DOCUMENT RESUME

ED 392 845 TM 024 693

AUTHOR Kahl, Stuart R.; And Others
TITLE Setting Standards for Performance Levels Using the Student-Based Constructed-Response Method.
PUB DATE 22 Apr 95
NOTE 23p.; Paper presented at the Annual Meeting of the American Educational Research Association (San Francisco, CA, April 18-22, 1995).
PUB TYPE Reports - Descriptive (141) -- Speeches/Conference Papers (150)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS Academic Standards; Achievement Tests; *Constructed Response; *Cutting Scores; Definitions; *Educational Assessment; Grade 8; Item Response Theory; Judges; Junior High Schools; Mathematics Tests; *Performance Based Assessment; Policy Formation; Reading Tests; State Programs; *Test Construction; *Testing Programs
IDENTIFIERS Maine Educational Assessment; Performance Levels; *Standard Setting

ABSTRACT

The assessment instruments of the Maine Educational Assessment emphasize extended constructed-response questions. The results from these assessments are reported in terms of percentages of students at four performance levels. The Student-Based Constructed-Response Method was used to establish performance standards for these levels on the eighth-grade reading and mathematics tests. First, a policy advisory group recommended general definitions for the four performance levels. Then, subject-area committees (including educators and noneducators) translated these definitions into brief subject-specific definitions. The members of these committees served as the judges for standard-setting, assigning individual students to proficiency levels based on a review of the students' answers to the extended constructed-response questions. Work from the full range of the item response theory ability scale was judged to identify the cut score intervals, and extensive review of those intervals pinpointed the cut scores. The judges attained high levels of agreement. (Contains 10 exhibits and 3 references.) (Author/SLD)

SETTING STANDARDS FOR PERFORMANCE LEVELS USING THE STUDENT-BASED CONSTRUCTED-RESPONSE METHOD

Stuart R. Kahl
Timothy J. Crockett
Charles A. DePascale
Sally L. Rindfleisch

Advanced Systems in
Measurement and Evaluation, Inc.
Dover, New Hampshire

Paper presented at the Annual Meeting of the American Educational Research Association, San Francisco, April 1995.

BACKGROUND

The need for new, effective standard-setting procedures for constructed-response tests has grown significantly because of two important developments in education. The first is the widespread recognition of the potential negative impacts on curriculum and instruction of high-stakes tests dominated by the multiple-choice format. The second development is the growing dissatisfaction with norm-referenced reporting of test results because of its failure to convey what it is that students understand and can do. Related to this second issue is educators' increased understanding that impressions left by seemingly positive normative test results can be inconsistent with students' ability (or lack thereof) to actually perform on more "authentic" and higher order tasks.

As a result of these developments, many states' accountability testing programs have begun to (1) make extensive use of constructed-response questions and (2) report test results in terms of percentages of students at various performance or proficiency levels. Such states include (but are not limited to) Delaware, Maryland, Kentucky, Massachusetts, New Hampshire, and Maine. The major purposes of this paper are to describe the procedures of the Student-Based Constructed-Response (SBCR) Method of standard setting, to share data from the application of this method to the Maine Educational Assessment, and to discuss the relative merits of the method.

The Maine Educational Assessment (MEA) is a statewide testing program designed to both evaluate school programs and facilitate improvement of curriculum and instruction in Maine schools. Through this program, grade 4, 8, and 11 students are assessed annually in six subject areas. Over the last nine years, the major emphasis in the assessment instruments has shifted from multiple-choice items to extended constructed-response questions. A recent decision to change the focus of reporting from average scaled scores to percentages of students at well defined performance levels necessitated the creation of performance standards (cut scores) to "separate" the levels. In recent years, there has been dissatisfaction with traditional methods of standard setting designed for multiple-choice tests and an increasingly strong belief that responses to extended constructed-response questions are more direct and far better indicators of what students know and can do. These views led to the refining of the SBCR Method, which was used in May and September of 1994 to establish performance standards for the MEA grade 8 reading and mathematics tests. The standard setting was conducted using the constructed-response components of the 1993-94 MEA instruments. The 1994-95 instruments, which were totally constructed-response (each response scored on a scale from 0 to 4), were equated as usual to the previous year's tests; then the cut scores previously established were applied to the 1994-95 scaled scores in reading and mathematics. Subsequent MEA reports displayed the percentages of students at various performance levels. The release of many of the assessment questions, along with their scoring guides and samples of student work, effectively communicated the standards to school personnel and the general public.

OVERVIEW OF THE SBCR METHOD

Prior to the actual standard setting, a policy advisory group, consisting of 50 to 75 educators and noneducators representing many walks of life and special interest groups, recommended general definitions for the four desired performance levels (novice, basic, advanced, and distinguished). Then, subject area committees (consisting of 20 to 25 on-grade teachers, other educators, and noneducators) translated these definitions into brief subject-specific definitions (still expressed in general terms). Later, the members of these committees also served as standard-setting judges. The

process required judges to assign students to performance levels based on their review of the students' answers to the extended constructed-response questions. There were two stages to the process. Initial judgments focussed on the work of students whose performance spanned the full range of the IRT ability scale. These probes facilitated the identification of probable intervals on the scale in which the cut scores would be located. The more extensive review of the work of students within those selected intervals helped to pinpoint the cut scores.

