DOCUMENT RESUME

ED 376 190 TM 022 210

AUTHOR Poggio, John P.; Glasnapp, Douglas R.
TITLE A Method for Setting Multi-Level Performance Standards on Objective Constructed Response Tests.
PUB DATE Apr 94
NOTE 20p.; Paper presented at the Annual Meeting of the National Council on Measurement in Education (April 4-8, 1994).
PUB TYPE Reports - Descriptive (141) -- Speeches/Conference Papers (150) -- Tests/Evaluation Instruments (160)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS Comparative Analysis; *Constructed Response; *Cutting Scores; Decision Making; *Educational Assessment; Evaluation Methods; *Objective Tests; Psychometrics; *Research Methodology; Standards
IDENTIFIERS Angoff Methods; Ebel Method; High Stakes Tests; *Standard Setting

ABSTRACT

This paper reports on a newly designed judgmental method for setting test performance standard that: (1) overcome many of the practical and psychometric problems associated with the Angoff and Ebel methods; (2) can be used to set multiple cut points on a score scale; (3) may be readily and efficiently implemented with assessments that use objective or constructed response items or both; and (4) allows participation in the standard setting process of persons who may not be educators or not necessarily familiar with the instruction of individuals with whom the examination will be used. In addition to describing the new approach, the paper reports on data gathered using the procedure in an actual standard setting process as part of a high stakes assessment program and provides comparative standard setting results in relation to the Angoff procedure. Results of the psychometric study and evaluation demonstrate the new approach to have decided benefits and features meriting its continued study as well as use. One table (sample rating form) is included. (Contains 4 references.) (Author)

# A METHOD FOR SETTING MULTI-LEVEL PERFORMANCE STANDARDS ON OBJECTIVE CONSTRUCTED RESPONSE TESTS*

John P. Poggio and Douglas R. Glasnapp
The University of Kansas

# A METHOD FOR SETTING MULTI-LEVEL PERFORMANCE STANDARDS ON OBJECTIVE OR CONSTRUCTED RESPONSE TESTS*

John P. Poggio and Douglas R. Glasnapp
The University of Kansas

**Abstract:** The paper reports on a newly designed judgmental method for setting test performance standards that: (1) overcomes many of the practical and psychometric problems associated with the Angoff and Ebel methods, (2) can be used to set multiple cut points on a score scale, (3) may be readily and efficiently implemented with assessments that use objective or constructed response items or both, and (4) allows participation in the standard setting process of persons who may not be educators or not necessarily familiar with the instruction of individuals with whom the examination will be used. In addition to describing the new approach, the paper reports on data gathered using the procedure in an actual standard setting process as part of a high stakes assessment program and provides comparative standard setting results in relation to the Angoff procedure. Results of the psychometric study and evaluation demonstrate the new approach to have decided benefits and features meriting its continued study as well as use.

"Standard setting" on tests continues to receive considerable deserved and needed attention. During the past decade, whether merited or not, mandated testing became the almost exclusive method of education accountability for policy makers. Whether setting "passing scores" on teacher licensing/ certification examinations, or making decisions regarding grade-to-grade promotions, or a determination of minimum competence for the awarding of a high school diploma, setting the examination's performance standard or "passing score" became a commonplace psychometric step in those assessment programs involved in the accountability scenario. And, the need for reliable and valid cutscore procedures is not likely to diminish - the use of tests in Education (as well as in other professions) as the gatekeeper is not lessening but appears to be increasing. Today's school reform initiatives calling for Outcomes Based accountability, national initiatives as NAEP and the emerging assessment regulations for Chapter 1 identification and evaluation, and the discussed American Achievement Test as well the numerous state programs moving toward performance assessment to evaluate higher-order skills are placing ever increasing expectations for approaches to help to establish test performance standards.

---

* Paper presented at the annual meeting of the National Council on Measurement in Education, New Orleans, Louisiana, April 1994. The work reported in this paper was supported in part by grants to the Center for Educational Testing and Evaluation, University of Kansas by the Kansas State Board of Education, Dr. Lee Droegemueller, Commissioner.

Which method to use to guide decisions regarding determination of a test's "cut-score" continues to attract the attention of the testing community. Though dozens of studies have been reported addressing the characteristics of particular approaches and the comparative properties of differing methods, interestingly we are little better off today as to the methods available for setting test performance standards than we were when the need for such procedures first became apparent (Livingston and Zieky, 1982). Today, two broad categories of standard setting methods are recognized: empirical and judgmental approaches. The empirical procedures characteristically require examinees to take the test, then their performance is coupled with independent evaluative judgments provided by instructors (e.g., contrasting groups, borderline identification, etc.). These data are then statistically manipulated and a recommendation for a cutscore is forthcoming.

