CRITERION ITEM STANDARD SETTING

Martin E. Grosse

National Board of Medical Examiners

Philadelphia, PA

Benjamin D. Wright

University of Chicago

ABSTRACT

A method for establishing a minimum passing standard for a multiple choice test is described. The procedure is based on expert judgment about the content of test items. The method can also be used to monitor or adjust a standard that is already in place. The procedure carries advantages that are not available from other approaches to standard setting. It is easy to implement and easy for participants to understand. Most important, the sources of a standard based on this approach can be inspected and evaluated easily by people who are not involved in the standard setting process.

ACKNOWLEDGMENT

The authors wish to express their thanks to the representatives of the American Board of Orthopaedic Surgery who participated in this study, and to the entire Board for its commitment to the study of this important issue.

INTRODUCTION

The American Board of Orthopaedic Surgery develops and administers a new multiple-choice test each year. This test assesses the general knowledge of orthopaedic surgery required of all candidates who will be certified by the Board.

In 1985 the Board conducted a study to confirm the validity of its norm referenced standard that has been located at a fixed standard deviation distance below the mean score of a selected "reference group" of candidates. This standard relies on the average performance and the variability of this group of candidates. The methodology described in this paper was employed to validate this standard. The method depends upon the judgment of content experts about the content of test items, rather than on the ability of candidates.

This paper describes an extension of a methodology developed by Grosse and Wright (1986). Their work described the use of the difficulty of a single test item to determine the standard for the test. The methodology described here differs significantly because it relies on the difficulties of a set of standard setting (or criterion) items, rather than on a single item, to arrive at a standard.

Implementation of this validation methodology requires a Rasch model calibration of the test. Rasch calibration yields a single measurement scale upon which test item difficulty and candidate ability are simultaneously located. The distance between an item's difficulty and a candidate's ability on this scale can be expressed as the probability of a correct response to the item by the candidate. Once item difficulties on the Rasch measurement scale have been established through calibration, any subset of test items can be used to estimate ability measures on the same scale as that of the total test scores. This permits a standard based on only a small set of items to be located on the measurement scale of the entire test.

METHOD

The Board previously employed a norm referenced approach to standard setting. Board members decided that the standard for the test would be set at one standard deviation below the mean of their reference group. This implies a failure rate of approximately 16% for reference group candidates. Board members felt able to make this judgment because of their experience as directors of residency training programs; they are familiar with the capabilities of individuals in the reference group. They also know the test content through their participation in writing and reviewing test items.

The Board was interested in gathering evidence that would provide additional information about the suitability of this minimum passing standard. It therefore undertook a study based on judgments about the content of test items. These sets of standard setting items are referred to as criterion items.

There are three phases to this criterion item approach. In the first, a set of criterion items is selected from the items in the test. In the second, a rule is developed to specify how these items will be used to set the standard. In the third, the consequences of applying the rule to the items are reviewed and the selection of items and the rule may be adjusted to produce an acceptable result.

Four judges were chosen by the Board to participate in the study. Criterion items were selected prior to the administration of the test. In the first phase of the study, each judge used the following rules to guide his independent selection of his personal set of criterion items from a test that contained 240 items.

1. The item is highly relevant to practice.
2. The item tests attitudes, skills, or knowledge required frequently that should be maintained at an efficient and effective level of quality by every practitioner.
3. The information tested by the item should be known by the lowest ability candidate who is clearly certifiable.
4. The item has one clearly correct answer. Avoid items with a "second best" answer that is not clearly wrong.

In the second phase, a rule was constructed to specify how the items selected would be used to set a standard. Each judge was allowed to specify his own percent correct minimum passing score for his personal set of criterion items. Because judges selected their own personal sets of criterion items and then set their own personal standards, the standards are uniquely individualized and independent.

In the third phase of the study, after administration of the test, an item analysis was conducted and each judge was provided with the percentage of candidates selecting each option of his criterion items. It is assumed that judges are correct in their initial decision that the content tested by each criterion item should be known to passing candidates. However, the structure of an item -- how its stem and alternative responses are expressed -- can distract candidates from selecting the correct option. Therefore, those items most difficult for candidates were reviewed. A review of the percentage of candidates selecting each alternative option helps judges to know how his criterion items were perceived by candidates.

During this review, each judge deleted some items from his criterion set. Following this deletion, each judge reviewed all of the criterion items remaining in his set and then adjusted his standard, if he felt that was necessary. Judges worked independently during all phases of this study.

Each judge expressed his standard as the percentage of his criterion items that he wanted answered correctly in order for candidates to pass. If candidates had been scored on a judge's particular set of criterion items, his standard would apply directly to that score. However, judges selected different sets of criterion items, and candidate were scored on all items, not on the criterion sets alone. Therefore, a method was needed to transfer each judge's standard from the scale of his personal set of items to the scale of the test as a whole. Rasch calibration enables this process.

Implementation of criterion item standard setting requires a Rasch calibration of the test (Rasch, 1960/1980; Wright and Panchapakesan, 1969; Wright and Stone, 1979; Wright and Masters, 1982). Several computer programs for calibrating items and measuring ability are available (Wright and Mead, 1976; Wright and Masters, 1981; Wright and Linacre, 1985; Wright et. al., 1985).

Calibration of the total test locates all item difficulty calibrations and person ability measures on a common logit scale. The Rasch model and the relationships among item difficulty, person ability, and the probability of a correct response are described in Appendix A. After calibration, it is a straight forward process to use the item difficulties from the criterion items of a particular judge to transfer his standard to the scale of the entire test. The mean and standard deviation of criterion item difficulties and the PROX estimation formulas presented by Wright and Stone (1979, page 21) enable each judge's personal criterion item standard to be expressed on the common scale of the entire test. The PROX formula applied is this.

$$b = H + (X * \ln(r/(L-r)))$$

Where b = criterion standard on the scale of the total test in logits

H = average difficulty of the criterion items

$$X = (1.0 + (w^{2} / 2.89))^{1/2}$$

Where w = standard deviation of criterion item difficulties

r = the percent correct standard selected by a judge

L = 100%

Unlike the procedure described by Grosse and Wright (1986), in which standards were based on the difficulty of the most difficult standard setting item, the standards derived in this study are based on complete sets of standard setting items.

RESULTS

Table 1 summarizes the results of judgments made about criterion items and the standards set. Judges varied considerably in the number of items selected and in the percent correct standards specified. Some judges specified an interval (e.g., 80% to 85%) in which the standard should be located. When this occurred, the midpoint (i.e., 82.5%) was used as the standard.

Insert Table 1 about here.

Figure 1 shows the consequences. The single Rasch ability-difficulty logit scale that results from the calibration is drawn 6 times to clarify the situation. Logit values to the left of zero are negative. On the top four lines, asterisks represent the difficulty estimates for each judge's set of criterion items. The fifth line shows the difficulty estimates for each item in the test. The bottom line shows ability measures for reference group candidates. The heavy vertical line marks a norm referenced standard for the test at one standard deviation below the reference group mean score.

The light vertical lines mark the independent standards set by the four judges. The standard for judge D is the same as the norm referenced standard. Candidates at or to the right of the vertical standard line pass.

Figure 1 shows the relationships among criterion items, all items, and the ability of reference group candidates. Most of the criterion items are to the left of the zero point on the logit scale. This means that the criterion items are easier, on the average, than the total set of items. The candidates are higher on the logit scale than most of the test items, meaning that candidates had a greater than 50/50 chance of answering most items correctly. The criterion referenced standards for the four judges are close to the norm referenced standard.

Insert Figure 1 about here

The judge's percent correct standards based on criterion items are converted to logit measures on the scale of the entire test using the PROX extimation formula described in the methodology section of this paper. Once this step is complete, the percent correct score on the scale of the entire test is found through a table look up procedure. The table, provided by all calibration programs, shows the logit equivalents for all possible raw scores on the test.

The results of this conversion process are shown in Table 2. The criterion standards on the scale of the 1985 test ranged from a low of 69% correct for judge C to a high of 75% correct for both judges A and B. The norm referenced standard for the test was at 72% correct. The standard set by Judge D was the same as the norm referenced standard.

Insert Table 2 about here

Table 2 also shows the candidate failure rates that could result from each standard. Because the standards are located in a dense part of the reference group ability distribution, small changes in a percent correct standard result in substantial changes in failure rates. This causes the failure rates to vary from judge to judge. But all of the standards set by judges were within 1.1 standard errors of the norm referenced standard.

For the data analyzed in this study, the norm referenced approach and the criterion item approach to locating a minimum passing standard are mutually validating. Expert judgments about test item content confirm the placement of the norm referenced standard. Individual differences in judges personal views about what level of knowledge should be required of candidates will always lead to some diversity in the individual standards set by judges. While a discussion among judges about the level of knowledge that ought to be required of candidates may diminish this variability, it is unlikely that it can ever be eliminated completely.

DISCUSSION

The procedure described in this paper may be used to establish, validate, or change the standard for a test. It has advantages over other widely used methods. These advantages and several potential misunderstandings of the procedure are discussed in this section.

EFFICIENT UTILIZATION OF CONTENT EXPERTS

The amount of time available from the content experts who serve as judges is always limited. Unlike the Angoff (1971), Ebel (1979), and Nedelsky (1954) procedures for setting standards, the criterion item procedure does not require judges to rate, or even to consider, every item in a test. For the orthopaedic surgery test, each judge would have needed to review 1200 (240 X 5) options to implement the Nedelsky procedure.

Procedures that require the considered judgment of a large number of items could produce a cursory review. When done carefully, this is a tedious and exhausting task. Considering just a small set of items allows judges to apply their talent and skill with much closer scrutiny of their items.

The criterion item approach does not require extensive training sessions for judges. Nor does it require all judges to travel to a central location for training before they begin their task. This study was conducted by mail.

ENCOURAGING DIVERSITY TO ACHIEVE CONSENSUS

Criterion item standard setting encourages the diversity and individuality of judges by eliminating group training, by permitting the selection of unique sets of standard setting items, and by allowing each judge to quantify his personal perception of how much knowledge is enough for passing. It emphasizes uniqueness rather than conformity, and in this respect is diametrically opposed to the procedures advocated for other approaches to standard setting. According to traditional thinking, this emphasis on diversity should produce divergent standards reflecting disagreement among judges. Instead, however, it produced remarkable unanimity.

This suggests that traditional thinking should be reconsidered. The criterion item approach permits each judge to define a personal, internally consistent frame of reference for the task, and then to work within this frame of reference without disruption from outside influences.

CAN A STANDARD BASED ON A SMALL SET OF ITEMS BE GENERALIZED TO THE ENTIRE TEST?

The orthopaedic surgery test is constructed to yield a useful degree of unidimensionality. All items share in common that they assess aspects of orthopaedic surgery that all candidates should know. Although every item assesses different content, each one is related to the practice of orthopaedic surgery. Therefore every possible subset of items should provide a reasonable representation of the entire test. The results of the analyses conducted suggest that this is a reasonable assumption.

STANDARDS BASED ON CONTENT ALL CANDIDATES SHOULD KNOW

In the present study, participants selected items testing content that they judged should be known by all passing candidates. A possible inference from this is that the performance standard for such important content should be 100% correct. But this inference overlooks something important.

When item content alone is considered there may be little disagreement among judges about whether it should be known. But when item structure is considered, there is ample room for doubt about whether competent and knowledgeable candidates will always choose the correct alternative. The intervening variable, item structure, refers to the clarity and precision of the text a candidate must read in the item stem, and to the length, complexity, readability, and similarity of the item's distractors.

An item-based standard setting procedure must take item structure into account. A judge's review of candidates' actual responses to a criterion item (P-values for each option) cultivates a better understanding of the impact of item structure upon candidate performance. Rasch item calibrations and fit statistics will also be helpful. When many candidates select a particular incorrect alternative, the item may contain an ambiguity or a subtlety not previously recognized. A popular incorrect response may reflect substantial partial knowledge about the content being tested.

It is difficult to imagine the instance where 100% correct would be a reasonable standard, even if there were no disagreement that the content of the items should be known by passing candidates. Item review with empirical data is necessary to arrive at a sound judgment about whether an item is useful for standard setting and to determine the level of performance that can reasonably be expected from candidates. Item review has the secondary benefit of increasing the sensitivity and skill of judges as item writers.

WHAT IF JUDGES PRODUCE WIDELY DISCREPANT STANDARDS?

What should one do if the judges set very different standards? The answer is to determine the source of the discrepancies. When judges select small sets of standard setting items (10 to 30), this is a manageable task. The results may be informative, and may lead to a resolution of the discrepancies.

A group of 20 board members can easily consider a set of 20 items, discuss them, and discuss the appropriateness of the standard that a judge has set. This has been done in practice. A testing authority different from the one discussed in this paper has conducted this exercise repeatedly. For this testing authority, the judgments of individual standard setters suggested consistently over a period of three years that the standard for passing the test should be higher. Sets of standard setting items were reviewed by all members of this testing authority. The levels of performance (percent correct standards) that were set by judges were considered. A consensus was reached that judges were expecting reasonable levels of knowledge from candidates. The standard for the test was then raised.

This kind of review procedure cannot be conducted with other approaches to standard setting. It is not possible, for example, when the Nedelsky method is used. Twenty board members can not review 1200 judgments (for 5 options on each of 240 items), much less discuss these judgments and arrive at a consensus.

The criterion item standard setting described here permits determination of the cause when a standard seems unreasonable. Criterion item standard setting is an improvement over other methods because the rationale for the standard can be made easily and is transparent to all who wish to inspect its source.

CONCLUSION

The application of criterion item standard setting has been enhanced by using the information provided by all of the standard setting items selected by each judge, rather than using only the most difficult standard setting item as reported for a previous application of similar methodology. The results of applying the method yielded results that supported the norm referenced standard for the test. This support was obtained in spite of the fact that the criterion item method differs markedly from other recommended procedures by encouraging the expression of diversity and individuality of judges. Rather than producing conflicting standards, this method resulted in a useful consensus.

Experience with the criterion item approach to standard setting has led to the conclusion that it has advantages over other methods. It requires significantly less time, effort, and cost. It fosters and capitalizes upon the uniqueness of judges by allowing them to select the content that will serve as the basis for their standards. The procedure is straightforward and uncomplicated for the judges. The approach permits the source of the standards to be inspected and evaluated by others who are not directly involved in the standard setting exercise.

REFERENCES

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

Ebel, R. L. (1979). Essentials of educational measurement. Englewood Cliffs, NJ: Prentice Hall.

Grosse, M. E. and Wright, B. D. (1986). Setting and maintaining certification standards with the Rasch model. Evaluation and the Health Professions, 9 (3), 267-285.

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

Rasch, G. (1960). Probabilistic models for some intelligence and attainment tests. Copenhagen: Danmarks Paedogogiske Institute. (Reprinted by the University of Chicago Press, 1980).

Wright, B. D. and Linacre, J. M. (1985). MICROSCALE: Rasch model measurement of change. Mediax Interactive Technologies, 21 Charles Street, Westport, CT 06880.

Wright, B. D. and Masters, G. N. (1981). CREDIT: Calibrating rating scales and partial credit scoring with the Rasch model. MESA Press, 5835 Kimbark Avenue, Chicago, IL 60637.

--- (1982). Rating scale analysis. MESA Press, 5835 Kimbark Avenue, Chicago, IL 60637.

Wright, B. D. and Mead, R. J. (1976). BICAL: Calibrating items with the Rasch model. Research Memorandum No. 23, Statistical Laboratory, Department of Education, University of Chicago.

Wright, B. D. and Panchapakesan, N. A. (1969). A procedure for sample-free item analyis. Educational and Psychological Measurement, 29, 23-48.

Wright, B. D., Rossner, M., and Congdon, R. (1985). MSCALE: calibrating rating scales with the Rasch model. MESA Press, 5835 Kimbark Avenue, Chicago, IL 60637.

Wright, B. D. and Stone, M. H. (1979). Best test design. MESA Press, 5835 Kimbark Avenue, Chicago, IL 60637.

Figure 1

Item Difficulties and Reference Group Abilities

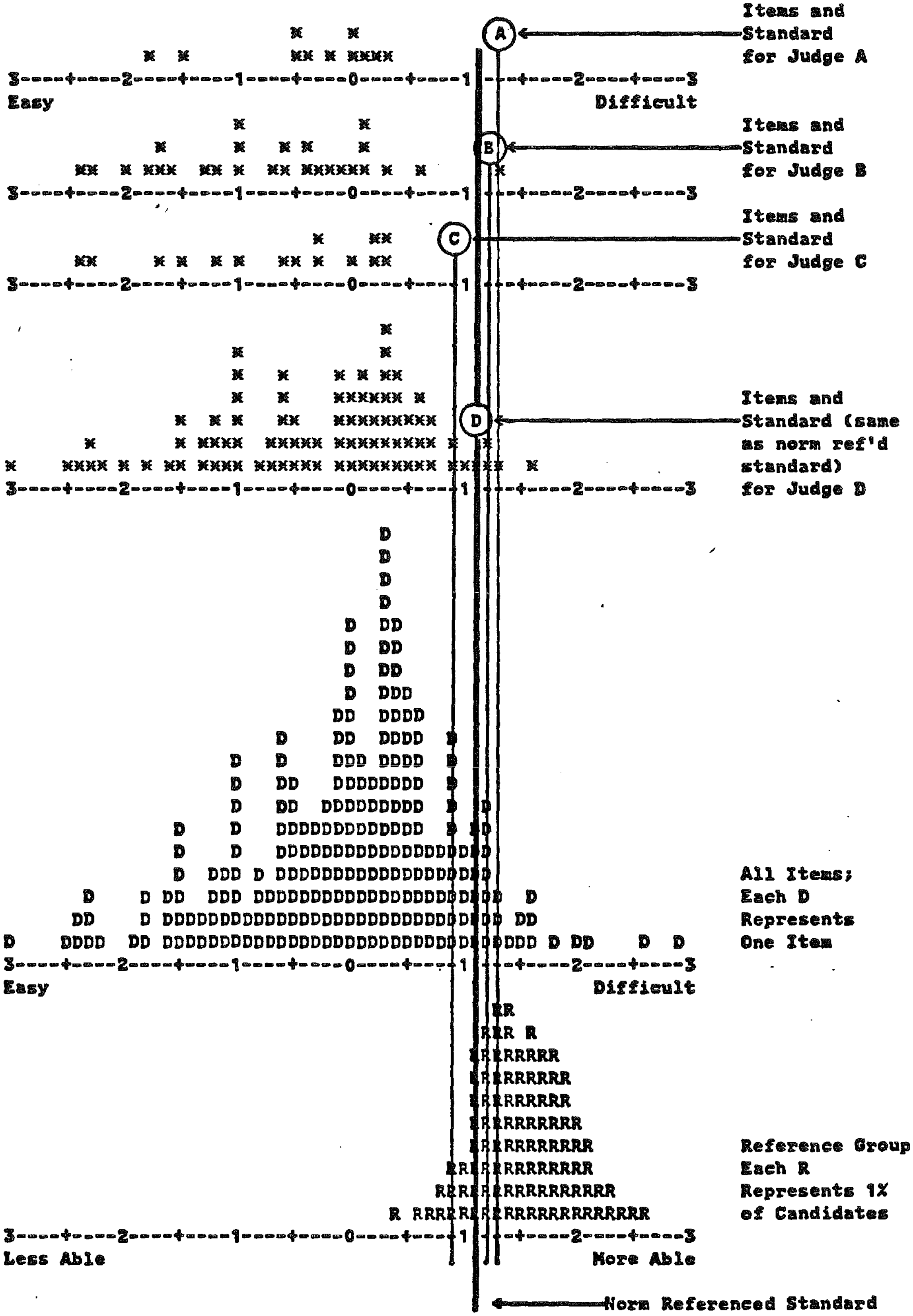

Table 1

Counts Of Criterion Items And Associated Standards

	Phase 1	Phase 2	Phase 3		
Judge	Number of Criterion Items Selected	Percent Correct Minimum Passing Standard	Number of Criterion Items Deleted	Number of Criterion Items Remaining	Percent Correct Minimum Passing Standard
A	15	90.0	3	11	82.5
B	30	90.0	4	27	85.0
C	16	81.0	1	15	81.0
D	95	77.5	7	88	77.5

Table 2

Minimum Passing Standards and Potential Failure Rates

	Minimum Passing Standards			
Judge	On the Criterion Item Sub-test (%)	On the Total Test Scale (Logits)	On the Total Test Scale (%) *	Reference Group Percentage Failure Rates (N = 476)
A	82.5	1.28	75	26
B	85	1.25	75	22
C	81	.93	69	5
D	77.5	1.12	72	13
Average (4 judges)	---	1.15	72	13
Norm Referenced Standard **	---	1.13	72	13

* Total Test standard obtained by using Rasch item difficulties to project the standard set by each judge on his individual set of criterion items onto the scale defined by the 240-item test.

** Set at one standard deviation below the reference group mean.

APPENDIX A

The relationship between the measures of item difficulty (D) and candidate ability (B) on a Rasch measurement scale and the probability (P) that an item of difficulty D will be answered correctly by a candidate with ability B is,

$$P = \frac{\exp(B-D)}{1 + \exp(B-D)} \tag{1}$$

where exp(B-D) represents the napierian coefficient e = 2.1718... raised to the power of the difference between B and D.

The table below gives the probability (P) of a correct answer and the difference (B - D) between ability and difficulty. The above formula can be used to compute other values.

Difference Between	
Person Ability and	Probability (P) of a
Item Difficulty	Correct Answer
in Logits	by Ability (B)
(B - D)	Against Difficulty (D)
-2.2	.10
-1.1	.25
.0	.50
+1.1	.75
+2.2	.90

This table shows that when item difficulty and candidate ability are equal, chances of a correct response by the candidate are 50/50. But when a candidate's ability is 2.2 logits higher than an item's difficulty, then chances are 90% that such an item will be answered correctly.