The rationale for the use of the SBCR Method has to do with the nature of the decisions standard-setting judges are asked to make. The modified Angoff method, used by the National Assessment Governing Board (NAGB) to establish cut scores for the predominately multiple-choice National Assessment of Educational Progress (NAEP) a few years ago, has been criticized for two major reasons. First, there is serious doubt that judges are able to make the judgments they are asked to make -- estimating item p-values for groups of "borderline" students. Second, there appears to be considerable inconsistency between what NAEP reports suggest students are able to do and what they are actually able to do, according to validation studies.

One recently developed method of standard setting for constructed-response tests requires judges to relate response categories of student work to predefined performance levels item by item. This method has been used in Maryland and in Maine for the subject areas in which the tests were fully matrix sampled. In Maine, judges found this task difficult because single items did not provide adequate evidence for such judgements. It is for this reason that the SBCR Method is so appealing. It enables judges to focus on complete sets of responses of students to many constructed-response questions. With such evidence, the matching of student work to performance level definitions appears to be a task that any judges are qualified and able to perform. The judges are typically impressed with the levels of agreement they attain and with the confidence they feel in the process. Defensible, believable standards result.

MEA STANDARD-SETTING STEPS

This section describes in detail the steps involved in the SBCR Method. These were the exact steps followed in applying the method to the MEA reading and mathematics assessments. Throughout this paper, data associated with the reading assessment are used for illustrative purposes. Because of the detail, some of the procedures may seem hard to follow on first reading. The reader is urged to refer to the appropriate exhibits attached to the paper as they are discussed in this section.

Meetings:

1. Convene policy advisory committee to create general definitions of performance levels. (See Exhibit A.)

2. Convene subject area committees (on-grade teachers, other educators, and noneducators) to translate general definitions of performance levels into subject-specific definitions. See Exhibit B. (The abbreviated definitions presented as exhibits have been expanded for release to the field with further explanation and student work samples. These materials are consistent with Maine's "Common Core of Learning" and curriculum standards developed by various groups at the national level -- e.g., NCTM, AAAS.)

3. Convene subject area committees to make judgments for use in standard setting.

Homework:

A complete set of questions and scoring guides must be provided to the judges at Meeting # 2. Before Meeting #3, subject area committees (judges) review constructed-response questions and the descriptors of the 4-point (top) responses from the scoring guides. (The purpose of the homework is simply to familiarize the judges with the questions, thereby saving time at the next meeting.)

Preparation for the SBCR Method for Reading and Mathematics:

1. Produce IRT scaled scores for students based on common questions. (These are questions to which all students respond. There were five such questions in the 1993-94 MEA reading instrument.)

2. Eliminate from the file student records with highly variable raw item scores, that is, with range greater than 2. (For example, 4,4,2,3,2 is acceptable, but 4,4,3,2,1 is not.)

3. Sample 50 students from each quarter logit. (The students' IRT ability scores ranged from -3.0 to +3.0 approximately. Thus, there were approximately 24 "quarter-logit" or quarter-unit intervals on that scale.)

4. Rank order students by scaled score.

5. Produce printout listing (in rank order) student name, scaled score, raw scores, lithocode (student serial number), and any other information that would facilitate the location of actual student responses in storage.

6. Identify 10 students in each quarter logit whose response sets are to be pulled: select the 1st student, the last student, and 8 students spaced at equal intervals in between. (Do not pull responses of students in quarter-logit ranges including students earning an average of one raw point or less per item. Based on the scoring rubrics and performance level definitions, students scoring 1 point on the test questions could not be considered above the lowest level of performance.)

7. Prepare "homogeneous" folders (one for each quarter logit), each of which includes responses of the 10 students identified in the step above. Place these student response sets in rank order from highest to lowest scaled score in the folder and attach a list of the student lithocodes in the same order to the inside front cover of the folder. Number the outside of the folders consecutively with "1" corresponding to the highest quarter-logit set.

8. Prepare the "heterogeneous" folder which should include copies of the top and the bottom student response sets from every quarter-logit folder. These should be in random order. (Only the leader's heterogeneous folder should list student lithocodes in order by scaled score in the inside front cover.)

9. Produce only a few copies of each homogeneous folder (since judges do not have to examine a particular homogeneous folder at the same time) and one copy of the heterogenous folder for every judge.

10. Prepare SBCR preliminary and final rating forms. (See Exhibits C and D.) The preliminary rating form lists in rank order by scaled score the lithocodes of the students whose response sets are in the heterogeneous folder. The final SBCR rating form is generic.

Running Meeting 3 - The Standard-Setting Meeting Using the SBCR Method:

1. Provide background, describe procedures, review definitions of performance levels. Distribute one heterogeneous folder to every committee member (judge).

2. Ask the judges to locate the work of a subset of students represented in the heterogeneous folder by giving them the lithocodes (in random order) of the top response set in every other homogeneous folder (folder 1, folder 3, folder 5, etc.). (NOTE: These response sets are already in their heterogeneous folders.) Have the judges independently rank order these response sets based on overall quality, keeping in mind the performance level descriptions. Have the judges record their rank orderings on a small slip of paper. This will not be turned in.