Judgmental standard setting approaches rely on knowledgeable participants, referred to as judges, who give their professional ratings as to the test item's characteristics as a test item's likely difficulty for examinees and the relevance of an item for the decision to be rendered. Ratings judgments are then statistically handled to arrive at the recommendation(s) for a test's cutscores. In practice, the judgmental approaches appear to be relied on most frequently to guide the cut-score decision-making process. Two approaches, the Angoff and the Ebel methods, are used in Education most commonly, with the Angoff approach perhaps used most often due to its perceived ease of implementation (Jaeger, 1989).

Yet the Angoff method is far from being without criticism or serious reservations regarding its appropriateness: a recent GAO report (1993) was particularly critical of the method because it does not allow judges who participate in the standard setting process to consider the relevance or importance of the test items; thus the method offers no consideration of or accommodation for the validity of the items when arriving at the performance standard. Researchers who have studied the method have been critical of the fundamental assumption of the approach that requires judges to estimate examinees performance, and specifically minimally qualified examinees, on the test's items. The practical utility of the method is often called into question

because only knowledgeable, experienced instructors/educators can implement the procedure (raters need to have experience with the instruction of the skill tested and ability of the examinee group); therefore, policy makers and committed constituents who themselves would have an interest and desire to participate in the standard setting process are excluded. Further, the Angoff method has those serving as judges work from a personal definition of the trait in question (e.g., minimal competence, advanced mastery of the trait, etc.) and to envision a hypothetical group of test takers who meet their personal description/ characterization; participants in the rating activity may have an extremely difficulty time maintaining their focus on such an atypical group as the central and singular referent, and the lack of communality across the judges' referent groups can be observed by the method's relatively lower reliability and the need to engage in judgmental iterations to achieve greater consistency among judges' standards. Further, the procedure is time consuming as participants provide numeric estimations of the likely performance of a hypothetical group of examinees on each test question; also, it is observed that the judgment task itself tends to result in judges adopting a pattern of responding which may compromise the independence of the item evaluations being made. Finally, the method was envisioned and has evolved for traditional multiple choice testing formats wherein a single cutscore is needed; its applicability and utility with assessments (especially performance assessments) that will employ extended, continuous, multinomial score scales are unknown. With the emergence of performance assessment, the use of constructed response test questions in combination with more traditional test formats, the movement toward multi-level performance standards (e.g., advanced, basic, inadequate, etc. classifications), and in consideration of the known shortcomings of the Angoff procedure, alternate standard setting approaches are deserving of consideration.

The paper reports on a newly designed judgmental method for setting test performance standards that: (1) overcomes many of the practical and psychometric problems associated with the Angoff and Ebel methods, (2) can be used to set multiple cut points on a score scale, (3) may be readily and efficiently implemented with assessments that use objective or constructed response items or both, and (4) allows the participation in the standard setting process of persons who may not be educators or not necessarily familiar with the instruction of individuals with whom the examination will be used. In addition

to describing the new approach, the paper reports on data gathered using the procedure in an actual standard setting process as part of a high stakes assessment program and provides comparative standards setting results in relation to the Angoff procedure.

## Overview and Description of the Method

The new proposed method requires each judge who is involved in the standard setting activity to specify the minimum acceptable score performance distribution (or maximum, depending on the standard setter's chosen referent) s/he would define as just barely acceptable in order to identify the referent group of examinees as having demonstrated performance in the target classification (e.g., competent, superior, acceptable, inadequate, etc.). This evaluation is applied by each judge involved in the standard setting process working independently of other judges. Based on the judge's specified minimum acceptable raw score distribution, a summary statistic (e.g., mean) is computed for the distribution which is then pooled with other judges' standards to produce a recommendation for the test's performance standard(s).

