DOCUMENT RESUME

ED 476 425 TM 034 927

AUTHOR DeMauro, Gerald E.

TITLE Developing a Theory of Performance: A Two-Stage Structure for the Psychology of Standard Setting.

PUB DATE 2003-04-00

NOTE 36p.; Paper presented at the Annual Meeting of the National Council on Measurement in Education (Chicago, IL, April 22-24, 2003).

PUB TYPE Reports - Research (143) -- Speeches/Meeting Papers (150)

EDRS PRICE EDRS Price MF01/PC02 Plus Postage.

DESCRIPTORS *Cognitive Processes; *Judges; Matrices; *Performance Based Assessment; Psychology ; *Standard Setting (Scoring)

ABSTRACT

An analysis was made of the cognitive processes that support the judgments made in standard setting activities. These processes were conceived as having two components: forming the domain needed to pass the test and identifying the criterion level of performance to pass the test. In fact, these processes are interactive, and were separated for the purposes of analysis and study. The judgments for the first component can be conceived of as a matrix in which the test questions provide information about information about whether or not their content is consistent with the test requirements. The second component can be conceived as a different type of matrix in which each test question is judged against a criterion value of difficulty. Analysis of judgments made by 38 panelists contributing to a mock standard setting suggests that the participating judges were consistently able to judge and use the judgments regarding the attributes of test items with respect to their conception of the required domain. They were also able consistently to judge the difficulty of the items against their criteria for passing the test. These findings suggest that this type of analysis may prove useful for future application in standard setting, and, certainly, for continued study of the process. (Author/SLD)

Developing a Theory of Performance:

A Two-Stage Structure

For the Psychology of Standard Setting

Gerald E. DeMauro,
New York State Education Department,
Coordinator of State Assessment
April, 2003

Revision of Paper presented as part of a symposium at the annual meeting of the National Council on Measurement in Education, Chicago, Ill., April, 2003.

Abstract

An analysis is made of the cognitive processes that support the judgments made in standard setting activities. Specifically, these processes are conceived as having two components: forming the domain needed to pass the test, and identifying the criterion level of performance to pass the test. In fact, these processes are interactive and are separated for the purposes of analysis and study.

The judgments for the first component can be conceived of as a matrix in which the test questions provide information about whether or not their content is consistent with the test requirements. The second component can be conceived as a different type of matrix in which each test question is judged against a criterion value of difficulty.

Analyses of judgments made by panelists contributing to a mock standard setting suggests that the participating judges were consistently able to judge and use the judgments regarding the attributes of test items with respect to their conception of the required domain. They were also able to consistently judge to difficulty of the items against their criteria for passing the test. These findings suggest that this type of analysis may prove useful for future application in standard setting, and, certainly, for continued study of the process.

Developing a Theory of Performance: A Two-Stage Structure

For the Psychology of Standard Setting

Gerald E. DeMauro,
New York State Education Department,
Coordinator of State Assessment

Overview

Anytime judgments are made about the magnitude of some stimulus, a conversion takes place between the physical properties and the perceived properties of the stimulus. Early work in psychophysical scaling attempted to find mathematical relationships to characterize the conversion. Standard setting lends itself to psychophysical analysis because the judges are asked to make a determination that involves perceived changes in item difficulty or in examinee skills in relation to more objective measures of those attributes. Standards are set by expert judges not by deciding where they believe cut scores should be placed, but rather, by degrees of judgments comparing the demand of test questions to the attributes possessed by hypothetical just passing examinees, or to the demands of test questions. Each of these judgments, in turn, reflects the theory constructed by the judge of minimally acceptable performance.

The Two Matrices

This study proposes to help our understanding of the cognitive processes underlying the standard setting judgments. In particular, the study breaks down these cognitive process into two interactive components: the first involves defining the skills and knowledge necessary for

passing the test, and the second compares test questions or examinees to determine whether or not they surpass the required characteristics.

These judgments have been described in studies of psychophysical scaling as matrices comparing perceptions of what is required of an examinee or test question with observed data concerning examinee performance or question content and difficulty. Specifically, the two components of judgments made in standard setting represent a proximity matrix describing how close the observation is to the psychological model, and a dominance matrix describing whether the observation exceeds or falls short of the psychological model.

Coombs, Dawes and Tversky (1970) describe a proximity matrix as follows:

> "All instances of labeling are examples of proximity relations between objects in distinct sets...In each such instance the correspondence could be between a set of points corresponding to the response categories and a set of points corresponding to the objects to be classified. The matching of a point in one set with a point in the other reflects their relative 'nearness' or proximity." (p. 34)

In analyzing the judgments made in standard setting activities, the first set of judgments are definitional. Through analyses of the skills of students or the content of test questions, subject matter experts give increasing operational meaning to required learning standards and component performance indicators. When presented with examples of student performance, and in particular, with the demands of the test questions, they judge how well these observed features of the test questions and student performance fit their theories, and they adjust the theories as needed. Once a proximity matrix is constructed from this process, judges can evaluate which items or examinees surpass their performance criterion for passing. These evaluations determine the dominance matrix.