3. Next, write the lithocodes of the response sets just reviewed on newsprint in order from highest to lowest actual performance based on scaled scores. Have the judges note the extent of agreement.

4. Ask the judges to now assign each of the response sets they ranked to a performance level. They should each write their decisions on a small slip of paper, again not to be turned in. Record their votes (based on shows of hands) next to the lithocodes on the newsprint.

5. Discuss in depth the response sets just rated as they relate to the performance levels definitions. Stimulate discussion with such questions as, "Why did most of you call this student's work 'advanced'?"

6. Have the judges reconsider their ratings of the student response sets and transfer their final ratings to a Preliminary SBCR Rating Form on which the lithocodes of all the response sets in the heterogeneous folder have been entered in order from highest to lowest actual performance.

7. Ask the judges to decide upon the performance levels of the rest of the sets in the heterogeneous folder and record their ratings on their preliminary rating forms.

8. Record the "votes" for all response sets on a "master" preliminary rating form based on shows of hands. Then gather the preliminary rating forms. (Data from "master" preliminary rating forms is shown in Exhibit E.)

9. Have the Chief of Standard Setting determine the homogeneous folder or folders that must be evaluated by the judges for determining each of the three cut points. (These would be the

folders representing the scaled score intervals in which the transition from one performance level to another must occur based on the aggregated ratings from the preliminary rating forms. An example is discussed in a later section.)

10. Divide the group of judges into thirds and have each small group examine the homogeneous folder or folders for one cut score. Have each judge complete a final SBCR rating form for each folder he/she is assigned. Rotate the materials so that all three small groups examine the homogeneous folder or folders for every cut point.

Using the Judgments to Determine Cut Points

After aggregating the final ratings from the SBCR Method, the determination of cut points is a relatively simple process. If the response sets from only one quarter-logit interval were examined for a particular cut point, the aggregated ratings will give us an average proportion of papers in a folder belonging to each of the two proficiency levels under consideration. If four-tenths of the papers are in the upper level, then the cut point would be the scale score within that quarter logit that separates the top four-tenths from the bottom six-tenths of the students within the quarter-logit range. If there is some doubt about which quarter logit "contains" the cut score, then two quarter-logit folders can be merged and the same approach applied to the new half-logit range. The process of determining cut scores based on the judges' ratings is illustrated in Exhibit G.

DISCUSSION OF SELECTED STEPS OF THE SBCR METHOD

The preparation of materials for use in the later steps of the SBCR Method is critical. While quite labor intensive, this preparation can make the last meeting, at which the judges match student work to performance level definitions, run particularly smoothly. The initial rank ordering of the work of selected students at that meeting is a real confidence builder. Since the response sets used in this step represent a wide range of performance, the judges are generally quite successful in rank ordering the work. These response sets also provide material for the discussion of how the capabilities described in the performance level definitions are represented in the students' responses. Thus, the steps involving this subset of response sets in the heterogeneous folder constitute important training. Also, once the judges have made their final independent judgments about the performance levels of these response sets, they have a portion of their preliminary rating forms already completed.

Exhibit E shows an aggregation of information from the judges' preliminary SBCR rating forms completed in the standard-setting meetings for reading. These are data from the "range-finding" activity which required the judges to rate student work in the "heterogeneous" folder. The response sets in that folder were the work of the high and low students in each of the ability intervals (.25 units or "logits" on the IRT scale). For each interval, there was a "homogeneous" folder containing the response sets of 10 students (including that interval's "representatives" to the heterogeneous folder). For each of the three groups of judges setting standards for the reading test, the preliminary ratings depicted in Exhibit E led to the identification of folders 2 and 3 as the folders with response sets requiring in-depth examination in order to pinpoint the cut score separating the distinguished (D) and advanced (A) levels. Exhibit F illustrates this "rangefinding" process

graphically. (Note: Exhibit F shows different ways the homogeneous folders can be identified. The exhibit is not intended to be consistent with the data from the actual standard-setting for MEA reading.)

By picking for the heterogeneous folder the response sets of the high and the low student in each ability interval, we are actually selecting pairs of response sets in which the performance is virtually identical. That is, the low student in one interval performed almost at the same ability level as the high student in the next interval. Thus, we have two indicators at each interval boundary to help determine which homogeneous folders need detailed examination. (NOTE: It is important that the response sets in the heterogeneous folder be ones that were scored very accurately. The computer has only the ratings the scorers assigned to responses to use in placing the students on the ability scale.)

There is little that can be said about the judges' final task, the matching of student response sets in the homogeneous folders to performance level definitions. The final SBCR rating form used in this step is shown in Exhibit D. Exhibit G is a compilation of the data from the final rating forms turned in at the end of the standard-setting process for MEA grade 8 reading. This exhibit also shows the relationship between the judges' ratings and the final cut scores. For example, the first group of judges found 50 percent of the response sets in Folders 2 and 3 to represent distinguished work. The cut score of 2.44 is the IRT ability score that cuts off 50 percent of the response sets in the IRT ability interval from 2.26 and 2.75. This interval is the half-logit interval corresponding to Folders 2 and 3.

