of Alignment Ratings: Is the Glass Half Empty or Half Full?

The study used multiple approaches to looking at reliability. Starting first with the full panel of 20 raters, results showed that, with modest training, raters achieved relatively high levels of agreement in the identification of specific topics and content categories assessed by individual items. The majority of raters in both groups agreed on the content classification of the great majority of items, and kappa coefficients confirmed moderate to good agreement amongst faculty and teachers on topic and category assignments. As might be expected, kappa coefficients were highest for category ratings, which meant that raters generally were able to differentiate between items addressing content in first-year algebra, second-year algebra, and geometry. Reliability slipped considerably when addressing specific topics within each of those courses. Moreover, study findings demonstrate substantial reductions in reliability as one moves from looking at agreement on a single item feature to that on multiple item features simultaneously. For example, the majority of panelists agreed on how barely one half of the items (22/42) should be classified with regard to topic and depth of knowledge, even though both features are essential to understanding content expectations. While results suggested that the depth-of-knowledge and dimensionality scales needed work to better define and differentiate the various values—analyses showed moderate agreement at best and large standard errors of measurement—study results did provide some empirical verification of the validity of depth-of knowledge-ratings. There was a strong relationship between depth-of-knowledge ratings and student performance on the items.

Using simulated panels of three educators and three faculty members constituted from the full panel, we both compared how the results from panels more typical of current practice would fare relative to our "gold standard" of 20 highly qualified panelists and examined the variability in item ratings across the 14,400 6-rater subsets so constituted. Here results showed considerable variability relative to both dimensions. For example, half of the 6-rater subsets reached agreement on at least 27 of the 30 item-topic correspondences identified by the 20-member panel, but only 7% reached agreement on all 30 items. Results were similar for the content category ratings, with half the subsets reaching agreement on at least 38 of 40 category-item correspondences reached by the full panel. Without regard to the benchmark item-content concordances established by the 20 panelists, the 6-rater subsets, on average, reached agreement on the content addressed by 32 items (although the specific items and content on which they agreed varied across groups), suggesting that the 6-member groups tended to overestimate alignment relative to the gold standard. Moreover, the content-item agreement across the subsets ranged from a low of 23 items to a high of 39 items. All of these analyses suggest the potential for considerable wobble in measures of alignment.

Alignment Measures: Variability and Necessary Limits

Indeed, we saw such wobble in examining two measures of the alignment between a test and a set of standards. The first measure, comprehensiveness of coverage, examined the percentage of topics in each content category and overall that were addressed by at least one item on the test, using only those topic and/or category designations on which there was agreement. Results from the 20-member panel benchmark suggested that the test addressed roughly 40% of the topics in each of four category areas considered essential for entering freshman—Variables, Equations, and Algebraic Expressions, Families of Functions and Their Graphs, Geometric Concepts, and Probability. For Data Analysis and Statistics, comprehensiveness of content coverage dropped to 25% (1 of 4 topics); and no items on the test represented the sixth category, Argumentation and Proof. Overall, the test represented about one third of the topics designated essential. In contrast, the 6-rater subsets, on average, found that 44% of the topics were represented on the test, with a range of 32% to 56%. Moreover, the specific profiles of topic coverage varied considerably across groups. As with the reliability ratings, this represents significant variation, depending on group membership.

The second measure of alignment, balance of coverage, addressed the relative representation of each of the six content categories. Results suggested that the majority of the test addressed algebra I and II and geometry, with nearly a quarter of the items classified each as Variables, Equations, and Algebraic Expressions, Families of Functions and Their Graphs, or Geometric Concepts. Less of the test addressed Probability (14%); Data Analysis and Statistics got scant treatment (2%); and Argumentation and Proof was absent. As with comprehensiveness of coverage, there was considerable variability across the 6-rater subgroups.