In measurement, examples of dominance matrices are readily available. Examinees form the rows of the matrix while items or stimuli form the columns. For standard setting, the rows

would be just passing examinees or the minimum required domain of knowledge and skills to pass, while the columns would be test items (Coombs et. al., 1970). The judgment of dominance is made on the sufficiency of the examinees' latent abilities in the content domain needed to pass the test.

Each cell of the dominance matrix, then, could contain a one or a zero depending upon whether the examinee dominated the item (one) or was dominated by the item (zero). Clearly, the characterization is extended to items in partial credit models by separating the possible points of performance and examining the relationship between examinees of defined levels of skills at each score point value (cf. Master, 1982; Linacre & Wright, 1998).

For standard setting, dominance relationships are estimated in the final judgments made by the subject matter experts. Ultimately, they decide that this examinee is above or below the theory each has of just passing performance, or that each item measures not enough or too much of what is needed to just pass.

The proximity and dominance judgments are very close. It is clear, though, that dominance cannot be decided without a formulation of what is dominated, and this, in turn, is made by comparison and reflection.

The Importance of the Study

Examining the components of the judgments made by the subject matter experts, should enable reflection on the process itself and suggest new methodologies and instructions, in which the implicit cognitive constructs are made explicit. Most important, the processes of judgment have a well-developed history in psychological measurement, one that needs to be refreshed in terms of the activities that support judgments about items and about student performance.

Ultimately, this elaboration of the meaning of performance adds to the construct validity of the instrument by supporting the meaning of the test score (AERA, NCME, & APA, 1999).

The evaluation of the cognitive demands of standard setting permits an examination of issues such as transitivity and consistency (Coombs, et. al., 1970) in judgments. These are important components for the interpretation of test scores, because they give operational meaning to the inferences available from test scores. For example, if the definition of passing includes the probability that some skill has been acquired, then it necessarily also includes a higher probability that a prerequisite skill has also been acquired.

Evaluation of the efficacy of this judgment model does not mean that the judges consciously construct these matrices. Rather, it assesses how consistently this model can describe the judgment process. Very simply, if the judges behave, in their judgments, as though they refer reliably to the existence of proximity relationships and then to dominance relationships, then this model becomes efficacious for analysis, for scaling, and for explicit formulation of standard setting procedures

Objective of Study

This study proposes a two step cognitive model to describe the judgments made in standard setting. The structure of the two matrices may not be consciously available to participants in standard setting. Nevertheless, the model will prove useful if it consistently accounts for the judgments made, both within judges and across judges. Specifically, we hypothesize the following:

a. the relationships between perceived proximity of test questions to passing will be consistent within subject matter experts across test items;

b. the perceived dominance of a just passing examinee over each test item will be consistent in terms of the difficulty of the items and the hypothesized proximity relationships..

Methods

The Exercise

A standard setting exercise was developed using released operational items from the New York State Mathematics A Regents examination (Math A), a test required for graduation. In all, five items were used: one each with Rasch difficulty values from the difficult (1.40) and easy (-1.33) ranges and three with Rasch difficulty values near that of the passing score (0.54, 0.66, and 0.70).

The judges each reviewed the State learning standards and performance indicators for commencement level mathematics. Instructions (Appendix) were provided for an item mapping exercise, including two rounds of judgments. The exercise was administered to the New York State Curriculum/Development Network, which included assessment specialists and regional curriculum, data analysis, and testing directors. In all, 38 judges participated although not all contributed each of the required judgments.

Note, this sample was atypical of standard setting panels because it was not chosen specifically to represent the State demographically or because the members possessed subject matter expertise. However, the exercise was not designed to set a standard, but rather to investigate the process of judgment.

Also, the spacing and presentation of the items in terms of difficulty was designed to facilitate the judgment task for people who were not subject matter experts. For this reason, the items were given in the order of least difficult, most difficult, below the cutoff, and increasingly closer to the cutoff. The Network participants were also quite familiar with this test content and

with student performance on this test, which is administered three times per year and released after each administration.

Intrajudge Consistency on the Proximity Matrix

Each judge was asked to make a decision concerning whether each test item is "one that a student should answer correctly to pass the Mathematics A examination to meet the requirements for a Regents diploma." After making that decision, for each question, each judge was asked:

> "How close are the content and cognitive demands of this question to those that should be required for a student with sufficient knowledge and skills to pass this examination?";
>
> "What is the probability that a student with minimum knowledge and skills required by the learning standards would answer this questions correctly by any means, including having the required knowledge and skill or guessing?"

In fact, this order of presentation is in reverse of the hypothesized formation of the theory. That is, the cognitive proximity matrix is constructed first and the dominance judgment follows with reference to the proximity matrix. However, making the yes/no decision about the item, assumes that the panelists can decompose the judgment and have the prerequisite proximity matrix available. That is, the yes/no dominance decisions imply the availability of a theory of content and difficulty. Remember that the proximity matrix is really a way of describing how close each stimulus is to meeting the judge's definition of what constitutes minimum item characteristics for passing or minimum examinee performance. The inability to decide which items are needed to pass may indicate either that these proximity judgments cannot be made, or that they cannot be applied yet to items or examinees.