To illustrate, consider needing to set performance standards to define Inadequate, Adequate and Advanced performance levels on a test that will be scored to have a maximum raw score of 75 points. As an examinee's score will lead to classification into one of three categories, only two (2) cut-scores, that is decision points, are need: the separation point between Inadequate and Adequate examinee scores, and the separation score between Adequate and Advanced performers. To begin, consider the process for determining the cut point between the Adequate and Advanced levels. The evaluative ratings data to be obtained from a judge participating in the standard setting activity involves having the judge envision 100 representative examinees and then specify the exact raw score distribution of the 100 examinees across the 75-point score scale that the judge would define as the minimum acceptable score performance distribution s/he would find barely acceptable and still be willing to classify the group of 100 examinees as demonstrating performance at the Advanced level. Following the specification of the minimum acceptable score frequency distribution described as Advanced, then the participant is to cast the minimum

score distribution for a group whose performance the judge would defined as just barely Adequate. This judgment is used to obtained data for the Inadequate-Adequate decision point. (As noted above, the frequency distribution specification could be made to identify the "maximum" score performance distribution yet the examinees performance would be considered Inadequate, etc.) In this way each judge in this example is to construct two hypothetical raw score frequency distributions, each distribution defining the minimum score performance distribution representing the judge's "standard" for that judgment category. For this method, like the Angoff and Ebel procedures, a judge specifies the absolute minimum performance for a group to satisfy a criterion; judgments are then manipulated statistically to yield a performance standard which may be used to qualify the performance of individuals or groups.

To determine a particular judge's standard(s), that is the likely location on the score scale where the cut point is being placed by a judge, procedures for computing summary statistics for grouped frequency distributions (means, medians, standard deviations, etc.) are used. For this procedure, the average score determined from the minimum acceptable score distribution (mean or median depending on distribution properties and user expectations) specified by the participant becomes that judge's decision point. With the ready access to technology, these once efficient "grouped statistics" procedures today are rarely presented in introductory statistics textbooks. Such earlier texts as those by Guilford, Walker and Lev, Garrett, and Ferguson provide ample treatment of procedures to compute statistics from frequency distributions. The "performance standard" resolved across participants may then be determined by computing the average across the judges for each category of classification. In the next section we present a more detailed presentation of the actual implementation of the proposed standard setting procedure.

## Implementation of the Procedure

The standards setting process for obtaining participants' judgments is to be enacted as a two step activity. First, those participating in the standard setting task are asked to review carefully each test question and provide a judgment as to each item's "cognitive demand or complexity." A rating scale needs to be

prepared and used to gather such evaluations (an example of such a process is provided on page 15). If preferred, participants can be queried to evaluate the extent of the curricular or instructional validity of each test item's content. This first step in the standard setting process does not directly contribute to the level of the standards to emerge as is the case with the Ebel standard setting methodology; it purpose is to require participants to become actively engaged in recognizing and thus aware of the properties of the test and its items, and thereby this step is expected to influence the standards that will emerge. This step ensures that participants are very familiar with the content, substance and format of the entire assessment on which performance standards are to be set.

Following this independent item validity appraisal which is in place to insure that judges are indeed familiar with the content and structure of the test, participants are briefed on the nature of the standard setting activity and are to be given experience and practice using one or two illustrative assessment exercises to gain experience in the key feature of this standard setting approach: the task of specifying the minimum acceptable raw score frequency distribution the judge defines as representing the performance of a group that meets the target classification descriptor, e.g., advanced, competent, borderline, etc. Operationally, judges are instructed as they establish their score performance distribution that, "Were even one of the examinees in your specified performance distribution to have scored lower than you have specified, you would change your appraisal of the entire groups' standing to the lower classification." The judgmental task for the participant is to specify the pattern of scores for a group of examinees that s/he would identify as just barely meriting the classification label to be assigned. The mean (or median) of the judge's grouped frequency distribution, computed from the judge's proposed minimum acceptable score distribution, represents the score point that distinguishes performance between the adjacent performance classifications for the participant.

On the next page is displayed a completed worksheet for the situation mentioned previously - an assessment having a possible 75 points for which performance standards are to be established to distinguish among Advanced, Adequate and Inadequate classifications. For this illustration in order to make the task more manageable for judges, a grouped frequency distribution utilizing a score interval of five (5) points has been used. Regardless of the interval width

employed, the task for the participant is to wrestle with the standards question by "casting" the frequency column(s) on the work sheet detailing the raw score frequency distribution pattern the participant would establish as just barely acceptable for a group deserving of each classification. The hand written frequencies on the sample worksheet demonstrate the pattern of scores one illustrative panelist provided as the minimum acceptable performance distribution for each classification. The assignment for the participant is to fill-in each frequency column with that pattern of scores s/he would establish as minimally acceptable given the entire group of examinees will qualify for the classification.