The questions that logically arise about the SBCR Method as it was applied for the MEA are the "are there enough" type questions. Are five constructed-response questions enough to match a student's response set to a performance level definition? If the questions solicit responses that illustrate the attributes identified in the performance level definitions, then the answer is "yes." Are there enough student response sets at the borders between quarter-logit intervals in the heterogeneous folder? If the responses were scored accurately, then the answer is "probably yes." Are there enough response sets in the homogeneous folders? The answer to this question is also "probably yes." Of course, in regard to all three of these issues, the numbers could be increased.

It is also important to recognize that the SBCR Method simply establishes cut scores. The final decisions about students for individual or school-level results may be based on more substantial measures. For example, in the 1994-95 MEA, each student answered eight common constructed-response questions and two matrix-sampled constructed-response questions. There were twelve test forms so that in every school 24 matrix-sampled questions were administered. For each student, two scaled scores were produced -- one based on the eight common questions and one based on the two matrix-sampled questions. For school level results, these two scaled scores were averaged. Then with the cut scores applied, percentages of students in the schools within each of the performance levels was determined. Exhibit H shows sections of a 1994-95 grade 8 school report. It is important to point out here that the results for different subgroups of students expressed in percentages at a particular level or above reveal the same relationships as the results that used to be reported in terms of scaled scores.

VALIDATION OF STANDARDS

A great deal of attention in the literature is given to procedural evidence of the validity of standards. Unfortunately, such evidence is relied on heavily, in part because of the lack of time and money budgeted for the collection of empirical evidence. Kane (1994) points out that procedural evidence is more useful in casting doubt about standards and cut scores that in supporting their use. He further states that "thorough implementation of the best available procedures does not guarantee that the resulting passing score is appropriate." Investigations of NAEP standard setting support this statement (National Academy of Education, 1993). It is also unfortunate that procedural evidence is often dominated by evidence that the judges for a particular standard setting effort included a cast of thousands. Whether large numbers of participants are really needed for standard setting may depend on the method used. It is not surprising that, given the nature of judges' decisions and the political nature and visibility of NAEP, NAGB used large numbers of judges for NAEP standard setting. However, those large numbers could not create standards (using the modified Angoff procedures) that would stand up to empirical validation. For a statewide testing program, Maine involved a relatively large number of people in standard setting activities. This was probably more important in the first phase of the effort -- creating the general definitions of performance levels. This step was the one that ultimately determined the approximate relative number of students at different levels. Using the SBCR Method, it may well be that translating general definitions to subject specific definitions and matching student work to definitions require fewer people.

One of the nice things about statewide assessments focusing on the quality and improvement of school programs, instead of high-stakes decisions about individual students, is that totally arbitrary cut scores would probably be defensible (e.g., cut scores initially separating quarters of the statewide distribution). What would then be important are increases in the percentages of students in the upper performance levels over time. The trend today, however, is toward performance standards that communicate expectations in terms of actual competencies. Assessment programs employing more performance-based methodologies have found actual student work most useful in communicating expectations. Unlike the NAEP of old, which addressed this need by identifying multiple-choice questions on which the student surpassing a cut score for a level had a high probability of answering correctly, the newer programs such as the MEA simply show sample work of students at the various levels. Whether this lack of precision is offset by the authenticity of the actual student work in considerations of defensibility remains to be tested. Nevertheless, evaluators of standard setting procedures should probably be wary of applying common criteria across different standard-setting methodologies and contexts.

Time and money constraints are no different in Maine that in other states. Still, a small validation study was conducted in the fall of 1994. Over two hundred student response sets from the 1994-95 grade 8 MEA reading assessment were selected for use. These response sets represented seven groups -- four mid-range groups and three borderline groups. Each of ten judges was asked to categorize a subset of the student response sets, assigning each student response set in the subset to one of the four reading performance levels. Each of the 200 students response sets was evaluated by two judges.

Exhibit I summarizes the raw data gathered from this effort. The high agreement rates for the mid-range response sets speak for themselves. In identifying the borderline response sets for this study, it was not specified that the sets should include equal numbers of sets just above and just below the cut scores. At the time of this writing it remains to determine how those borderline

response sets were actually distributed. If they were evenly distributed, then the results on the advanced/distinguished borderline response sets are what one should find. Otherwise, the results for the other borderline sets may be more reasonable.

The reader should keep in mind that this small investigation compares direct ratings (i.e., performance levels) of 1994-95 common-item response sets to the ratings those sets were assigned by computer applying the cut scores established using an entirely different test -- the 1993-94 common reading items. (Recall, the two years' tests were statistically equated before the cut scores were applied to the 1994-95 results.) Thus, there is evidence of the validity of the standards, the standard setting process, and the tests themselves. Further investigation of the borderline response sets and of estimates of standard errors at the cut points are needed to further the case for the validity of the standards.

The raw data from a similar validation study in mathematics, while revealing some consistency in the assignments of response sets to performance levels, did not yield as high a level of consistency for the midrange response sets as the reading study did. A likely explanation of this has to do with the differing natures of reading and mathematics. Reading competency appears to be a more generalizable competency across reading tasks than mathematical competency does across mathematical tasks. If a person can read one type of reading material effectively, then he or she can probably read other types effectively. However, the subdomains of mathematical competencies tend to be more distinct. This raises concerns about the adequacy and consistency over time of content coverage of relatively short constructed-response tests. Of course, the reliability and generalizability of such tests, augmented by many matrix-sampled questions, can be more than sufficient for purposes of a program focusing on school results. In the MEA, there is greater consistency in content coverage over time at the school level because a total of 32 common and matrix-sampled constructed-response questions are used for MEA school results; and many of them are reused for purposes of equating.

DISCUSSION

One of the commonalities of assessment results in states employing alternative forms of assessment (e.g., open-response questions, extended performance events, portfolios) and performance level reporting is the low percentage of students at the upper performance levels. A few years ago when states responded to the Lake Wobegon scandal in the testing industry by developing their own testing programs using their own current user norms, educators in schools really scoring in the bottom half of their states' score distributions were shaken. Now it appears that educators in schools scoring in the upper half of the distributions are experiencing their rude awakenings. Despite their high scores from norm-referenced reporting, they are being told that the vast majority of their students are performing at low levels. They may claim that their students are simply not used to the types of measures being administered by them. That's all right. If increased exposure to the new formats in regular classroom instruction results in improved results, those results will reflect legitimate performance gains.

REFERENCES

Kahl, S.R., Crockett, T.J., DePascale, C.A., Rindfleisch, S.L. (1994). Using actual student work to determine cut scores for proficiency levels: new methods for new tests. Paper presented at the National Conference on Large Scale Assessment, sponsored by the Council of Chief State School Officers, Albuquerque, NM.

Kane, M. (1994). Validating the performance standards associated with passing scores. *Review of Educational Research*, 64(3), 425-461

Shepard, L., Glaser, R., Linn, R., & Bohrnstedt, G. (1993). *Setting performance standards for student achievement*. National Academy of Education, Stanford University, Stanford, CA.

Comments: The score of 2.44 is the IRT scaled score that cuts off 50 percent of the students in the half-logit interval spanned by Folders 2 and 3. Because of tied scores, the same score of 2.44 cuts off the top 61 percent of the students in that interval also. (The figure 61 percent was the result obtained from a second group of judges.) The cut scores were computed separately for the different groups of judges. This is necessary when the groups identify different homogeneous folders in the rangefinding step. Nevertheless, the data above shows strong agreement among the three groups of judges. Final cut scores are weighted averages of the cut scores found by the groups. For example, the weighted average of 2.44, 2.44, and 2.67 is the final cut score of 2.51, the IRT scaled score separating the distinguished students from the advanced students.

MAINE
EDUCATIONAL
ASSESSMENT

PERFORMANCE LEVELS

Distinguished

Distinguished Maine students demonstrate in-depth understanding of information and concepts. The students grasp "big ideas" and readily see connections among ideas beyond the obvious. These students are insightful, can communicate complex ideas effectively (and often creatively) and can solve challenging problems using innovative, efficient strategies.

Advanced

Advanced Maine students successfully apply a wealth of knowledge and skills to independently develop new understanding and solutions to problems and tasks. These students are able to make important connections among ideas and communicate effectively what they know and are able to do.

Basic

Basic Maine students demonstrate a command of essential knowledge and skills with partial success on tasks involving higher-level concepts, including applications of skills. With some direction, these students make connections among ideas and successfully address problems and tasks. Their communications are direct and reasonably effective, but sometimes lack the substance or detail necessary to convey in-depth understanding of concepts.

Novice

Novice Maine students display partial command of essential knowledge and skills. With direction, these students apply their knowledge to complete routine problems and well-defined tasks. The students' communications are rudimentary, and sometimes ineffective.

September 14, 1994

MAINE EDUCATIONAL ASSESSMENT

DRAFT DRAFT

PERFORMANCE LEVELS IN READING

Distinguished

Distinguished Maine readers demonstrate the ability to see implications and extend applications and connections beyond the obvious. These students are insightful, understand complex ideas, control reading strategies needed to construct meaning from various types of material, and use reference skills effectively.

Advanced

Advanced Maine readers demonstrate full understanding of the text and can link ideas within and among texts. These readers' answers to questions are complete, demonstrate control of reading strategies needed to construct meaning from various types of material, and show knowledge of reference skills.

Basic

Basic Maine readers demonstrate better understanding of some types of texts than others. These students may make important connections among ideas within some texts or in some responses, but the demonstration of this ability may not be consistent across texts. Some readers may be consistent in making obvious connections and relatively low level inferences across texts. These readers demonstrate some control of reading strategies needed to construct meaning from various types of material and know standard reference skills.

Novice

Novice Maine readers demonstrate limited understanding of reading material beyond obvious stated facts. These readers' control of strategies appears to be dependent on the particular type or difficulty level of the text. These students demonstrate limited ability to use reference skills independently.

May 8, 1995

EXHIBIT C

STUDENT BASED CONSTRUCTED RESPONSE
PRELIMINARY RATING FORM

Judge: _____ Session: A.M. P.M.

Reading - High and Low

1. High 3.16 ID# 1021048 _____
1. Low 3.16 ID# 1121234 _____

2. High 2.74 ID# 1020713 _____
2. Low 2.60 ID# 1041031 _____

3. High 2.48 ID# 1051398 _____
3. Low 2.26 ID# 1010212 _____

4. High 2.24 ID# 1010596 _____
4. Low 2.00 ID# 1120125 _____

5. High 1.88 ID# 1021383 _____
5. Low 1.75 ID# 1021133 _____

6. High 1.73 ID# 1101514 _____
6. Low 1.50 ID# 1040571 _____

7. High 1.44 ID# 1030022 _____
7. Low 1.25 ID# 1110753 _____

8. High 1.22 ID# 1080301 _____
8. Low 1.00 ID# 1050775 _____

9. High .99 ID# 1070899 _____
9. Low .76 ID# 1120555 _____

10. High .74 ID# 1071300 _____
10. Low .50 ID# 1040601 _____

11. High .49 ID# 1021397 _____
11. Low .28 ID# 1110255 _____

12. High .21 ID# 1081552 _____
12. Low .02 ID# 1111522 _____

13. High -.01 ID# 1010784 _____
13. Low -.25 ID# 1120423 _____

14. High -.28 ID# 1011584 _____
14. Low -.50 ID# 1020198 _____

15. High -.52 ID# 1020085 _____
15. Low -.75 ID# 1010403 _____

16. High -.76 ID# 1100147 _____
16. Low -1.00 ID# 1011249 _____

17. High -1.02 ID# 1060409 _____
17. Low -1.25 ID# 1121231 _____

18. High -1.26 ID# 1121713 _____
18. Low -1.50 ID# 1111464 _____

19. High -1.51 ID# 1010296 _____
19. Low -1.73 ID# 1101420 _____

EXHIBIT D

STUDENT BASED CONSTRUCTED RESPONSE RATING FORM

SUBJECT:_____

FOLDER #:_____

JUDGE'S NAME:_____ SESSION: AM PM

LITHOCODE #	RATING
_____	_____
_____	_____
_____	_____
_____	_____
_____	_____
_____	_____
_____	_____
_____	_____
_____	_____
_____	_____
_____	_____
_____	_____
_____	_____

EXHIBIT E

Aggregated Ratings From SBCR Preliminary Rating Form

Folder (¼ logit range)	Response Set	Ratings (5/26/94 a.m.)				Ratings (5/26/94 p.m.)				Ratings (5/27/94)			
		D	A	B	N	D	A	B	N	D	A	B	N
1	high	12	–	–	–	8	–	–	–	12	–	–	–
(>2.75)	low	8	1	–	–	7	1	–	–	11	1	–	–
2	high	7	3	–	–	4	3	1	–	5	6	1	–
(2.51 to 2.75)	low	4	4	–	–	5	3	–	–	7	5	–	–
3	high	10	2	–	–	6	2	–	–	4	8	–	–
(2.26 to 2.50)	low	–	7	1	–	1	6	–	–	–	8	4	–
4	high	1	8	–	–	1	7	1	–	–	11	1	–
(2.01 to 2.25)	low	–	7	1	–	1	7	1	–	–	10	2	–
5	high	–	10	2	–	0	7	1	–	–	10	2	–
(1.76 to 2.00)	low	–	9	1	–	0	7	1	–	1	8	3	–
6	high	1	6	1	–	–	8	–	–	–	9	3	–
(1.51 to 1.75)	low	–	7	1	–	–	7	1	–	–	9	3	–
7	high	–	8	3	–	–	8	–	–	1	10	1	–
(1.26 to 1.50)	low	–	8	–	–	1	7	1	–	–	8	4	–
8	high	–	7	3	–	–	5	3	–	–	7	5	–
(1.01 to 1.25)	low	–	5	5	–	–	7	–	–	–	7	5	–
9	high	–	3	9	–	–	6	2	–	1	5	6	–
(.76 to 1.00)	low	–	3	5	–	–	3	4	–	–	–	12	–
10	high	–	–	9	–	–	–	8	–	–	–	12	–
(.51 to .75)	low	–	–	9	–	–	–	8	–	–	–	10	2
11	high	–	–	11	–	–	1	7	–	–	3	9	–
(.26 to .50)	low	–	–	8	–	–	1	7	–	–	–	12	–
12	high	–	–	8	–	–	–	6	2	–	–	9	3
(.01 to .25)	low	–	–	8	–	–	2	6	–	–	–	11	1
13	high	–	–	9	1	–	–	6	1	–	–	11	1
(–.24 to .00)	low	–	–	6	2	–	1	6	–	–	1	10	1
14	high	–	–	4	3	–	–	3	5	–	–	4	8
(–.49 to –.25)	low	–	–	5	1	–	–	4	4	–	–	7	5
15	high	–	–	5	7	–	–	5	3	–	–	1	11
(–.74 to –.50)	low	–	–	2	6	–	–	–	8	–	–	2	10
16	high	–	–	–	10	–	–	–	8	–	–	2	10
(–.99 to –.75)	low	–	–	–	8	–	–	–	8	–	–	–	12
		Folders 2, 3, 8, 9, 14, 15 were selected for further examination.				Folders 2, 3, 9, 14, 15 were selected for further examination				Folders 2, 3, 8, 9, 13, 14 were selected for further examination			

Note: D = Distinguished, A = Advanced, B = Basic, N = Novice

Rangefinding in the SBCR Method

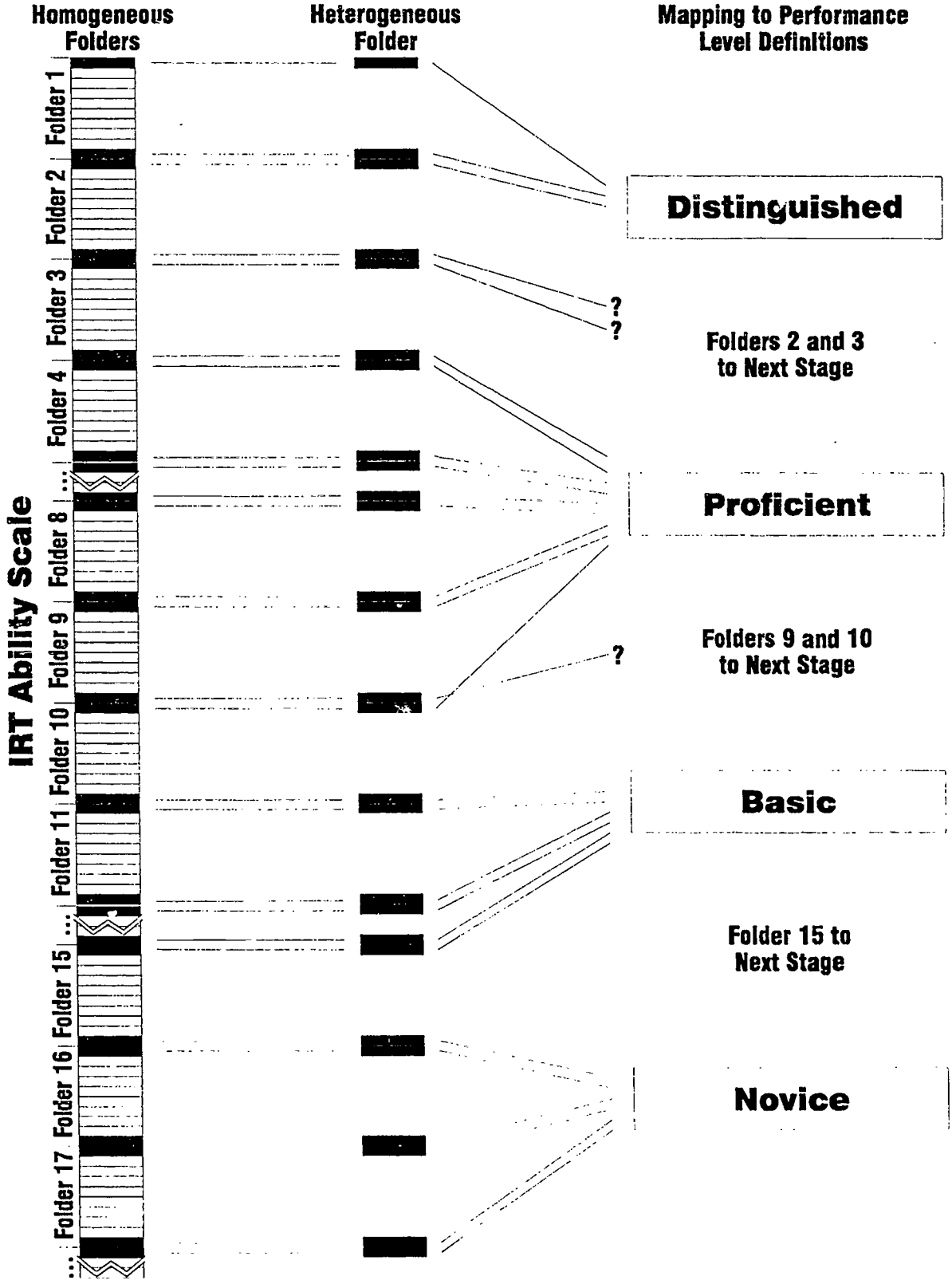

EXHIBIT G

Proportion of Response Sets Rated at Upper Level by Folder(s) by Judge

Judge	5/26/94 a.m. group			5/26/94 p.m. group			9/27/94 group		
	% D in Fldrs. 2 & 3	% A in Fldrs. 8 & 9	% B in Fldrs. 14 & 15	% D in Fldrs. 2 & 3	% A in Fldr. 9	% B in Fldrs. 14 & 15	% D in Fldrs. 2 & 3	% A in Fldrs. 8 & 9	% B in Fldrs. 13 & 14
1	.58	.24	.30	.79	.20	.30	.58	.35	.50
2	.58	.70	.35	.95	.40	.40	.60	.25	.20
3	.84	–	–	.32	.60	0.00	.26	.36	.30
4	.53	.45	.15	.37	.60	.10	.37	.30	.25
5	.63	.65	.55				.58	.35	.45
6	.42	.10	.45				.32	.25	.55
7	.53	.55	.10				.21	.45	.80
8	.37	.23	.55					.15	.65
9	.37	.15	–					.25	.40
10	.42	.30	.45					.05	
11	.21	.00	.20						
$\bar{x}$	.50	.34	.34	.61	.45	.20	.42	.28	.46
Cut point (logits)	2.44	.94	–.31	2.44	.83	–.31	2.67	.94	–.28

Note: D = Distinguished, A = Advanced, B = Basic, N = Novice

Comments: The score of 2.44 is the IRT scaled score that cuts of 50 percent of the students in the half-logit interval spanned by Folders 2 and 3. Because of tied scores, the same score of 2.44 cuts off the top 61 percent of the students in that interval also. (The figure 61 percent was the result obtained from a second group of judges.) The cut scores were computed separately for the different groups of judges. This is necessary when the groups identify different homogeneous folders in the rangefinding step. Nevertheless, the data above shows strong agreement among the three groups of judges. Final cut scores are weighted averages of the cut scores found by the groups. For example, the weighted average of 2.44, 2.44, and 2.67 is the final cut score of 2.51, the IRT scaled score separating the distinguished students from the advanced students.

EXHIBIT H

See attached grphics file: EXHIBITH.EPS

EXHIBIT I

MEA GRADE 8 READING — STANDARDS VALIDATION

	Novice	Basic	Advanced	Distinguished
Novice	49	4	0	0
Novice/Basic	5	53	0	0
Basic	0	48	7	0
Basic/Advanced	1	39	17	0
Advanced	0	4	46	1
Advanced/Distinguished	0	1	29	29
Distinguished	0	1	6	45

Rows: levels actually assigned to 1994 grade 8 students after equating to 1993 and applying standards established using 1993 test — midrange and borderline students' response sets used

Columns: levels assigned to students' response sets by participants in standards validation study

MAINE EDUCATIONAL ASSESSMENT

READING RESULTS

School:
District:
Grade:
Date:

PERFORMANCE LEVELS		STUDENTS AT EACH PERFORMANCE LEVEL				
		School		District		State
		N	%	N	%	%
Distinguished Maine readers demonstrate the ability to see implications and extend applications and connections beyond the obvious. These students are insightful, understand complex ideas, control reading strategies needed to construct meaning from various types of material, and use reference skills effectively.	1994-1995	1	2	1	1	1
	1995-1996					
	1996-1997					
	Cumulative Average	1	2	1	1	1
Advanced Maine readers demonstrate full understanding of the text and can link ideas within and among texts. These readers' answers to questions are complete, demonstrate control of reading strategies needed to construct meaning from various types of material, and show knowledge of reference skills.	1994-1995	26	45	35	27	22
	1995-1996					
	1996-1997					
	Cumulative Average	26	45	35	27	22
Basic Maine readers demonstrate better understanding of some types of texts than others. These students may make important connections among ideas within some texts or in some responses, but the demonstration of this ability may not be consistent across texts. These readers demonstrate some control of reading strategies needed to construct meaning from various types of material and know standard reference skills.	1994-1995	26	45	77	60	54
	1995-1996					
	1996-1997					
	Cumulative Average	26	45	77	60	54
Novice Maine readers demonstrate limited understanding of reading material beyond obvious stated facts. These readers' control of strategies appears to be dependent on the particular type or difficulty level of the text. These students demonstrate limited ability to use reference skills independently.	1994-1995	5	9	16	12	24
	1995-1996					
	1996-1997					
	Cumulative Average	5	9	16	12	24

Reporting Categories	SCHOOL			STATE		
	% Students in category	% Basic or above	% Advanced or above	% Students in category	% Basic or above	% Advanced or above
GENDER						
Male	41	88	25	50	67	15
Female	59	94	62	50	85	30
PARENT EDUCATION						
Not a high school graduate	3			4	56	8
High school graduate	17	70	40	20	70	15
Some college	19	100	55	22	79	23
College graduate	47	93	63	41	85	31
I don't know.	14	100	0	13	62	10
GRADE FIRST ATTENDED SCHOOL IN DISTRICT						
Kindergarten or first grade	70	90	48	69	77	24
Second or third grade	11	83	17	9	76	21
Fourth or fifth grade	9	100	60	8	76	21
Sixth or seventh grade	9	100	60	10	73	19
Eighth grade	2			4	71	14
HIGH SCHOOL CAREER PATHWAY PLANS						
College Prep	72	95	56	77	83	27
Tech Prep	15	88	50	11	61	9
Apprenticeship Programs	2			3	52	7
Occupational Prep	11	67	17	9	57	10

Questionnaire Items	SCH.	STATE		
	% Students in category	% Students in category	% Basic or above	% Advanced or above
Most useful techniques to understanding what is read:				
Teacher/peer respnse to reading journal	21	18	72	20
Conferencing with teacher	25	20	72	20
Conferencing with peers	11	22	78	24
Conferencing with family members	14	13	77	25
Reading aloud	29	28	79	23
How do you receive vocabulary instruction?				
I keep a notebook.	0	14	65	14
We get vocabulary lists.	54	29	77	22
We work on words while reading.	27	28	78	25
All of the above	19	14	80	27
We do not receive vocabulary instruction.	0	16	76	21
How many books have you read in past two months?				
None	14	8	56	10
One	43	13	67	16
Two or three	21	37	75	20
Four or five	18	22	82	26
Six or more	4	21	82	30

EXHIBIT H