Evaluating the Proximity Matrix

Analysis was made of each judge's proximity matrix by first identifying for each, what constitutes the required domain. This degree of knowledge and skill is represented byθ, which is the implicit ability level for examinees that possess the required knowledge and skill. To determine this value for each judge, for each item, the rated probabilities and the item difficulties were inserted into the Rasch probability relationship to solve for θ (Wright & Stone, 1979):

$$p(xi=1 / \theta) = \frac{\exp(\theta - bi)}{1 + \exp(\theta - bi)} \quad (1)$$

Where "xi = 1" is the occurrence of a correct answer to item i, "exp" is e raised to the given power, and "bi" is the difficulty (in logits) of item i.

Because five items were judged, each judge's implicit criterion contains expected estimation error in relation to the items:

$$\hat{\theta}ij = \theta ij + \varepsilon ij \quad (2)$$

Where "$\hat{\theta}ij$" is the each judge's (j) implicit cut score in formula (1), "θij" is the true skill level for that judge, and "εij" is the estimation error associated with that judge on item i. Over judges, then, the implicit cut score is the mean observed $\hat{\theta}ij$:

$$\frac{\Sigma j \Sigma i \hat{\theta}ij}{nij} = \frac{\Sigma j \Sigma i(\theta ij + \varepsilon ij)}{nij} \quad (3)$$

Note, the implicit passing criterion for each judge is based on its proximity to the difficulty of the item (bi) and the probability of a correct response. That proximity, in turn, is a function of the item content and demands in relation to the required domain.

The proximity matrix, then, emerges from the closeness of the item demands and the judge's theory of what constitutes passing:

$$cij = f(\theta ij - bi) \quad (4)$$

Where cij is the estimated proximity (or "closeness") of these demands to those of the minimal requirements for passing for this judge and this item.

If the proximity matrix is available for each judge, then, to be useful, it should reflect the judge's theory of just passing and therefore influence the judge toward consistency in judgment. If a psychological theory exists of the required domain, as it must for standard setting to occur, then the task of the judges is to estimate how far or near each item or each student is from possessing that domain. In a sense, the proximity matrix contains the very initial judgments of the constitute elements of passing, either in terms of student performance or in terms of item content or cognitive demand.

The question of the availability of that matrix for judgment is addressed by the consistency of judgment within judges across items and across judges. These may be evaluated through regression models designed first to define the relationship in (4), above. For example, because reliability is defined as the true score variance divided by the error variance, it follows that reliability can be estimated through a regression model in which the closeness and distance

functions are each dependent variables in models where the judge and item effects are dependent variables (c.f., Kerlinger,1964).

Secondly the slopes are analyzed to determine the extent to which that relationship varies in relation to the judge. Consistency of that relationship could be interpreted as evidence that the proximity matrix is universally available to these judges. For the purposes of this study, correlations of .3 (r-square of .09) or higher are deemed evidence of consistency.

To evaluate within judge and across judge consistency, general linear model regressions will be used to define the function in equation (4), above. This relationship ultimately addresses the regression of the estimated probability (independent variable) and the closeness estimate (dependent variable). To evaluate the across judge consistency, regression analyses will be made in which judge and difference between each judge's implicit theta and item difficulty were independent variables and the closeness estimate is the dependent variable. The error component to evaluate whether the slopes are consistent across judges will be the interaction of the difference between the implicit theta and item difficulty and the judge.

Actually, two analyses were conducted as described above. They differed as to the choice of the implicit theta used to define the distance function. The first used the value obtained for each judge for each item (as in formula 2 above), while the second used the theta obtained by averaging the values within judge and over items (as in 3 above, assuming that the individual item values incorporate errors in judgments related to items).

Closeness was rated on an 11-point scale, ranging from –5 to +5. Like item difficulty, these ratings range from "substantially below" the required demands for passing to "substantially above" the required demands for passing. Nevertheless, because of the range of the scale, with "just right" having a value of zero, preliminary analyses were conducted estimating the

contribution not only of the linear relationship between closeness and the difference between θij and bi, but also the quadratic relationship, ((θij and bi)**(2)). Because the latter did not add significantly to the model, only the linear relationship was explored.

Evaluating the Dominance Matrix

For each of the five items, judges made summative judgments of whether or not the item should be passed by students who just meet the demands of the New York State learning standards, in this case, in mathematics. Because there were only five items judged, it would be difficult to develop standardized distances or discriminant functions within judge comparing the means and standard deviations of items judged as needed to pass to the means and standard deviations of items judged as beyond what is needed to pass. In some cases, one or the other group would only contain an item. Therefore, a more reliable method was used, first identifying the items as one of two kinds within judge: needed to pass or beyond what is needed to pass. Analyses were then made of the group of items across judges comprising each of these two item groups. No assumptions were made on the actual difficulties of the items, so it was possible for some to be more difficult than others, but be judged as needed while less difficult items could be judged as beyond what is needed.