*SAMPLE STANDARD SETTING WORKSHEET*

| Test Scores | **Minimum Acceptable Score Distribution for the GROUP to be Identified ADVANCED*** | **Minimum Acceptable Score Distribution for the GROUP to be Identified ADEQUATE*** |
|---|---|---|
| | frequency | frequency |
| 71 - 75 | 10 | 1 |
| 66 - 70 | 17 | 2 |
| 61 - 65 | 23 | 3 |
| 56 - 60 | 27 | 7 |
| 51 - 55 | 14 | 14 |
| 46 - 50 | 6 | 20 |
| 41 - 45 | 2 | 23 |
| 36 - 40 | 1 | 17 |
| 31 - 35 | - | 7 |
| 26 - 30 | - | 3 |
| 21 - 25 | - | 2 |
| 16 - 20 | | 1 |
| 11 - 15 | | - |
| 6 - 10 | | - |
| 0 - 5 | | - |
| Total N | 100 | 100 |

*The mean (or median) of the particular specified score distribution serves as the judge's decision point for that classification. For the pattern of frequencies indicated on this example worksheet and using means to determine the cutscores for the categories, then the standards are 60.55 and 45.20 respectively for this panelist. Thus, scores of 61 or greater result in a classification of Advanced, scores of 46 to 60 result in a classification of Adequate, and scores of 45 or lower would result in the classification of Inadequate performance.

The underlying assumption of the proposed method is that a judge in specifying the minimum acceptable score distribution for a category, then the mean of that distribution (or median if distribution shape is a consideration) is the typical or common point that can be considered to provide a stable and reasonable estimate for a cutscore when an examinee's performance or the performance of a group is to be defined by the classification. In the actual utilization of this approach, participants are to be informed that the mean (or median) of the frequencies they specify serve as their definition of the test classification performance standard.

A consideration of the proposed standard setting procedure reveals the following features regarding its utilization:

- unlike other judgmental methods (i.e., Angoff, Ebel and Nedelsky), the proposed approach does not depend on establishing cut-scores based on the tenuous assumption that judges are capable of forecasting the likely performance of examinees on each test item;

- as other judgmental procedures focus attention on estimating performance on each test item, response patterns and bias likely affect the actual standards that result; the proposed procedure forces participants to wrestle and come to grips with a single decision making process - suggesting the score distribution of examinee performance that represents, in the participant's judgment, a specific level of performance; and,

- using the proposed procedure given that performance standards are derived based on judgments utilizing score distributions, the procedure may be applied to a multitude of assessment formats or configurations of items including constructed response or objective testing formats or a combination of such formats. Judgments regarding standards can be obtained for individual multi-point performance assessment items or score composites formed across items.

To further illustrate implementation of the proposed methodology attached are actual worksheets used with an application of the proposed procedure. Included are exemplar documents to illustrate:

(1) instructions and a rating worksheet used to obtain item validity/cognitive complexity evaluations (page 15);

(2) a standard setting worksheet used to obtain a participant's raw score distribution estimates wherein the assessment used objective type test items and performance standards were to be set for the Knowledge Base (5-score points maximum ) and Non-Routine Problem Solving (12-point maximum) scales (page 16);

(3) a standard setting worksheet used for two separate performance assessment items wherein for each item the maximum score was 5 points and multi-level standards were to be established (page 17); and,

(4) a worksheet used to gather judgments to set multi-level standards for total scores on the entire (objective and performance items combined) assessment battery (page 18).

These worksheet forms were used to arrive at test performance standards information on a series of mathematics assessments for which multi-level cut points were needing to be determined as described below. Those wishing to obtain a complete set of the directions and instructions used to obtain judgments for this project should contact the authors.

## A Tryout of the Proposed Procedure

To study the adequacy of the proposed approach for standard setting, an actual field tryout was carried out. During spring 1993, the Kansas State Board of Education was needing to establish performance standards on its state mathematics assessments. The assessments were constructed to measure outcomes modeled after the NCTM curriculum standards; the examinations developed were comprised of assorted and varying objective format test questions and extended performance assessment/constructed response items. Performance assessment questions were scored separately using a multi-point scale (e.g., 0 to 5 points maximum); traditional objective items are dichotomously

scored, while non-traditional multiple correct objective test items were scored using a three point scale (0 to 2 points maximum). The state had developed an elaborate and extensively defined five classification proficiency scale (Excellence-Strong-Progressing-Borderline-Inadequate) to be used for classifying student scores and building and district averages on the mathematics skills (problem-solving, reasoning, communication, composite score, etc.) measured by the assessments.

Fifty five (55) individuals were impaneled to assist with establishing the needed multi-level test performance standards. The newly devised procedure as described previously was implemented concurrent with utilization of the Angoff approach. During the standard setting activity the newly devised method was not distinguished or discussed in any way with participants as a "trial" or an experimental procedure. It was put in place with the same attention and consideration given the securing of judgmental rating information for the Angoff approach. Results from the comparative application of the two approaches is presented in the next section.

## Findings from the Tryout

The Kansas mathematics assessments was designed to produce six distinct scale scores in addition to a composite (total) score. The performance of student's, buildings and each school district, was to be classified into one of five proficiency scale categories. As such a considerable amount of standard setting data were assembled. Approximately 30 experience mathematics educators participated in the standard setting activities at each grade.

The table provided on page 14 reports comparative statistics from the standard setting procedures at two of the tested grades. The table only provides comparative results for the objective portion of the assessments at grades 4 and 10 where both the Angoff and the newly devised procedure could both be applied to the assessments (the Angoff procedure can only be applied to dichotomously score objective test items). Reported in the table are summary statistics for the standards being set across the individual judges/panelists. Included for each assessment under each procedure are: an indication of the

skewness of the individual standards being produced by judges, the minimum and maximum standards assigned by individual judges, select percentile point values, and the mean and standard deviation of the standards produced by a method across judges for the objective portion of the assessment. To summarize results and findings:

- the newly devised standard setting procedure was found to have an average interrater reliability of .89, while the Angoff procedure was found to be at .80.

- in completing the rating tasks three of the 55 judges when using the new procedure were unable to follow the procedure as called for, whereas under the Angoff method, ratings information provided by seven panelists had to be discarded as unusable due to improper completion of the task.

- the variance of judges cut-scores produced by the Angoff method was always considerably greater than the variance of judges cut-scores using the new procedure at all score classification points. Angoff ratings by the participants tended to produce more pronounced skewed distributions as well.

- as unsolicited anecdotal qualitative reports, participants reported understanding the nature and intent of the standard setting activity under the proscribed instructions of the new standard setting procedure, while many judges commented they were unsure about the basis of the Angoff method and thus questioned its credibility to produce trustworthy and accurate standards. A frequent comment also noted regarding application of the Angoff method was how to handle specification of difficult items that approach a chance level of response. No such anomalies were noted by participants with reference to the new procedure.

- cut-scores computed as means across judges were systematically higher based on application of the Angoff method in all performance

categories studied; the cut-scores determined based on the new procedure were observed repeatedly to be one to three score points below the Angoff standards. When trimmed means (90 and 95%) were evaluated for determining cut scores, differences between the approaches were narrowed with the Angoff standards moving toward the new procedure standards, there being little shift in the new method trimmed mean and untrimmed results.

- correlations of the judges' standards derived from the new procedure and the Angoff method with the panelists' evaluation of the appropriateness of items as measures of the constructs being assessed by the examination favored the new method. When standards produced by judges using the new method were correlated with judges ratings of the appropriateness of the items for the assessment correlations were observed to range from .30 to .50, whereas correlations with the Angoff standards ranged from .10 to .35. When judges' ratings of item complexity/task difficulty were correlated with their standards no difference in extent of relationships between the methods were observed. For both procedures, the correlations tended to ranged from .35 to .65. Both approaches appear to yield standards sensitive to difficulty, but standards secured from the new procedure related stronger to an independent measure of content validity evidence.

- the cutscores that resulted when the newly devised procedure was used with multi-point scored performance assessments and when setting standards for the total assessment composite score were comparable in level and reliability to the cutscores produced when the new procedure was used with objective, dichotomously score test items. This finding suggests that the new method is appropriate and viable when used for setting standards on other than objective assessment devices.

## Discussion

Deficiencies often cited regarding the Angoff and Ebel approaches to setting test performance standards have been noted. The newly devised approach offers a correction or an accommodation to many of these shortcomings. Further, initial psychometric study and evaluation suggests the new approach has decided benefits meriting its continued study and evaluation. Standards setting procedures only serve as guides toward likely, reasonable, defensible and credible decision points. The method being proposed does appear to be promising in consideration of such criteria.

## BIBLIOGRAPHY

Jaeger, Richard M. (1989). "Certification of Student Competence." In R. Linn, ed. Educational Measurement, 3rd ed. New York: American Council on Education and Macmillan Publishing Co.

Kansas Mathematics Assessments, Grades 4, 7 and 10 (1993). Center for Educational Testing and Evaluation, The University of Kansas, Lawrence, Kansas, 66045.

Livingston, S.A. and M. Zieky. (1982). Passing Scores: A Manual for Setting Standards of Performance on Educational and Occupational Tests. Princeton, N.J.: Educational Testing Service.

United States General Accounting Office. (1993). Educational Achievement Standards. Program Evaluation and Methodology Division, (B-251957), Washington, D.C., U.S. Government Printing Office.

**Comparative Standards based on Application of Procedures at Two Grade Levels**

***Grade 4 Mathematics Standards***

**Poggio/Glasnapp Procedure (26 pts)**

| *for classification into Category* | Skew | Min | P25 | P50 | P75 | Max | Mean | SD | % Correct |
|---|---|---|---|---|---|---|---|---|---|
| Excellence | -0.45 | 15.40 | 18.10 | 22.00 | 23.80 | 25.00 | 21.02 | 2.24 | 80.85 |
| Strong | -0.64 | 11.02 | 16.15 | 17.65 | 19.00 | 22.00 | 17.44 | 2.76 | 67.08 |
| Progressing | 0.71 | 8.35 | 12.25 | 13.00 | 14.05 | 19.75 | 13.38 | 2.38 | 51.45 |
| Borderline | 0.91 | 4.60 | 5.65 | 7.00 | 9.10 | 14.65 | 7.75 | 2.55 | 29.81 |
| Inadequate | | | | | | | | | |

**Angoff Procedure (26 pts)**

| *for classification into Category* | Skew | Min | P25 | P50 | P75 | Max | Mean | SD | % Correct |
|---|---|---|---|---|---|---|---|---|---|
| Excellence | -1.46 | 14.05 | 21.13 | 24.00 | 24.33 | 25.51 | 22.39 | 3.02 | 86.13 |
| Strong | -0.72 | 12.77 | 17.07 | 19.84 | 21.84 | 23.60 | 19.40 | 3.04 | 74.63 |
| Progressing | -0.06 | 10.59 | 13.65 | 16.10 | 19.13 | 21.69 | 16.25 | 3.60 | 62.49 |
| Borderline | 0.70 | 7.01 | 9.40 | 10.82 | 14.77 | 18.96 | 12.10 | 3.74 | 46.54 |
| Inadequate | | | | | | | | | |

***Grade 10 Mathematics Standards***

**Poggio/Glasnapp Procedure (29 pts)**

| *for classification into Category* | Skew | Min | P25 | P50 | P75 | Max | Mean | SD | % Correct |
|---|---|---|---|---|---|---|---|---|---|
| Excellence | -0.90 | 15.19 | 21.88 | 23.73 | 25.08 | 27.10 | 22.78 | 2.25 | 78.57 |
| Strong | -0.38 | 13.48 | 17.84 | 19.23 | 21.29 | 23.80 | 19.30 | 2.60 | 66.56 |
| Progressing | 0.59 | 10.90 | 13.76 | 15.49 | 16.38 | 22.60 | 15.28 | 2.59 | 52.68 |
| Borderline | 0.83 | 3.40 | 6.66 | 9.10 | 11.50 | 19.36 | 9.36 | 3.90 | 32.26 |
| Inadequate | | | | | | | | | |

**Angoff Procedure (29 pts)**

| *for classification into Category* | Skew | Min | P25 | P50 | P75 | Max | Mean | SD | % Correct |
|---|---|---|---|---|---|---|---|---|---|
| Excellence | -0.52 | 21.74 | 23.36 | 26.80 | 27.49 | 27.95 | 25.59 | 2.13 | 88.25 |
| Strong | -0.69 | 14.68 | 19.30 | 23.97 | 25.31 | 27.95 | 22.47 | 3.33 | 77.49 |
| Progressing | -0.59 | 9.17 | 14.66 | 19.84 | 21.38 | 27.95 | 18.25 | 4.77 | 62.94 |
| Borderline | -0.18 | 3.91 | 8.28 | 13.92 | 17.44 | 27.95 | 12.84 | 5.51 | 44.28 |
| Inadequate | | | | | | | | | |

## Cognitive Complexity/Demand of the 1993 Test Items

*Directions*: Follow the instructions below. To carry out this activity you will use the portions of the test booklets that are attached. Note that the "correct" answers have been circled. Your responses to this rating activity are to be recorded on the response form that follows on the next page.

The mathematics items in Parts 1 and 3 of the 1993 Kansas Assessment focus on measuring mathematics knowledge, skill and abilities in the areas of estimation, non-routine problem solving, reasoning and communications as well as the core knowledge base which includes traditional routine problem solving.

Consider the cognitive demand, complexity and difficulty characteristics that a test item can represent to a student. *For each test item* in Parts 1 and 3, you are to rate the cognitive complexity (that is, the item's task demand) in terms of the quantitative thinking requirements called upon for the student to respond correctly to the item. Review an item, then evaluate the complexity of the mathematical thinking in which a student at this grade would need to engage to answer the item correctly. Record your complexity judgment using the following 10-point scale. On the response form, rate each item from 1 (very low cognitive complexity) through 5 and 6 (moderate) to 10 (very high cognitive complexity).

| Very low complexity; Little thinking required; Only rote memory needed to answer item correct; Automatic, common knowledge leads to answer. | Moderate Complexity | Extremely high complexity; Higher order thinking required; Many decisions; critical analyses; Quantitatively ingenious, clever or taxing. |
|---|---|---|

1 2 3 4 5 6 7 8 9 10

**10-POINT SCALE**

Record your rating for each item on the response form that follows. The test questions for Part 1 and Part 3 follow the response form. Bear in mind that the Part 1 Estimation was timed and students permitted only six (6) minutes to work on this part of the test. Part 3 includes objective test questions which was administered without restrictive time limits.

**Begin by FIRST reviewing and going over all test items before you begin your ratings of the individual items.** Once you are familiar with the range of test items, then start your evaluation. You may separate the pages to facilitate the rating task.

# Activity -- Response Form

*Directions*: For each subscale, given the number of score points for that subscale, create the score distribution that in your professional opinion should be established to define the minimum/maximum performance of a class of students who should as a group be identified by the category. What should be required as the minimum (maximum) acceptable performance distribution for a group of 100 regular education students to be judged as performing at each Proficiency Scale level?

Objective Items

**SUBSCALE: KNOWLEDGE BASE**

| (minimum) EXCELLENCE | | (minimum) STRONG | | (minimum) PROGRESSING | | (maximum) INADEQUATE | |
|---|---|---|---|---|---|---|---|
| Score | % | Score | % | Score | % | Score | % |
| 5 | _____ | 5 | _____ | 5 | _____ | 5 | _____ |
| 4 | _____ | 4 | _____ | 4 | _____ | 4 | _____ |
| 3 | _____ | 3 | _____ | 3 | _____ | 3 | _____ |
| 2 | _____ | 2 | _____ | 2 | _____ | 2 | _____ |
| 1 | _____ | 1 | _____ | 1 | _____ | 1 | _____ |
| 0 | _____ | 0 | _____ | 0 | _____ | 0 | _____ |
| | 100 | | 100 | | 100 | | 100 |

Objective Items

**SUBSCALE: NON ROUTINE PROBLEM SOLVING**

| (minimum) EXCELLENCE | | (minimum) STRONG | | (minimum) PROGRESSING | | (maximum) INADEQUATE | |
|---|---|---|---|---|---|---|---|
| Score | % | Score | % | Score | % | Score | % |
| 12 | _____ | 12 | _____ | 12 | _____ | 12 | _____ |
| 11 | _____ | 11 | _____ | 11 | _____ | 11 | _____ |
| 10 | _____ | 10 | _____ | 10 | _____ | 10 | _____ |
| 9 | _____ | 9 | _____ | 9 | _____ | 9 | _____ |
| 8 | _____ | 8 | _____ | 8 | _____ | 8 | _____ |
| 7 | _____ | 7 | _____ | 7 | _____ | 7 | _____ |
| 6 | _____ | 6 | _____ | 6 | _____ | 6 | _____ |
| 5 | _____ | 5 | _____ | 5 | _____ | 5 | _____ |
| 4 | _____ | 4 | _____ | 4 | _____ | 4 | _____ |
| 3 | _____ | 3 | _____ | 3 | _____ | 3 | _____ |
| 2 | _____ | 2 | _____ | 2 | _____ | 2 | _____ |
| 1 | _____ | 1 | _____ | 1 | _____ | 1 | _____ |
| 0 | _____ | 0 | _____ | 0 | _____ | 0 | _____ |
| | 100 | | 100 | | 100 | | 100 |

## Response Form - Activity

Given the number of score points that can be awarded (5-points) to a response on the Performance Assessment (Part 2) portion, create the score distributions that in your professional opinion should be established to define the minimum (maximum) performance of a class of students who have as a group demonstrated and can be identified as: Excellence, Strong, Progressing, and Inadequate. Record your score performance distributions below for each test item.

Performance Assessment #1

| (minimum) EXCELLENCE | | (minimum) STRONG | | (minimum) PROGRESSING | | (maximum) INADEQUATE | |
|---|---|---|---|---|---|---|---|
| Score | % | Score | % | Score | % | Score | % |
| 5 | _____ | 5 | _____ | 5 | _____ | 5 | _____ |
| 4 | _____ | 4 | _____ | 4 | _____ | 4 | _____ |
| 3 | _____ | 3 | _____ | 3 | _____ | 3 | _____ |
| 2 | _____ | 2 | _____ | 2 | _____ | 2 | _____ |
| 1 | _____ | 1 | _____ | 1 | _____ | 1 | _____ |
| 0 | _____ | 0 | _____ | 0 | _____ | 0 | _____ |
| | 100 | | 100 | | 100 | | 100 |

---

Performance Assessment #2

| (minimum) EXCELLENCE | | (minimum) STRONG | | (minimum) PROGRESSING | | (maximum) INADEQUATE | |
|---|---|---|---|---|---|---|---|
| Score | % | Score | % | Score | % | Score | % |
| 5 | _____ | 5 | _____ | 5 | _____ | 5 | _____ |
| 4 | _____ | 4 | _____ | 4 | _____ | 4 | _____ |
| 3 | _____ | 3 | _____ | 3 | _____ | 3 | _____ |
| 2 | _____ | 2 | _____ | 2 | _____ | 2 | _____ |
| 1 | _____ | 1 | _____ | 1 | _____ | 1 | _____ |
| 0 | _____ | 0 | _____ | 0 | _____ | 0 | _____ |
| | 100 | | 100 | | 100 | | 100 |

You have now reviewed and studied all test items and test questions in the 1993 Kansas Mathematics Assessment. Based on your study if a total composite mathematics standard has to be set combining all parts of the test, what would your performance distributions be for the different Proficiency Scale classifications when all test items are considered. Record that which in your professional opinion should be established as the threshold score performance distributions for the entire set of test items in the assessment (multiple choice, estimation and performance assessment). What should be required as the minimum (maximum) acceptable performance distribution for a group of 100 regular education students to be judged as performing at each Proficiency Scale level on the total set of items?

For this score performance distribution evaluation, we have used a generic interval width of 5 percent across which you need to distribute student performance. For example the range 95 - 100% means that students in this interval obtained 95 to 100 percent of the points on the total mathematics assessment. Record your performance distributions for the Proficiency Scale categories shown.

ALL ITEMS IN THE 1993 KANSAS MATHEMATICS ASSESSMENT

| (minimum) EXCELLENCE Percent Correct | % | (minimum) STRONG Percent Correct | % | (minimum) PROGRESSING Percent Correct | % | (maximum) INADEQUATE Percent Correct | % |
|---|---|---|---|---|---|---|---|
| 95-100 | _____ | 95-100 | _____ | 95-100 | _____ | 95-100 | _____ |
| 90-94 | _____ | 90-94 | _____ | 90-94 | _____ | 90-94 | _____ |
| 85-89 | _____ | 85-89 | _____ | 85-89 | _____ | 85-89 | _____ |
| 80-84 | _____ | 80-84 | _____ | 80-84 | _____ | 80-84 | _____ |
| 75-79 | _____ | 75-79 | _____ | 75-79 | _____ | 75-79 | _____ |
| 70-74 | _____ | 70-74 | _____ | 70-74 | _____ | 70-74 | _____ |
| 65-69 | _____ | 65-69 | _____ | 65-69 | _____ | 65-69 | _____ |
| 60-64 | _____ | 60-64 | _____ | 60-64 | _____ | 60-64 | _____ |
| 55-59 | _____ | 55-59 | _____ | 55-59 | _____ | 55-59 | _____ |
| 50-54 | _____ | 50-54 | _____ | 50-54 | _____ | 50-54 | _____ |
| 45-49 | _____ | 45-49 | _____ | 45-49 | _____ | 45-49 | _____ |
| 40-44 | _____ | 40-44 | _____ | 40-44 | _____ | 40-44 | _____ |
| 35-39 | _____ | 35-39 | _____ | 35-39 | _____ | 35-39 | _____ |
| 30-34 | _____ | 30-34 | _____ | 30-34 | _____ | 30-34 | _____ |
| 25-29 | _____ | 25-29 | _____ | 25-29 | _____ | 25-29 | _____ |
| 20-24 | _____ | 20-24 | _____ | 20-24 | _____ | 20-24 | _____ |
| 15-19 | _____ | 15-19 | _____ | 15-19 | _____ | 15-19 | _____ |
| 10-14 | _____ | 10-14 | _____ | 10-14 | _____ | 10-14 | _____ |
| 5-9 | _____ | 5-9 | _____ | 5-9 | _____ | 5-9 | _____ |
| 0-4 | _____ | 0-4 | _____ | 0-4 | _____ | 0-4 | _____ |
| | 100 | | 100 | | 100 | | 100 |