In addition to suggesting variability in these alignment measures, although not a major focus of our study, our findings also serve as a reminder of the limits of what and how many topics (or standards) can be addressed by a single test. Even in the case here, where the test spanned two classroom periods and a very modest criterion of "coverage" (a single item) was used, the number of test items severely limits breadth and depth of coverage. How comprehensive, deep, and balanced can an assessment be, given the reality of available testing time? Test purpose and values, as well as practical considerations, should figure in such decisions. Whose values are/should be represented by the alignment of the test? To what extent are decisions made in advance and based on value decisions, as opposed to what items survive an empirical field test or happen to be on an off-the-shelf test? These are some of the many questions that need to be addressed early in the test design process so that specifications firmly aligned with test purpose(s) can be a solid basis for item and test development. Alignment considerations need to precede the test development or selection process, not trail it. Waiting for the results of an after-the-fact alignment study clearly is too late.

Factors That Influence Reliability and Alignment

Our study is limited in identifying factors that influence judgments about alignment, although two key ones can be highlighted. Obvious but worth underscoring: Who does the ratings matters. Study results clearly show the effects of different combinations of raters on alignment. While it could be argued that our results indicate the need for additional rater training to assure more consistent results—and indeed they do—it is also the case that our panelists were highly qualified in terms of content knowledge and prior experience with K-12 mathematics to engage in the rating process.

Another issue related to who does the ratings is the representation of various constituencies in the process. We purposively composed our 6-rater subsets to assure equal representation of high school educators and college faculty, both because we believe that alignment procedures should represent multiple constituencies and because we suspected that these two groups might have different perspectives. Though our data show that faculty and educators' topic ratings were very similar, there were significant differences in how the two groups perceived item dimensionality and depth of knowledge. High school educators tended to see complexity in terms of rating items as multiple dimensional and higher in depth of knowledge where faculty saw simplicity—a single topic and lower depth of knowledge. Perhaps these differences reflect experience with students at different levels of developing expertise. That is, at the high school level, the students with whom educators interact are novices with regard to algebra and geometry content and need to acquire and learn to integrate various dimensions of subject matter content, whereas for University faculty, the students tend to be more expert and their knowledge more integrated and automated (Glaser & Baxter, 2000). This speculation highlights the reality that the cognitive complexity of any given item may not be fixed, but depends on students' developmental levels and prior instructional experience, making depth of knowledge a slippery and difficult rating.

The depth-of-knowledge ratings also demonstrate the ways in which the method can influence results. For example, treating depth of knowledge as a categorical variable and looking at the extent of agreement for the 20-rater panel, reliability looked at least moderate. Raters agreed on the depth of knowledge of two thirds of the items, and based on these ratings, 26% of the items were considered of medium complexity (a rating of 2 or higher). Treating depth of knowledge as a continuous variable and including all raters' ratings, regardless of agreement, revealed only a modest index of dependability and substantial error variance that showed much inconsistency among raters. Yet looking at dependability relative to some threshold (e.g., whether or not depth of knowledge is at least a 2) revealed high consistency. Moreover, averaging depth of knowledge ratings across all raters for each item and then classifying each as high, medium or low based on average rating revealed quite a different distribution of depth of knowledge—by this method 59% of the items were considered medium. The characterization of the test differed, then, according to whether all items and ratings were analyzed or only those items reaching our threshold of agreement.

Practical Implications

The challenge of assuring consistency in item classification and alignment measures has important parallels in bringing current reform ideas to fruition in practice. We noted in the introduction the important lever assessment serves in today's school improvement efforts and the socio-political purposes served by assessment in communicating to teachers and students what is expected so that they can prepare for success. Agreement in the assignment of ratings can be considered an indicator of the extent to which common understandings are shared. Similarly, lack of agreement on the relationship between content topics and items suggests that educators operate with diverse definitions of the meaning of standards in terms of content and depth-of-knowledge expectations. Given uneven understandings, teachers will have difficulty translating expectations into effective classroom teaching and learning experiences. Our findings suggest that even highly experienced educators with solid content credentials can experience difficulty applying standard definitions of content and cognitive demand. True, the majority of our raters were able to agree on what topic was assessed by the majority of items, but that still leaves a significant proportion of items on which there was no such agreement, and a significant minority of raters on each item who viewed the content differently. Moreover, as we noted above, the levels of agreement dipped precipitously when we looked at agreement across multiple item features—for example, content and depth of knowledge—which seem essential if instruction is to be appropriately aligned to standards and assessments. Considering the expertise and experience of the educators and faculty who were involved in this study and the fact that typical classroom teachers may get virtually no training in alignment, our findings may well represent a best case, and suggest the need for substantial action to assure that practicing teachers get the help they need to understand what is expected and to share common expectations.

In summary, our study has identified and demonstrated standard techniques that we believe should be used to assure the measurement quality of alignment measures. Findings from our case study, while clearly limited in generalizability, raise important questions about the reliability of the alignment process and its implications for practice. A challenge for future research and development is the further exploration and solution of these knotty questions.

References

Ananda, S. (2003). Achieving alignment. *Leadership, 33*(1), 18-21.

Bhola, D. S., Impara, J. C., & Buckendahl, C. W. (2003). Aligning tests with states' content standards: Methods and issues. *Educational Measurement: Issues & Practice*, *22*(3), 21-29.

Brennan, R. L. (2001). *Generalizability theory*. New York: Springer-Verlag.

Brennan, R. L., & Kane, M. F. (1977). An index of dependability for mastery tests. *Journal of Educational Measurement, 14,* 277-289.

Buckendahl, C. W., Plake, B. S., Impara, J. C., & Irwin, P. M., (2000, April). *Alignment of standardized achievement tests to state content standards: A comparison of publishers' and teachers' perspectives*. Paper presented at the annual meeting of the National Council on Measurement in Education, New Orleans.

California Department of Education. (1997). *Mathematics content standards for California public schools: Kindergarten through grade twelve.* Sacramento, CA: Author.

California Department of Education. (10/04/01). *High school mathematics blueprint, (*draft). Sacramento, CA: Author.

California Department of Education. (10/16/01). *Overview specifications for high school mathematics* (draft). Sacramento, CA: Author.

Cicchetti, D. V. (1994). Guidelines, criteria, and rules of thumb for evaluating normed and standardized assessment instruments in psychology. *Psychological Assessment*, *6*, 284-290.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement, 20,* 37-46.

Fleiss, J. L. (1971). Measuring nominal scale agreement among many raters. *Psychological Bulletin, 76,* 378-382.

Fleiss, J. L. (1981). Balanced incomplete block designs for inter-rater reliability studies. *Applied Psychological Measurement, 5,* 105-112.

Glaser, R., & Baxter, G. P. (2000). *Assessing active knowledge* (CSE Tech. Rep. No. 516). Los Angeles: University of California, National Center for Research on Evaluation, Standards, and Student Testing (CRESST).

Herman, J. L. (2004). The effects of testing on instruction. In S. Fuhrman & R. Elmore (Eds.), *Redesigning accountability systems for education* (pp. 141-166). New York: Teachers College Press.

Herman, J. L., Webb, N. M., & Zuniga, S. A. (2003). *Alignment and college admissions: The match of expectations, assessments, and educator perspectives* (CSE Tech. Rep. No. 593). Los Angeles: University of California, National Center for Research on Evaluation, Standards, and Student Testing (CRESST).

Koretz, D. M., Barron, S., Mitchell, K. J., & Stecher, B. M. (1996). *Perceived effects of the Kentucky Instructional Results Information System (KIRIS).* Santa Monica, CA: RAND.

Koretz, D. M., Mitchell, K. J., Barron, S., & Keith, S. (1996). *Perceived effects of the Maryland State Assessment Program* (CSE Tech. Rep. No. 409). Los Angeles: University of California, National Center for Research on Evaluation, Standards, and Student Testing (CRESST).

Lane, S., Stone, C., Parke, C., Hansen, M., & Cerillo, T. (2000, April). *Consequential evidence for MSPAP from the teacher, principal and student perspective*. Paper presented at the annual meeting of the National Council on Measurement in Education. New Orleans.

McDonnell, L. M., & Choisser, C. (1997). *Testing and teaching: Local implementation of new state assessments* (CSE Tech. Rep. No. 442). Los Angeles: University of California, National Center for Research on Evaluation, Standards, and Student Testing (CRESST).

No Child Left Behind Act of 2001, Pub. L. No. 107-110, 115 Stat. 1425 (2002).

Olson, L. (2003, Spring). Standards and tests: Keeping them aligned [Entire issue]. *Research Points: Essential Information for Education Policy, 1*(1).

Porter, A. C. (2002). Measuring the content of instruction: Uses in research and practice. *Educational Researcher, 31*(7), 3-14.

Porter, A. C., & Smithson, J. L. (2001). *Defining, developing, and using curriculum indicators* (CPRE Research Rep. Series No. RR-048) . Philadelphia: University of Pennsylvania, Consortium for Policy Research in Education.

Rothman, R., Slattery, J. B., Vranek, J. L., & Resnick, L. B. (2002). *Benchmarking and alignment of standards and testing* (CSE Tech. Rep. No. 566). Los Angeles: University of California, National Center for Research on Evaluation, Standards, and Student Testing (CRESST).

Shavelson, R. J., Baxter, G. P., & Gao, X. (1993). Sampling variability of performance assessments. *Journal of Educational Measurement, 30,* 215-232.

Shavelson, R. J., & Ruiz-Primo, M. A. (2000). On the psychometrics of assessing science understanding. In J. J. Mintzes & J. H. Wandersee, et al. (Eds.), *Assessing science understanding* (pp. 303-341). San Diego, CA: Academic Press.

Shavelson, R. J., Ruiz-Primo, M. A., & Wiley, E. W. (1999). Note on sources of sampling variability in science performance assessments. *Journal of Educational Measurement, 36,* 61-71.

Shavelson, R. J., & Webb, N. M. (1991). *Generalizability theory: A primer*. Newbury Park, CA: Sage Publications.

Stecher, B., Barron, S., Chun, T., & Ross, K. (2000). *The effects of the Washington State Education Reform on schools and classrooms* (CSE Tech. Rep. No. 525). Los Angeles: University of California, National Center for Research on Evaluation, Standards, and Student Testing (CRESST).

Watkins, M. W., & Pacheco, M. (2001). Inter observer agreement in behavioral research: Importance and calculation. *Journal of Behavioral Education, 10,* 205-212.

Webb, N. L. (1997). *Criteria for alignment of expectations and assessments in mathematics and science education* (National Institute for Science Education NISE Res. Monograph No. 6). Madison: University of Wisconsin, Wisconsin Center for Education Research.

Webb, N. L. (1999). *Alignment of science and mathematics standards and assessments in four states* (National Institute for Science Education NISE Res. Monograph No. 18). Madison: University of Wisconsin, Wisconsin Center for Education Research.

Webb, N. L. (2002, April). *An analysis of the alignment between mathematics standards and assessments for three states.* Paper presented at the annual meeting of the American Educational Research Association, New Orleans.

Wixson, K. K., Fisk, M. C., Dutro, E., & McDaniel, J. (2002). *The alignment of state standards and assessments in elementary reading.* Ann Arbor: University of Michigan, Center for the Improvement of Early Reading Achievement.

Appendix A

Specific Topics Considered Essential for All Entering College Freshmen in the *UC Statement on Competencies in Mathematics* (see www.universityofcalifornia.edu/senate/reports/mathcomp.html)

Variables, Equations, and Algebraic Expressions
1. Algebraic symbols and expressions
2. Evaluation of expressions and formulas
3. Translation from words to symbols
4. Solutions of linear equations and inequalities
5. Absolute value
6. Powers and roots
7. Solutions of quadratic equations
8. Solving two linear equations in two unknowns including the graphical interpretation of a simultaneous solution

Families of Functions and Their Graphs
1. Applications
2. Linear functions
3. Quadratic and power functions
4. Exponential functions
5. Roots
6. Operations on functions and the corresponding effects on their graphs
7. Interpretation of graphs
8. Function notation
9. Functions in context, as models for data
10. Polynomials

Geometric Concepts
1. Distances, areas, and volumes, and their relationship with dimension
2. Angle measurement
3. Similarity
4. Congruence
5. Lines, triangles, circles, and their properties
6. Symmetry
7. Pythagorean Theorem
8. Coordinate geometry in the plane, including distance between points, midpoint, equation of a circle
9. Introduction to coordinate geometry in three dimensions
10. Right angle trigonometry

Probability
1. Counting (permutations and combinations, multiplication principle)
2. Sample spaces
3. Expected value
4. Conditional probability

5. Area representations of probability

Data Analysis and Statistics
1. Presentation and analysis of data
2. Mean, median and standard deviation
3. Representative samples
4. Using lines to fit data and make predictions

Argumentation and Proof
1. Mathematical implication
2. Hypotheses and conclusions
3. Direct and indirect reasoning
4. Inductive and deductive reasoning

Appendix B

Detailed Description of Depth-of-Knowledge Levels

Level 1. Recall and Reproduction

At Level 1 is the recall of information such as a fact, definition, term, or a simple procedure, as well as performing a simple algorithm or applying a formula. That is, in mathematics a one-step, well-defined, or straight algorithmic procedure should be included at this lowest level. Assessment items and expectations that require **students** to compute a sum, difference, product, or quotient are considered Level 1 items. Simple word problems that can be directly translated into a number sentence and solved by computation are considered Level 1. Some examples that represent, but do not constitute, all of Level 1 performance are:

1. Recall or recognize a fact, term, or property
2. Compute a sum, difference, product, or quotient
3. Represent in words, pictures, or symbols a mathematical object or relation
4. Provide or recognize a standard mathematical representation for a situation
5. Provide or recognize equivalent representations
6. Perform a routine procedure such as measuring length
7. Evaluate an equation or formula for one of its items

Level 2. Skills and Concepts

Level 2 includes the engagement of some mental processing beyond recalling or reproducing a response. A Level 2 assessment item requires students to make some decisions as to how to approach the problem or activity, whereas Level 1 requires students to demonstrate a rote response, perform a well-known algorithm, follow a set procedure (like a recipe), or perform a clearly defined series of steps. Keywords that generally distinguish a Level 2 item include "classify," "organize," "estimate," "make observations," "collect and display data," and "compare data." These actions imply more than one step. For example, to compare data requires first identifying characteristics of the objects or phenomenon and then grouping or ordering the objects. Some action verbs, such as "explain," "describe," or "interpret," could be classified at different levels depending on the object of the

action. For example, interpreting information from a graph, requiring reading information from the graph, is a Level 2. Item interpreting information from a complex graph that requires some decisions regarding features of the graph that need to be considered and how information from the graph can be aggregated is a Level 3. Other Level 2 activities include making observations and collecting data; classifying, organizing, and comparing data; and organizing and displaying data in tables, graphs, and charts. Some examples that represent, but do not constitute all of Level 2 performance, are:

1. Specify and explain the relationship between facts, terms, properties, or operations
2. Describe and explain examples and non-examples of mathematical concepts
3. Describe how different representations can be used for different purposes
4. Represent a situation mathematically in more than one way
5. Coordinate different representations depending on situation and purpose
6. Select a procedure according to specified criteria and perform it
7. Formulate a routine problem given data and conditions
8. Compare statements such as definitions, examples, or arguments
9. Compare given strategies or procedures
10. Solve a routine problem that requires some interpretation with multiple steps

Provide an *informal* justification of one or more steps in a routine procedure

Level 3. Problem Solving and Strategic Thinking

Level 3 requires reasoning, planning, using evidence, and a higher level of thinking than the previous two levels. In most instances, requiring students to explain their thinking is at Level 3. Requiring a very simple explanation should be a Level 2. Activities that require students to make conjectures are also at this level. The cognitive demands at Level 3 are complex and abstract. The complexity does not result only from the fact that there are multiple answers, a possibility for both Levels 1 and 2, but because the task requires more demanding reasoning. An activity, however, that has more than one possible answer and requires students to justify the response they give would most likely be a Level 3. Other Level 3 activities include

drawing conclusions from observations; citing evidence and developing a logical argument for concepts; explaining phenomena in terms of concepts; and using concepts to solve non-routine problems. Some examples that represent, but do not constitute all of Level 3 performance, are:

1. Analyze similarities and differences between procedures
2. Analyze similarities and differences between problem-solving strategies
3. Formulate an original problem, given a situation
4. Provide *formal* justification for the steps in a solution process
5. Solve non-routine problems
6. Formulate a mathematical model for a complex situation
7. Analyze the assumptions made in a mathematical model
8. Analyze a deductive argument, including proofs of various types

Level 4. Extended Thinking

Level 4 requires complex reasoning, planning, developing, and thinking that will probably require an extended period of time. The extended time period is not a distinguishing factor if the required work is only repetitive and does not require applying significant conceptual understanding and higher order thinking. For example, if a student has to take the water temperature from a river each day for a month and then construct a graph, this would be classified as a Level 2 activity. However, if the student conducts a river study that requires taking into consideration a number of variables, this would be a Level 4.

At Level 4, the cognitive demands of the task are high and the work very complex. Students are required to make several connections—relate ideas within the content area or among content areas—and have to select one approach among many alternatives on how the situation can be solved, in order to be ranked at this highest level. Many on-demand assessment instruments will not include any assessment activities that could be classified as Level 4. However, standards, goals, and objectives can be stated in such a way as to expect students to perform extended thinking. "Develop generalizations of the results obtained and the strategies used and apply them to new problem situations," is an example of a Grade 8 objective that is a Level 4. Many, but not all, performance assessment and open-ended

assessment activities requiring significant thought will be Level 4. Some examples that represent but do not constitute all of a Level 4 performance are:

- Develop a generalization from a mathematical situation
- Apply mathematics in order to model and illuminate a practical problem or situation
- Conduct a project requiring specifying a problem, identifying a number of solution paths, selecting the most effective solution path, solving the problem, and reporting the results
- Prove an original theorem
- Design a mathematical model to inform and solve a practical or abstract situation

Appendix C

Technical Notes on the Generalizability Studies

Absolute error variance for the decision (D) study: $\hat{\sigma}^2_{Abs}$. For the item x rater design, estimated error variance for absolute decisions is:

$$\hat{\sigma}^2_{Abs} = \frac{\hat{\sigma}^2_r}{n'_r} + \frac{\hat{\sigma}^2_{ir,e}}{n'_r}$$

where n'_r is the number of raters in the decision (D) study, here six.

Index of dependability for the decision (D) study: $\hat{\Phi}$. For the item x rater design, the index of dependability is:

$$\hat{\Phi} = \frac{\hat{\sigma}^2_i + \hat{\sigma}^2_{Abs}}{\hat{\sigma}^2_{Abs}}$$

Index of dependability for threshold λ: $\hat{\Phi}(\lambda)$. For the item x rater design, the index of dependability for deciding whether items meet the threshold λ is:

$$\hat{\Phi}(\lambda) = \frac{\hat{\sigma}^2_i + est(\mu - \lambda)^2}{\hat{\sigma}^2_i + est(\mu - \lambda)^2 + \hat{\sigma}^2_{Abs}}$$

where $est(\mu - \lambda)^2 = (\overline{X} - \lambda)^2 - \hat{\sigma}^2(\overline{X})^2$,

$\overline{X}$ = the mean rating over the sample of items and raters in the generalizability (G) study (42 items and 20 raters), and

$$\hat{\sigma}^2(\overline{X}) = \frac{\hat{\sigma}^2_i}{n'_i} + \frac{\hat{\sigma}^2_r}{n'_r} + \frac{\hat{\sigma}^2_{ir,e}}{n'_i n'_r}$$

For the depth-of-knowledge ratings, $\overline{X} = 1.5917$; for centrality ratings, $\overline{X} =$ 2.6690. Using the estimated variance components from Table 2 and $n'_i = 42$ and $n'_r = 6$, $\hat{\sigma}^2(\overline{X}) = .0077$ for depth of knowledge ratings, and $\hat{\sigma}^2(\overline{X}) = .0177$ for centrality ratings.