If their dominance matrices are available for each judge, then their judgments of which items needed to meet the standards should be consistent with the implicit cut scores (based on the probability estimates). The questions remain whether the judgments are consistent both between and within judges. A methodological challenge, here, is to avoid using the probability ratings to validate the implicit cut scores, since the latter are generated by the former:

$$\theta j = \frac{\Sigma i(\log(p) - (\log(q)) + (bi))}{nj} \quad (5)$$

Where θj is each judge's implicit cut score, p is the probability estimate made by the judge (j), q is 1 – p, bi is the item difficulty, and nj is the number of items for which the judge gave probability estimates.

Very simply, having designed an accessible proximity matrix enables the judges to rate both the probabilities of a correct response by a just passing student and the closeness of each item to the standards. It follows, that summative judgments form the dominance matrix refer to the domain built on the proximity judgments. This relationship is evaluated by how consistent those summative judgments are with the proximity construct.

Two types of analyses addressed consistency of judgments. The first was a general linear regression in which the yes/no summative judgment and the judge were treated as independent variables. The closeness estimate and the difference between the implicit cut score and the item difficulty (θj - bi, hereafter called the "distance function"), both elements of the proximity matrix, were each treated as dependent variables.

The second analysis actually consisted of several classificatory discriminant analyses. The analyses identified the values of the distance functions and of the closeness ratings for each question that were the implicit cut scores for the summative yes (it must be passed)/ no (it need not be passed) determinations. This classificatory analysis identifies the implicit theta that dividing the yes items from the no items.

Finally, analyses are provided of the agreement among the various indicators. Agreement between the summative yes/no judgment and the distance function was evaluated by counting

each positive distance function as a yes decision and each negative as a no decision and then computing the percentage agreement of the two judgments. Correlational analyses and stepwise analyses are also provided to define the relationships among the proximity variables and the summative judgments.

Results

The Proximity Matrix

Implicit thetas. The implicit thetas and the distance functions (average of the differences between thetas and item difficulties) are shown in table 1. The mean squares for the item effect were 19.59 and 6.65 for the closeness estimate and distance function, respectively. The mean squares for the judge by item interaction were 2.68 and 0.62, respectively. Therefore, the reliability estimates for the judgments across judges were 0.86 and 0.91, respectively. These are quite respectable.

The closeness function. The regression of the closeness estimate onto the distance function was .235. The linear function was:

$$\text{Closeness} = -0.931 * (\text{distance function by judge by item}) + 0.445 \qquad (6)$$

The magnitude of this relationship supports the existence of a judgment of proximity to a domain, manifest consistently in the closeness estimate and the implicit theta.

The regression of the closeness estimate onto judge and the distance function produced an r-square value of 0.713. The F-ratio for the homogeneity of the slopes over judges revealed that the relationship between these two variables was homogenous across judges ($F(df=31, 31) = 0.68$, ns). This finding supports the universal accessibility of the proximity matrix. Moreover, this consistency in the relationship across judges also lends some support to aggregating data across judges to evaluate the dominance matrix (below).

Using the average (over items) theta value, the relationship between the closeness estimate and the distance function yields an r-square value of .093:

Closeness = -0.487 * (distance function by judge) + 0.258 (7)

The regression of the closeness estimate onto both judge and the distance function produced an r-square value of 0.528. As with the average theta values, there were no slope differences related to judge ((F(df=34, 34) = 1.03, ns). The degree of relationship shows that, in fact, the implicit theta varied for the judges by item, but was of sufficient magnitude ($r > 0.3$) to support the existence of the proximity judgment. Again, the nominal slope variation suggests universal access to the matrix and supports aggregating the data across judges to evaluate the dominance matrix (below).

For the purposes of the dominance matrix analyses, the average theta over items for each judge are used. As shown above in formulae (2) and (3), the existence of error attributable to items make these analyses somewhat conservative.

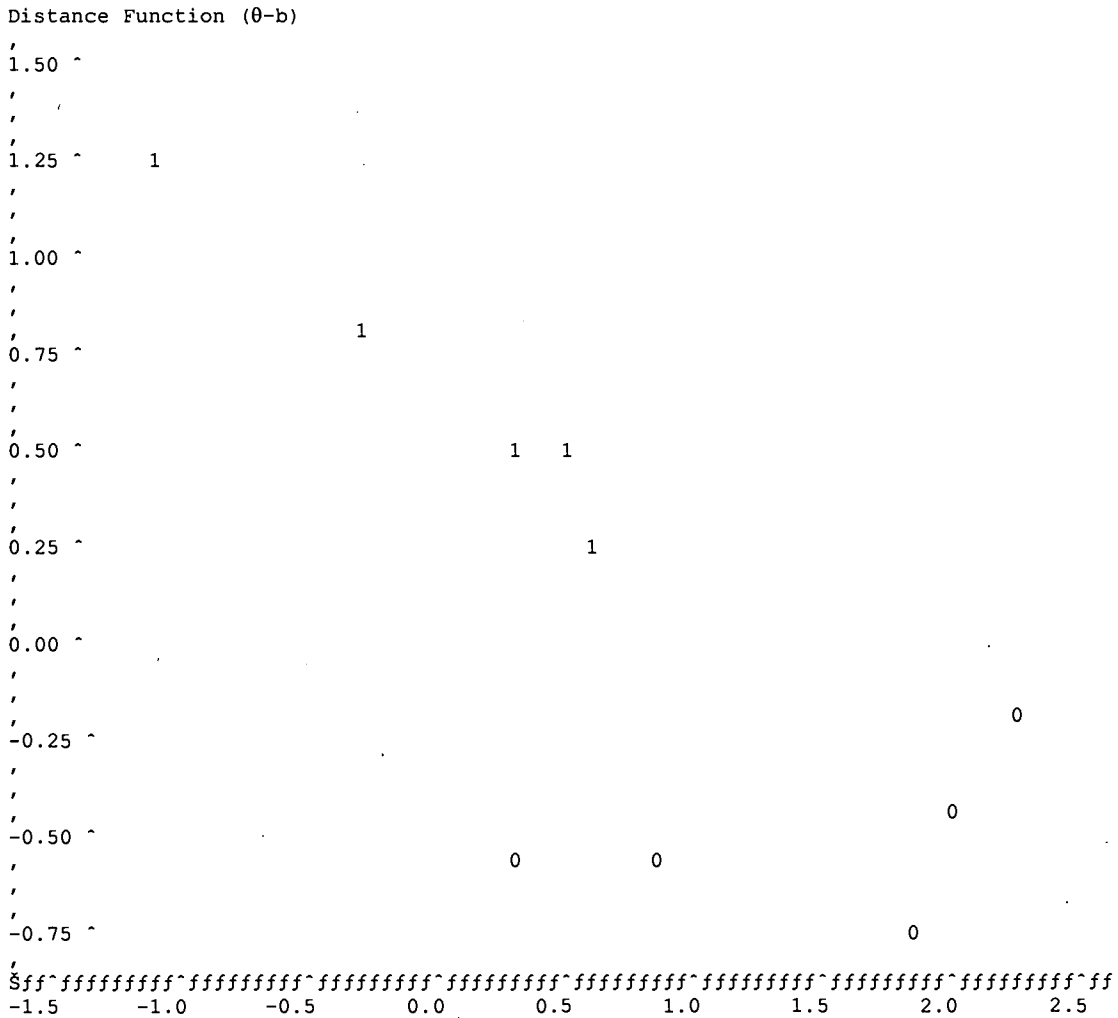
The Dominance Matrix

General linear regression. There was a significant relationship between the yes/no judgment about the necessity of the item for passing both for the distance function, using the mean theta for each judge over items (r-square = 0.377, F(df=1, 107)=9.88,p<.01) and for the closeness estimate (r-square = 0.439, F(df=1, 98)=18.68, p<.0001). Figure 1 shows that as the distance function increases (greater likelihood to pass the item), the judgment that the item

should be passed is increasingly positive and the closeness judgment that the item is well within the domain also increases.

Figure 1

Plot of Distance Function by
Closeness Estimate
(0 = item need not be passed,
1 = item should be passed)



The relationships between the yes/no judgment and both the distance function and closeness estimates did not vary across judges ($F(df=20, 107)=0.15$, ns; and $F(df=17, 99)=1.33$, ns; respectively). Again, this shows there was consistency in access of the hypothetical dominance matrix across judges. The overall mean distance function for the no judgment (beyond the passing requirement) was –0.14, and the mean distance function for the yes judgment was 0.77 (sd=1.13, or 0.81 standard deviations difference). The mean closeness estimate for the no judgment (beyond what is needed) was 1.23, and the mean closeness estimate for the yes judgment was –0.32 (sd=1.80, or 0.86 standard deviations difference). Clearly, the dominance judgment is consistent with both the implicit theta (Table 1, below) and distance function and with the closeness construct.

Figure 1 shows the mean closeness ratings and distance functions for each of the five items, by whether it was rated as needed to pass (1) or not needed to pass (0). Note that each of the items rated needed to pass had mean distance functions of .25 or higher. Only one of the items rated as not needed to pass had a mean closeness rating lower than 0.5, and for this item, the distance function was –0.50.

Table 2a and 2b show that the mean differences between items deemed necessary to pass and those deemed beyond necessary to pass both for the closeness ratings and for the distance function values was above a standard deviation for each of the questions judged. Clearly, the yes/no decision, item by item, was consistent both with the closeness estimate and with the implicit thetas for the judges, supporting the hypothesis that the dominance yes/no judgment builds on the foundational proximity judgments. Table 3 presents the interrelationships among the judgments made.

Note from Table 3 that the distance function and the closeness estimates were both related to the yes/no decision about the inclusion of the item as a passing criterion. Note, as well, that the probability judgment was also significantly related. However, this judgment, at least logically, appears to be more an outcome of the judgment of the ability of the examinee to conquer the item, and therefore would come after deciding on the minimum competency level (the implicit theta drives this estimate) and the item difficulty.

