ESTIMATING CRITERION-REFERENCED STANDARDS FOR MULTIPLE-CHOICE EXAMINATIONS

Francis P. Hughes, Ph.D.
Charles F. Schumacher, Ph.D.
Benjamin D. Wright, