JOURNAL OF APPLIED MEASUREMENT, *12*(3), 298-309

Selecting Cut Scores with a Composite of Item Types: The Construct Mapping Procedure

Karen Draney

Mark Wilson

University of California, Berkeley

In this paper, we describe a new method we have developed for setting cut scores between levels of a test. We outline the wide variety of potential methods that have been used for such a process, and emphasize the need for a coherent conceptual framework under which the variety of methods could be understood. We then describe our particular method, based on an item response modeling framework, which uses the Wright Map, a graphical model of item and threshold difficulties, and a piece of computer software that provides probabilities of various responses for scores under consideration as cut scores. Finally, we describe a study we conducted for the Golden State Examination in Chemistry, in which we investigate the classification agreement for two groups using the method, and also investigate the reactions of the committee members to the procedure and the software, and the lessons we learned from this process.

Requests for reprints should be sent to Karen Draney, POME, Graduate School of Eductation, University of California, Berkeley, Berkeley, CA 94720-1670, USA, e-mail: kdraney@berkeley.edu.

When tests are given, especially for large-scale or high-stakes purposes, there is often the need to classify the resulting range of scores into a smaller number of (ordered) performance categories. This may be a simple two-category decision (e.g., pass/fail), or a more complex set of decisions (e.g., NAEP's performance levels: Basic, Proficient, Advanced). This involves the choosing of one or more score levels to define the boundaries between the levels.

Traditional procedures for standard setting, such as Nedelsky's (1954), may be appropriate for dichotomously-scored multiple choice tests, but are not necessarily straightforward to extend to tests composed of polytomous items (Cizek, 1996). This observation is even more germane when the test is a composite of both multiple choice and polytomously-scored items, as is the case for many state and test publisher's tests today. Consider an example from the California Department of Education (CDE). The Golden State Examination (GSE) program (discontinued in 2002) was a set of 16 content-based end-of-course examinations offered at the high school level to students who had finished appropriate coursework (e.g., algebra, economics, chemistry, etc.). These examinations were composed of multiple choice items and a variety of open-ended essay and performance items. For several years the CDE developers had used a so-called "mapping matrix" method of setting cut scores where a committee of teachers, subject matter experts and other relevant educators, designed a two-dimensional matrix that maps the total scores on the multiple choice items and the scores on open-ended items onto six "performance levels." This method was seen as a crucial way to bring the judgments of professionals into the arena of test interpretation. The cut score selection committee was expected to use a criterion-referenced framework for setting these performance levels. Thus, the actual frequencies of student scores on the multiple choice items and the open-ended items were not usually shown to the committee until the end of the process. Note that the open-ended items already involved teacher professionals as scorers.The mapping matrix procedure was designed to use professional judgments to balance the technical usefulness of the multiple-choice items with the assumed educational validity of the open-ended items.

While this approach was satisfactory for a number of years, ultimately several problems developed that led the CDE to seek an alternative that provided a more flexible technical solution, and yet preserved the original commitment to relying on teacher professional judgment. The problems were two-fold. First, the committees had a difficult time maintaining a consistent standard from year-to-year using the matrix method. Second, the test specifications were altered, so that, where originally a single (fairly long) open-ended item was thought to be sufficient, the newer design specifications called for several shorter open-ended items, which proved difficult to analyze with the mapping matrix approach.

In this context, a practical solution by Wilson and Draney (2002) was developed, capitalizing on recent technological and conceptual advances in item response modeling. We refer to the method as the "Construct Mapping" method. The underlying conceptual framework is an item response modeling approach. The two item types are scaled together to estimate the best-fitting composite, using either pre-set weights derived from the previous year's matrix-based cut scores, or using weights decided on by the cut score selection committee, or another appropriate committee. We use an item response model for this that allows for pre-set weights for sets of items (Wu, Adams and Wilson, 1998). We will elaborate more on the issue of weights later in the article, but essentially this allows test developers to give sections of the test with different raw scores the same overall weight in determining a student's proficiency level.

This calibration is then used to create an item map combining the locations of the multiple choice items and the thresholds for the score levels of all the performance items. The committee, through a consensus-building process, sets up cut points on this map, using the item response

calibrations to give interpretability in terms of predicted responses to both multiple choice items and open-ended items. Locations of students on the scaled variable are also available for interpretative purposes. This procedure allows both criterion-referenced interpretations of cut scores and norm-referenced interpretations. A piece of software, called "ConstructMap" that displays the item difficulties of the multiple-choice items and the thresholds of the open-ended items together on a Wright map of the variable, and also gives instant calculations of expected responses at chosen points, has been designed, and is used by the committee to help set the cut scores.

In the following sections, we (a) review the literature, with a particular eye toward the type of situation where items are of mixed types, (b) give details of the Construct Mapping technique, and (c) describe a study we have done on the new technique. We conclude with some reflections on what we have learned, and some considerations of future directions.

Background

Methods for selecting cut scores can be sorted into descriptive categories in several ways (Jaeger, 1995). One common method for doing so is to sort them into those focused on individual items and those focused on groups of persons. The traditional procedure used in the GSE program is like neither, because it does not focus on specific groups of examinees, nor does it focus on specific items, but rather on scores of subsets of the items.

An example of a cut score selection method focused on items is the Angoff method (Angoff, 1971), which asks judges to estimate the probability of passing an item for a borderline or minimally competent examinee There are a number of modifications and extensions of this technique (see, for example, Cizek and Bunch, 2007, Chapters 6 and 7). Other such methods include those described by Ebel (1972), Jaeger (1982), and Nedelsky (1954). Item-focused cut score selection methods, particularly those based on the Angoff method, have been criticized as being complex and difficult for the judges involved (e.g., Impara and Plake, 1998). In particular, judges' understanding of what constitutes a "borderline" person is problematic; establishing an explicit understanding in the minds of the judges would seem to assume that the problem the technique is intended to address is already solved. Also, these methods tend to give different results with different sets of judges (Sireci, Robin, and Patelis, 1999).[1]

An example of a cut score selecting method focused on groups of persons is the Borderline Groups method (Livingston and Zieky, 1982), which asks judges to select sets of persons as representing "borderline" persons; the median score of the resulting sets of persons is taken as the cut score. The Contrasting Groups method (Livingston and Zieky, 1982) is another such example. In this method, groups above and below the cut score are selected; the score used as the cutoff is the one that results in the lowest number of misclassifications of persons in the two groups. These examinee-centered methods have also been criticized. The sampling variability of the groups involved is usually unknown; therefore issues such as the reliability with which persons can be classified is also unknown. In addition, the resources involved in finding appropriate groups are often large (e.g., groups must often be chosen using some criterion other than the examination in question) (Sireci, Robin, and Patelis, 1999).

In recent years, a number of new or modified procedures have been developed to cope with the increasingly complex information resulting from examinations that include or are based completely on performance-style items. The data from such examinations is usually complex, consisting of a number of different scores, such as multiple scores derived from a single item, or scores derived from different item types. The methods described in the previous paragraphs were originally developed for items that can be scored right or wrong, and whose scores can be summed to form a single examination score. Several examples of cut score selection methods derived to deal with

1 Although some such variation may be inevitable, and our task will be to *control* the variation, not eliminate it.

data from complex examinations that include written response or performance items rather than, or in addition to, multiple choice items include the Analytic Judgment method (Plake and Hambleton, 2001), the Integrated Judgment procedure (Jaeger and Mills, 2001), and the Multistage Dominant Profile method (Putnam, Pence, and Jaeger, 1995).

Perhaps the most common method described for use with sets of constructed response items, called the "body of work" method (Kingston, Kahl, Sweeney, and Bay, 2001), was developed by Kahl, Crockett, DePascale, and Rindfleisch (1994). In this method, IRT estimates for person proficiency are obtained based on a set of constructed response items. Representational sets of persons across the entire proficiency scale are then chosen, and arranged into folders representing intervals on the scale. Cutoffs are then chosen based on folders, and papers within folder. The result of this method is a scale value (on the IRT logit scale) to be used as a cutoff. Cut score selecting judges, however, use only the papers in the folders; they do not use (or indeed, need to know about) the numerical values of the logit scale.

The method we will describe in this article is a type of item-mapping method (Loomis and Bourque, 2001). Two other examples of this type of method are the method discussed by Wang (2003), and the Bookmark method (Lewis, Mitzel, and Green, 1996; Mitzel, Lewis, Patz, and Green, 2001). In the latter method, items are placed on an achievement continuum using an item response model. A test booklet is then created by arranging the items in order of increasing difficulty. Judges, working individually, place bookmarks in the ordered test booklet where they judge that, in order to be classified into the proficiency level under consideration, an examinee must be able to pass all of the items preceding the bookmark with some pre-specified probability (e.g., 80%). Following this, there may be several rounds of consensus-building, in which judges come to agreement about the locations of the bookmarks.

In addition to new methods of selecting cut scores, recent work has also focused on ways to evaluate and/or augment existing methods with more empirically-based analysis. One example of this is work by Sireci et al. (1999), which uses cluster analysis of the subscores to be combined in the cut score selection process. Cutoff scores are derived from resulting clusters of students. Such a method could be used either as a cut score selecting tool in its own right, or to evaluate cut scores set by other methods. Another method that could be used to evaluate the work of judges selecting cut scores was developed by Engelhard and Anderson (1998). This method, based on the binomial trials model (Wright and Masters, 1981), can be used to assess the variability between judges, and the stability over time, of borderline examinee passing rates for items, such as those used in the Angoff model and its extensions.

One of the major issues that becomes apparent when one reviews studies on cut score selecting methodology is the need for a coherent theoretical framework from which to examine both the methods and results of their use. Several examples of such a framework exist, such as the one given by Reckase (1998). Reckase describes a latent-trait based theoretical framework from which to discuss cut score selecting methods. Scores are assumed to be on a numerical scale that will be used to determine whether a person is above the required cut score. Tasks, methods, judges, and possibly other features are assumed to affect the process. The cut score chosen by a given judge is assumed to be a value on the numerical scale used to measure the persons (often a latent trait scale). This framework requires the explicit specification of a number of assumptions about cut score selecting in general (e.g., are the judgments made by different judges to be regarded as drawn from a distribution representing a single underlying judgment, but made with error, or does each judge have a potentially different latent value representing his or her judgment?).

Another example of such a framework can be built around Rasch models. The history of this is given by Stone (2009), and includes work by such researchers as Wright and Grosse (1981), Stone (1995), Wang (2003), and others. Our method is based on a Rasch-based underlying framework,

uses many of the Rasch model's unique features, and makes the process of the selection of a cut score on the latent trait scale explicit, rather than inferred.

The Traditional GSE Procedure: The Mapping Matrix

The GSE program in the state of California consisted of a set of high school honors examinations. These were end-of-course examinations in a number of subjects, including mathematics (Algebra, Geometry, High School Mathematics), language (Reading and Literature, Written Composition, Spanish Language), science (Physics, Chemistry, Biology, Coordinated Science), and social science (US History, Government and Civics, Economics). Each examination consisted of a set of multiple choice items and at least one written response item.

Based on the GSE, examinees were categorized into one of six performance levels—descriptive categories of student performance in each subject (see general performance levels in Figure 1). Each individual subject had its own more detailed performance descriptions. The top three levels (4, 5, and 6) were considered "honors" levels (School Recognition, Honors, and High Honors, respectively). If a student achieved one of these honors levels on six exams (including US History, Reading and Literature or Written Composition, a mathematics exam, and a science exam), the student was also eligible for a State honors diploma. Thus, the procedure for GSE consisted of five non-independent standard settings.

As mentioned above, the correspondence between scores on the various item types of an examination and each of the six performance levels was established by a cut score selection committee. The committee consisted of approximately ten people who had been part of the GSE process in some way. These may have included persons who were part of the test development team, leaders for the scoring of the written response section (who may also have been involved in the scoring of open ended items, selection of exemplar papers known as "anchor papers," and other processes), raters who had scored the written response section, and other persons involved in some way in the GSE program.

Committees met for a day-long session, during which time they reviewed all of the specific test materials: The test guide developed for teacher use (which includes the performance

Level 6 (High honors)	Student work demonstrates evidence of rigorous and in-depth understanding of key ideas
Level 5 (Honors)	Student work demonstrates evidence of solid and full understanding of key ideas
Level 4 (School Recognition)	Student work demonstrates evidence of substantial understanding of key ideas
Level 3	Student work demonstrates evidence of a basic understanding of key ideas
Level 2	Student work demonstrates evidence of limited understanding of key ideas
Level 1	Student work demonstrates little or no evidence of understanding of key ideas

Figure 1. Descriptions of general GSE performance levels.

level descriptions), the multiple choice items, the written response items with their scoring guides, and examples of student responses at each score level (referred to as *anchor papers*). They spent about 30 minutes looking through this material individually. Once they had completed these familiarity exercises, they were then shown a two-way matrix, known as a *mapping matrix*, with rows defined by all possible scores on the multiple choice section and columns defined by all possible scores on the written response section of the examination. The committee was then asked, for each possible score combination (represented by a cell of the matrix), to determine the most appropriate performance level. Group members discussed their differing opinions, led by a committee chair. Comments and topics of discussion tended to focus on total scores and test-specific issues (e.g., "The second written response item was quite difficult this year, so students shouldn't need to get such a high written response score to get a performance level of 4" or "The first 10 multiple choice items are about basic knowledge that is covered by every class, so students should have to get a multiple choice score higher than 10 to receive a performance level higher than 1"). At the end of the discussion, if consensus had not been reached, the cutoffs were decided by majority vote.

While this approach had been satisfactory for a number of years, several problems developed that led the CDE to seek an alternative that provided a more flexible technical solution, and yet preserved the original commitment to relying on teacher professional judgment. These included changes in the testing program to include more open-ended items (thus complicating the two-dimensional matrix), and technical difficulties with year-to-year equating. These led us to develop a new method.

The Construct Mapping Approach

The Construct Mapping approach to the problem of cut score selection in a mixed item type situation is designed to reduce the amount of arbitrariness in the traditional cut score selecting process, without losing the essential element of human judgment on which it was based. The new method allows committee members to use the item response scale as a guide to what a student at a given level knows and can do. The modeling technique used is a Rasch-family item response modeling approach, where the two item types are scaled together using judged weights to estimate the best-fitting composite (Wu, Adams and Wilson, 1998). The weights are decided beforehand in one of two ways: (a) the test construction committee decides on weights based on criteria such as (i) amount of time allocated to each section, or (ii) judged importance of each section, or (iii) a combination of both; (b) the weights are decided from an earlier mapping matrix results using a form of regression (see Draney, Wilson, and Hoskens, 2000).

A weighted calibration (Wu, Adams, and Wilson, 1998)[2] is then used to create an item map combining the locations of the multiple choice items and the score levels of the performance items. An example showing a section of such an item map for a test with two written response items is given in Figure 2. The column on the far left contains a numerical scale that allows the selection and examination of a given point on the map, and the selection of the eventual cut scores for the performance levels. This scale is a transformation of the original logit scale, designed to have a mean of 500, and to range from 0 to approximately 1000; we refer to the scale in this example as the GSE scale. The next two columns contain the location of the multiple choice items (labeled by number of appearance on the examination form), and the probability that a person at a given point would get each item correct (in this case, a person at 500 on the GSE scale). The next four sets of columns display the thresholds for the components of the written response item; in the case of Chemistry, a lab exercise scored on four different aspects of performance referred to as "components" (each component is scored on a scale of 1 to 4 on this particular examination), and the probability that a person at 500 on the GSE scale would score at that particular score

2 A similar calibration is described in considerable detail in Wilson and Wang (1996).

level on each item. The software also displays, for a person at the selected point on the GSE scale, the expected score total on the multiple choice section and the expected score on each of the written response items (the Figure does not show this part of the display). Note that, using this software, it is possible to give different sections of the test different overall weights (weights being the percentage of the total test score accounted for by the points on a particular section). On this test, the multiple choice section accounts for 40% of the total and the lab exercise for 60%; the various scores on the lab exercise also have different weights.

The cut score selection committee undergoes a set of exercises designed to familiarize them with the content of the test for which standards are being set. Then the committee is led through a series of exercises to help the members understand the use of the Construct Mapping map as shown in Figure 2, and to familiarize them with the ConstructMap software. First, the multiple choice portion only of the map is displayed, and the meanings of the item difficulties explained; then the item thresholds are added and explained. After this, the software is introduced, along with the set of commands used to operate it. A worksheet with example proficiency levels is

GSE Chemistry, Spring 2000 (*N* = 12,596)

MC total : Component total = 40:60; Component weights 20:25:35:20

	Items									
GSE	Multiple Choice		Component 1		Component 2		Component 3		Component 4	
Scale		Prob		Prob		Prob		Prob		Prob
680							3.4	.00		
670										
660										
650					2.4	.03				
640	7	.32								
630										
620	20	.34								
610	24	.36								
600	4 14 29	.37								
590	12 31	.38								
580	32	.40							4.3	.20
570	18 34	.41					3.3	.13		
560										
550	5	.43								
540										
530										
520	6 19	.47			2.3	.34				
510	13 17 27	.49								
500	28	.50								
490	33 35	.52								
480	9 15 26	.53								
470	23	.54								
460										
450	3	.56	1.3	.72						
440	25 30	.58								
430	8 10 16 22	.59								
420										
410									4.2	.64
400										
390	11	.65					3.2	.81		
380	21	.66			2.2	.56				
370										
360	2	.68								
350										
340									4.1	.13

Figure 2. Illustration of a cut score setting item map.

passed out, and committee members try out the commands with those proficiency levels and experiment with others.

The display of multiple choice item locations in ascending difficulty, next to the written response thresholds, helps to characterize the scale in terms of what increasing proficiency "looks like" in the pool of test-takers. For example, if a committee was considering 500 (see Figure 2) as a cut score, then they could note that it is a point at which items like 28 are expected to be answered correctly about 50% of the time, a harder item like 7 is expected to be answered correctly about 30%, and easier items like 2 are expected to be answered correctly 68% of the time. The set of multiple choice items, sorted so they are in order of ascending difficulty, is available to the committee so that the members can relate these probabilities to their understanding of the items. The committee could also note that a student at that point (i.e., 500), would also have about a 34% chance of scoring a 3 on the second written response item and a 64% chance of scoring a 2 on the fourth. Examples of student work at these levels are available to the committee for consideration of the interpretation of these scores.

The committee then, through a consensus-building process, sets up cut points on this map, using the item response calibrations to give interpretability in terms of predicted responses to both multiple choice items and open-ended items. Locations of students on the scaled variable are also available for interpretative purposes. This is an intensive process involving individuals or pairs choosing initial sets of cut points, and then rounds of intensive discussion until all committee members are satisfied with the result.

Use of the maps available from the item response modeling approach not only allows the committees to interpret cut-offs in a criterion-referenced way, it also allows maintenance of similar cut scores from year to year by equating of the item response scales.

We can also allow the committee to consider a variety of weights as they set cut-offs for the performance levels. ConstructMap allows the weights to be dynamically varied. Because recalibration would be too slow in real-time, we do this by performing a series of calibrations beforehand with relative weights spanning the expected range. On the day of the standard setting, we use interpolation between results from adjacent pairs of relative weights to produce approximations to the calibrated maps. There are enough pre-calibrations done to make this sufficiently accurate for our purposes—effectively that means that the maps we produce by the approximation do not differ from those produced by a fresh calibration.

Lastly, impact data (in the form of number and percent of students at a given point on the scale, and in the form of cumulative percentiles) can be shown. This is generally done after the original rounds of standard setting are completed, in order that the initial decisions may be made on a more criterion-referenced basis. This step is not an essential part of our process, but as we were working with an established set of examinations with considerable history, it was added to prevent unacceptably large shifts in test score meaning as committees became used to the new procedure.

The Chemistry Study

We now turn to a description of our study of the method and accompanying software. This study, undertaken with the 2000 Chemistry GSE, included two committees, both of whom participated in the Construct Mapping process. The first of the two sessions was a "live" standard setting session (i.e., the cut scores were used for assigning performance levels in 2000), and the second a replication for research purposes only. The first committee was selected to be the "live" standard setting committee, and the second committee was made up of persons who would have been eligible to participate in the official standard setting, but had not been selected. In the case of the Chemistry GSE, as in the other sciences, the written response component has traditionally been a single, 45-minute long laboratory task, which has been assigned a single holistic score. However, in 2000, the laboratory tasks were scored on four separate components using four separate scoring guides, some on a scale of 1 to 3, and some on a scale of 1 to 4 (see Wiley, 2000, for a descrip-

tion of these scales). Thus, the traditional matrix method was no longer a practical way to set cut scores for these tests.

As part of both the live procedure and the replication study in Chemistry, committees were shown impact data, both in the form of cumulative percentiles and in the form of the resulting numbers of students at each of the six performance levels, for both the current and the previous year, and allowed to adjust their cutoffs if they wished. This comparison and adjustment phase had long been a characteristic of live standard setting sessions.

During a number of pilot studies of the method and accompanying software (see Draney, Wilson, and Hoskens, 2000), we learned several things that were useful in the design of our study. First, various aspects of the leadership and composition of the standard setting group can potentially influence the outcome of a given standard setting session (e.g., more experienced leadership leads to more group satisfaction; more teaching experience in group members leads to more realistic results), so it is important to keep as many aspects as possible of the group composition and leadership constant throughout any study of the method. Thus, both committees were led by the same person, a member of the development team.

Second, although at one point we set the relative weights of the different sections of the test (e.g., written response/lab section vs. multiple choice), and although the software is able to display the results of changing the relative weights on the resulting maps, we found that (a) the committee used only substantive considerations when deciding on weights and (b) they often found the effect of the weights on the maps confusing. We therefore set the relative weights of the various portions of the exams (both for the overall lab task versus multiple choice, and for the various components of the lab task relative to each other) in a separate meeting prior to the standard setting session. These weights are shown in Table 1.

Results

The two committee groups in the Chemistry study showed a very high rate of agreement, with 92% of all students classified into the same performance level by both committee groups. The second group classified the remaining 8% of students one level higher than did the first group. These proportions are shown in Table 2. In this table, the two committee groups are referred to as "group 1" and "group 2"; the rows of the table show the proportions of students who were classified 2 performance levels higher, 1 higher, the same, 1 lower, and 2 lower, by the cut scores chosen by the first committee group as compared to the second. The proportions of students classified into each performance level overall by the two groups are shown in Table 3.

One might ask whether the similarity between these two groups had to do with adjustments made once the groups had been shown impact data, and compared this data to the number of students receiving the various performance levels in the previous year. However, inspection of the original and final sets of performance levels (before and after the impact data were made available) shows that the judgments made by the groups were actually more similar prior to inspection of impact data. Four of the five original cutoffs were identical for the two groups; the final cutoff differed by only one unit. This would have resulted in only 0.7% of students being classified differently by the two groups.

Table 1

Relative weights for the different item types from the Chemistry study

Total points	multiple choice[a]		component 1		component 2		component 3		component 4	
m.c./w.r.	weight	points	weight	points	weight	points	weight	points	weight	points
35/15	1.00	35.00	3.50	10.50	3.28	13.125	4.59	18.375	2.63	10.5

Note: m.c. = multiple choice; w.r. = written response section
[a] Used as a reference category.

Discussion

This study, although small, has provided information about the cut score selecting process as it was used in the GSE program, and also useful feedback about new developments for the process. Through this and other previous studies with the GSE, we have learned about the factors that affect the decisions about performance level cutoffs that these committees make, including the effect of changes in attitude over time, effects of different styles of leaders, and effects of differences in teaching and GSE experience.

The agreement between the two groups in this study looks especially good when compared with the rate of agreement between GSE standard setting groups obtained from previous pilot studies in which the traditional method was also studied. In these pilot studies, the proportion of students classified into the same performance level was never higher than 75% (Draney, Wilson, and Hoskens, 2000). The Construct Mapping procedure shows promise as a cut score selection method in this context.

Through interviews with the individual committee members after the use of this method, we have learned that most committees and individual committee members felt more comfortable with the expected scores associated with a given proficiency level than with the probabilities of passing individual items displayed in the ConstructMap software, especially for multiple choice items; only one member of one committee discussed the probability of passing individual multiple choice items. Most committee members also took into consideration that, given that the multiple choice items had four possible responses, an expected score of one-fourth of the total number of multiple choice items would be the average score expected by someone who was guessing on all of the items. They thus felt that a student should have a multiple choice score higher than this to receive a performance level higher than one. In both committees, some members felt that the expected multiple choice scores were the more important of the two kinds of scores, as they reflected breadth of knowledge, while other members felt that the written response items were more important, as they reflected depth of understanding.

Based both on the commentary of the committee members, and on feedback from several observers of the process, it has become clear that a more comprehensive training process needs to be designed. We are in the process of developing this, emphasizing greater comprehension of the criterion-referenced meaning of locations on the item map, as well as designing an improved software package; one which we hope will operate more smoothly and intuitively.

For example, we plan to add a facility to the ConstructMap program that will display the locations on the map of individual score combinations (e.g., 2 scores of 4 on the written response items and a total of 25 on the multiple choice section). Committee members have expressed uneasiness

Table 2

Proportions of matches and measures of similarity

	Proportion
1st group 2 lower	0.00
1st group 1 lower	0.08
both groups same	0.92
1st group 1 higher	0.00
1st group 2 higher	0.00
Kappa	0.90

Table 3

Comparison of percentages classified into performance levels

		Group 1 (live)		Group 2 (replication)	
N=12,596	Level	Percent	Cumulative	Percent	Cumulative
	6	3.79%	3.79%	5.15%	5.15%
	5	8.72%	12.51%	12.02%	17.17%
	4	17.54%	30.05%	15.05%	32.22%
	3	29.47%	59.52%	27.29%	59.52%
	2	29.87%	89.39%	29.87%	89.39%
	1	10.61%	100.00%	10.61%	100.00%

at not knowing exactly where on the map—and hence in which performance level—certain score combinations would appear. We would also like to incorporate certain aspects of the method reported by Kahl, et al. (1994), in which the logit scale is made concrete through the use of folders containing sets of actual examinee papers representing various points on the student proficiency scale.

Finally, we have developed a facility for entering each multiple choice item along with its correct response, and written response items along with sets of written response work for each of the associated score levels for each item, making it possible for committee members to see these sets of written responses, along with actual multiple choice items, as they are working with the software.

Acknowledgements

We wish to acknowledge the generous support of the California Department of Education, the Sacramento County Office of Education, and the many teachers and other educational professionals without whose support these studies could not have been accomplished. Any errors or omissions are the responsibility of the authors.

References

Angoff, W. H. (1971). Scales, norms, and equivalent scores. In R. L. Thorndike (Ed.), *Educational measurement* (2nd ed., pp. 508-600). Washington, DC: American Council of Education.

Cizek, G. J. (1996). Cut score-setting guidelines. *Educational Measurement: Issues and Practice, 15*, 13-21.

Cizek, G. J., and Bunch, M. B. (2007). *Standard setting: A guide to establishing and evaluating performance standards on tests.* Thousand Oaks, CA: Sage.

Draney, K., Wilson, M., and Hoskens, M. (2000, April). *Standard setting on a composite of item types: The "standard mapping" procedure.* Paper presented at the annual meeting of the American Educational Research Association, New Orleans, LA.

Ebel, R. L. (1972). *Essentials of educational measurement.* Englewood Cliffs, NJ: Prentice Hall.

Engelhard, G., and Anderson, D. W. (1998). A binomial trials model for examining the ratings of cut score-setting judges. *Applied Measurement in Education, 11*, 209-230.

Impara, J. C., and Plake, B. S. (1998). Teachers' ability to estimate item difficulty: A test of the assumptions in the Angoff standard setting method. *Journal of Educational Measurement, 35*, 69-81.

Jaeger, R. M. (1995). Setting cut scores for complex performances: An iterative, judgmental policy-capturing strategy. *Educational Measurement, Issues, and Practice, 14*, 16-20.

Jaeger, R. M., and Mills, C. M. (2001). An integrated judgment procedure for setting standards on complex, large-scale assessments. In G. J. Cizek (Ed.), *Setting performance standards: Concepts, methods, and perspectives* (pp. 313-338). Mahwah, NJ: Lawrence Erlbaum.

Jaeger, R. M. (1982). An iterative structured judgment process for establishing cut scores on competency tests: Theory and application. *Educational Evaluation and Policy Analysis, 4*, 461-475.

Kahl, S. R., Crockett, T. J., DePascale, C. A., and Rindfleisch, S. L. (1994, June). *Using actual student work to determine cutscores for proficiency levels.* Paper presented at the CCSSO national conference on Large-Scale Assessment, Albuquerque, NM.

Kingston, N. M., Kahl, S. R., Sweeney, K. P., and Bay, L. (2001). Setting performance standards using the body of work method. In G. J. Cizek (Ed.), Setting *performance standards: Concepts, methods, and perspectives* (pp. 219-248). Mahwah, NJ: Lawrence Erlbaum.

Lewis, D. M., Mitzel, H. C., and Green, D. R. (1996, June). *Standard setting: A bookmark approach.* Paper presented at the CCSSO national conference on Large-Scale Assessment, Phoenix, AZ.

Livingston, S. A., and Zieky, M. J. (1982). *Passing scores: A manual for setting cut scores of performance on educational and occupational tests.* Princeton, NJ: Educational Testing Service.

Loomis, S. C., and Bourque, M. L. (2001). From tradition to innovation: Standard setting on the National Assessment of Educational Progress. In G. J. Cizek (Ed.), *Setting performance standards: Concepts, methods, and perspectives* (pp. 175-218). Mahwah, NJ: Lawrence Erlbaum.

Mitzel, H. C., Lewis, D. M., Patz, R. J., and Green, D. R. (2001). The bookmark procedure: Psychological perspectives. In G. J. Cizek (Ed.), *Setting performance standards: Concepts, methods, and perspectives* (pp. 249-282). Mahwah, NJ: Lawrence Erlbaum.

Nedelsky, L. (1954). Absolute grading practices for objective tests. *Educational and Psychological Measurement, 14*, 3-19.

Plake, B. S., and Hambleton, R. K. (2001). The analytic judgment method for setting standards on complex performance assessments. In G. J. Cizek (Ed.), *Setting performance standards: Concepts, methods, and perspectives* (pp. 283-312). Mahwah, NJ: Lawrence Erlbaum.

Putnam, S. E., Pence, P., and Jaeger, R. M. (1995). A multi-stage dominant profile method for setting cut scores on complex performance assessments. *Applied Measurement in Education, 8*, 57-83.

Reckase, M. D. (1998, April). *Analysis of methods for collecting test-based judgments.* Paper presented at the annual meeting of the National Council on Measurement in Education, San Diego, CA.

Sireci, S. G., Robin, F., and Patelis, T. (1999). Using cluster analysis to facilitate standard setting. *Applied Measurement in Education, 12*, 301-325.

Stone, G. E. (1995). *Objective standard setting.* Unpublished doctoral dissertation, The University of Chicago, 1995.

Stone, G. E. (2009). Introduction to the Rasch family of standard setting methods. In E. V. Smith, Jr. and G. E. Stone, (Eds.), *Criterian referenced testing: Practice analysis to score reporting using Rasch measurement models* (pp. 138-147). Maple Grove, MN: JAM Press.

Wang, N. (2003). Use of the Rasch IRT model in standard setting: An item-mapping approach. *Journal of Educational Measurement, 40*, 231-253.

Wiley, D. (2000). *Final report on GSE component scoring pilot study.* Evanston, IL: Northwestern University Research Report.

Wilson, M., and Draney, K. (2002). A technique for setting standards and maintaining them over time. In S. Nishisato, Y. Baba, H. Bozdogan, and K. Kanefuji (Eds.), *Measurement and multivariate analysis* (pp. 325-332). New York: Springer-Verlag.

Wilson, M. and Wang, W.-C. (1996). Complex composites: Issues that arise in combining different modes of assessment. *Applied Psychological Measurement, 19*, 51-71.

Wright, B. D., and Grosse, M. E. (1981). *Part II item bank and standard setting study.* Philadelphia, PA: National Board of Medical Examiners.

Wright, B. D., and Masters, G. N. (1981). *Rating scale analysis.* Chicago: MESA Press.

Wu, M., Adams, R. J., and Wilson M. R. (1998). ACER-ConQuest [Computer program and manual]. Melbourne, VIC, Australia: ACER Press.