Discriminant functions. The first discriminant analysis demonstrated differences in the criterion values for both the closeness estimates and the distance functions in relation to the items. The question-by-question values are presented in tables 4a (closeness) and 4b (distance functions), respectively, below. These variations could actually be related to the differences in the linearity of the relationships among the probability estimates and the closeness estimates. As the difficulties of the items change, the probabilities might not change linearly. This is true, as well, of the closeness ratings. Tables 2a and 2b present the mean closeness ratings and mean distance functions in relation to the yes/no decision, showing clear differences in each of the dependent variables.

Finally, Table 5 presents the agreements of the yes/no decision with whether items should be passed based on the implicit theta being higher than the item difficulty (the distance function) computed in two ways:

1. using the item by item implicit thetas for each judge, and
2. using the mean theta (the more conservative test) over items for each judge.

As well, the agreement between the yes/no decision and the distance function exceeding values yielding probabilities of .67 or higher (as was the instruction to the judges) was computed

for each item. The results show very consistent agreement, indicating reliability among the judges in referencing the proximity matrix to make the dominance decisions.

Table 1

Mean Difference between
Theta and Item Difficulty
by Judge

Judge	Theta	Implicit Distance Function
1	0.50	0.11
2	1.56	1.16
3	0.83	0.44
4	-0.86	-0.90
5	1.76	0.88
6	1.01	0.69
7	0.43	0.03
8	1.00	0.61
9	0.48	0.08
10	0.89	0.56
11	0.92	0.59
12	-0.37	-0.69
13	0.95	0.55
14	1.28	1.03
16	2.79	1.39
17	0.81	0.55
18	0.76	0.37
19	0.53	0.49
20	1.02	0.69
21	0.95	0.55
22	2.44	1.39
23	0.50	0.11
24	0.58	0.19
25	1.64	1.24
26	0.31	-0.52
27	1.14	0.75
28	1.00	0.61
29	0.69	-0.13
30	1.36	1.39
31	1.17	0.35
32	0.88	0.55
33	0.78	-0.04
34	1.03	0.64
36	0.56	0.16
37	1.94	1.06
Mean	0.95	0.48

Table 2a

Mean Closeness Values, by Item, for

Judges Deeming the Item Necessary to Pass

and For Judges not Deeming

the Items Necessary To Pass

	Needed to Pass			Not Needed to Pass			All Judges		
Item	N	Mean	S.d.	N	Mean	S.d.	N	Mean	S.d.
8	23	0.04	1.82	11	1.55	1.29	34	0.53	1.80
11	33	-1.33	1.41	1	2.00	----	34	-1.24	1.50
12	25	-0.28	1.14	4	2.25	0.50	29	0.07	1.39
18	26	0.23	1.77	6	0.83	2.14	32	0.34	1.82
20	15	0.33	1.35	7	0.29	2.98	22	0.32	1.94

Table 2b

Mean Distance Function Values, by Item, for

Judges Deeming the Item Necessary to Pass

and For Judges not Deeming

the Items Necessary To Pass

	Needed to Pass			Not Needed to Pass			All Judges		
Item	N	Mean	S.d.	N	Mean	S.d.	N	Mean	S.d.
8	24	-0.37	0.69	11	-0.64	0.63	35	0.45	0.67
11	26	2.33	0.60	1	0.47	----	35	2.28	0.67
12	26	0.42	0.62	4	0.48	0.15	31	0.43	0.57
18	27	0.39	0.60	6	-0.31	0.74	34	0.25	0.47
20	16	0.33	0.49	12	0.11	0.44	28	0.24	0.48

Table 3

Correlation Matrix between Yes/No

Judgment for Items, Closeness Estimates,

Probabilities of Passing, and Distance

Function Estimates

Variable:	Needed for Passing	Probability Correct	Closeness	Distance Function
Needed For Passing	-----	0.600	-0.336	0.333
P(correct)	-----	-----	-0.420	0.643
Closeness	-----	-----	-----	-0.305

Table 4a

Criterion Closeness Values for Items to be Deemed Necessary for Passing the Test, by Item (Discriminant Analysis)

(-5 = substantially below requirement, +5 = substantially beyond requirement)

Item	Criterion Closeness
8	2.5
11	above 2
12	1.5
18	5.5 (all judges)
20	-3

Tables 4b

Criterion Distance Values for Items to be Deemed Necessary for Passing the Test, by Item (Discriminant Analysis)

Item	Criterion Distance	Corresponding Theta
8	-2.01	-0.61
11	0.25	-1.08
12	2.25*	2.79
18	–0.73	-0.03
20	-0.03	0.63

* All judgments for this item were reclassified as needed to pass.

Table 5

Agreement between the Yes/No Decision on Item Necessity

for Passing and the Distance Function,

For Criterion Values of .50 Chance of Passing

and .67 Chance of Passing,

for Each Judge's ImplicitTheta Computed

across Items (Mean) and within Items

	Theta Computed			
	Across Items		Within Items	
Item	p=.50	p=.67	p=.50	p=.67
8	.417	.361	.722	.583
11	.971	1.000	.771	.629
12	.625	.313	.844	.563
18	.686	.371	.714	.571
20	.483	.552	.793	.483

Conclusion

All standard setting techniques involve some hypothetical construction on the part of judges. Either there is the formation of a knowledge and skills domain with criteria for inclusion of elements in the domain, or there is the more difficult location of a body of knowledge and skills within a hypothetical examinee of minimum competence. It follows that the definition of this domain or of this examinee is increasingly sharpened by interaction with test questions: this question goes beyond what is the domain; the domain contains this one; this question contains what a minimally competent examinee should know.

These judgments imply that each item is evaluated for content and difficulty and compared to the required domain. Essentially, this comparison is what has been called a proximity matrix, and the process of making the comparison increases the operational definition, for each judge, of the domain. This clarifying of the definition enables each judge not only to judge the contents and difficulty of each item in terms of congruence with the domain, but also to judge whether the passing examinee would answer the question correctly. This superiority of the examinee to the test question determines what has been called a dominance matrix.

If in fact, these cognitive processes underlie the judgments of standard setting, then these models of comparison and contrast may be helpful for understanding the process. For example, it is clear from these results that the behavior of judges is consistent with the existence of a cognitive construct against which the content and difficulty of test questions may be compared. It is also clear that judgments are consistently made about what constitutes sufficient content and difficulty to be required for passing.

The findings of these analyses support the existence of a set of consistent judgments, some involving proximity to the requirements for passing and others building on that proximity to buttress summative judgments. We note that there are differences among items and judges on the values of the implicit thetas. These can be related to the nonlinear relationships underlying a psychophysical scale, e.g., as the item difficulties vary, the closeness or probability ratings vary nonlinearly. These differences may well be related to the lack of specific subject matter expertise among these judges, as compared to the normal composition of standard setting panels.

Finally, defining the psychophysical structure may prove useful as an explicit process in standard setting, supported by directions about how to construct the required domain and how to compare the characteristics of test questions to that domain. Follow up research is needed on a standard setting exercise in which the cognitive constructs involved in judgment are made explicit, and their contribution to the process is measured and evaluated.

APPENDIX

Standard Setting Exercise

In this exercise, you will be asked to review the key ideas and performance indicators for the Mathematics A test. You will then review a series of actual test items to decide whether each item is one that a student should answer correctly to pass the Mathematics A examination to meet the requirements for a Regents diploma. Once you have made that decision, you will be asked to make two other decisions:

(1) how close each question is in content and cognitive level required, to measuring the body of knowledge and skills required for this key idea;
(2) the probability that a just passing examinee (at a scale score of 65) would answer this question correctly.

Please complete the following questions first:

How many years' experience do you have teaching:

1. Mathematics

___ ___ ___
0 1-4 5 or more

2. High School Mathematics

___ ___ ___
0 1-4 5 or more

3. School Testing

___ ___ ___
0 1-4 5 or more

Instructions

In your considerations below, please consider that a passing student (at 65) should demonstrate mastery of all of the content required by the learning standard, including that for each key idea and performance indicator. A student passing with distinction should demonstrate mastery of all of the content required by each key idea and also demonstrate ability to go beyond that requirement for each key idea.

Remember, if a student who is right at the passing level should be able to demonstrate mastery, then all of the students above that level (65-100) should also be able to demonstrate that capability. If a student right at the level of passing with distinction is able to demonstrate mastery, then all of the students above that level (65-100) should also be able to demonstrate that capability.

For this exercise, you will be asked to make three judgments, as described above. The first will be a simple dichotomous judgment: Should this question be answered correctly by 2/3 of the time a student who has demonstrated sufficient knowledge and skills required by the learning standards?
The second judgment is: How close are the content and cognitive demands of this question to those that should be required for a student with sufficient knowledge and skills to pass this examination?
Finally, the third judgment is: What is the probability, that a student with the minimum knowledge and skills required by the learning standards would answer this question correctly by any means, including having the required knowledge and skill or guessing?

Round 1 Judgments

Question Number:

1. Should this question be answered correctly by a just passing student? ____ ____
yes no

2. How close are the content and cognitive demands of this question to those required for a student to pass this examination?

-5	0	+5
Substantially Below	Just Right	Substantially Above

3. What is the probability that a just passing student, at a scale score of 65, would answer this question correctly (or achieve this number of points for an open-ended question)?

0	.2	.4	.5	.6	.8	1.0
Not at All						Certain

Round 2 Judgments

Question Number:

1. Should this question be answered correctly by a just passing student? ____ yes ____ no

2. How close are the content and cognitive demands of this question to those required for a student to pass this examination (circle one)?

-5 -4 -3 -2 -1 0 +1 +2 +3 +4 +5

Substantially Below — Just Right — Substantially Above

3. What is the probability that a just passing student, at a scale score of 65, would answer this question correctly (or achieve this number of points for an open-ended question) (circle one)?

0 .2 .4 .5 .6 .8 1.0

Not at All — Certain

References

American Educational Research Association, American Psychological Association, National Council on Measurement in Education (1999), Standards for Educational and Psychological Testing. Washington, DC: American Educational Research Association.

Coombs, C. H., Dawes, R. M., & Tversky, A. (1970), Mathematical Psychology: An Elementary Introduction. Englewood Cliffs, NJ: Prentice Hall, Inc.

Kerlinger, F. N. (1964) Foundations of Behavioral Research: Educational and Psychological Inquiry. New York, N.Y.: Holt Rinehart and Winston, Inc., pp. 436-437.

Linacre, M. & Wright, B. D. (1998), A User's Guide to Winsteps, Bigsteps, Ministeps Rasch Model Computer Programs. Chicago, Ill.: MESA Press.

Masters, G. (1982) , A Rasch model for partial credit scoring. Psychometrika, Vol. 47, no. 2, June 1982, pp. 149-174.

Wright, B. D. & Stone, M. H. (1979), Best Test Design: Rasch Measurement. Chicago, Ill.: MESA Press.

U.S. Department of Education
Office of Educational Research and Improvement (OERI)
National Library of Education (NLE)
Educational Resources Information Center (ERIC)

REPRODUCTION RELEASE

(Specific Document)

TM034927

I. DOCUMENT IDENTIFICATION:

Title: Developing a Theory of Performance: A Two-Stage Structure for the Psychology of Standard Setting

Author(s): Gerald E. DeMauro

Corporate Source: New York State Education Department

Publication Date: April, 2003

II. REPRODUCTION RELEASE:

In order to disseminate as widely as possible timely and significant materials of interest to the educational community, documents announced in the monthly abstract journal of the ERIC system, *Resources in Education* (RIE), are usually made available to users in microfiche, reproduced paper copy, and electronic media, and sold through the ERIC Document Reproduction Service (EDRS). Credit is given to the source of each document, and, if reproduction release is granted, one of the following notices is affixed to the document.

If permission is granted to reproduce and disseminate the identified document, please CHECK ONE of the following three options and sign at the bottom of the page.

The sample sticker shown below will be affixed to all Level 1 documents

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL HAS BEEN GRANTED BY

Sample

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

1

Level 1

Check here for Level 1 release, permitting reproduction and dissemination in microfiche or other ERIC archival media (e.g., electronic) *and* paper copy.

The sample sticker shown below will be affixed to all Level 2A documents

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL IN MICROFICHE, AND IN ELECTRONIC MEDIA FOR ERIC COLLECTION SUBSCRIBERS ONLY, HAS BEEN GRANTED BY

Sample

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

2A

Level 2A

Check here for Level 2A release, permitting reproduction and dissemination in microfiche and in electronic media for ERIC archival collection subscribers only

The sample sticker shown below will be affixed to all Level 2B documents

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL IN MICROFICHE ONLY HAS BEEN GRANTED BY

Sample

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

2B

Level 2B

Check here for Level 2B release, permitting reproduction and dissemination in microfiche only

Documents will be processed as indicated provided reproduction quality permits.
If permission to reproduce is granted, but no box is checked, documents will be processed at Level 1.

I hereby grant to the Educational Resources Information Center (ERIC) nonexclusive permission to reproduce and disseminate this document as indicated above. Reproduction from the ERIC microfiche or electronic media by persons other than ERIC employees and its system contractors requires permission from the copyright holder. Exception is made for non-profit reproduction by libraries and other service agencies to satisfy information needs of educators in response to discrete inquiries.

Sign here, please

Signature: Gerald E. DeMauro

Printed Name/Position/Title: Gerald E. DeMauro / Coordinator of State Assessment, NY

Organization/Address: New York State Education Dept. 89 Washington Ave, 775 EBA Albany, NY 12234

Telephone: 518-474-5902

FAX: 518-474-1989

E-Mail Address: GDeMauro@mail.NYSED.gov

Date: 5-14-03

(Over)

III. DOCUMENT AVAILABILITY INFORMATION (FROM NON-ERIC SOURCE):

If permission to reproduce is not granted to ERIC, or, if you wish ERIC to cite the availability of the document from another source, please provide the following information regarding the availability of the document. (ERIC will not announce a document unless it is publicly available, and a dependable source can be specified. Contributors should also be aware that ERIC selection criteria are significantly more stringent for documents that cannot be made available through EDRS.)

Publisher/Distributor: New York State Education Department, Office of State Assessment
Address: 89 Washington Ave., 775 EBA Albany, NY 12234
Price: —

IV. REFERRAL OF ERIC TO COPYRIGHT/REPRODUCTION RIGHTS HOLDER:

If the right to grant this reproduction release is held by someone other than the addressee, please provide the appropriate name and address:

Name:
Address:

V. WHERE TO SEND THIS FORM:

Send this form to the following ERIC Clearinghouse:

ERIC CLEARINGHOUSE ON ASSESSMENT AND EVALUATION
UNIVERSITY OF MARYLAND
1129 SHRIVER LAB
COLLEGE PARK, MD 20742-5701
ATTN: ACQUISITIONS

However, if solicited by the ERIC Facility, or if making an unsolicited contribution to ERIC, return this form (and the document being contributed) to:

ERIC Processing and Reference Facility
4483-A Forbes Boulevard
Lanham, Maryland 20706

Telephone: 301-552-4200
Toll Free: 800-799-3742
FAX: 301-552-4700
e-mail: ericfac@inet.ed.gov
WWW: http://ericfacility.org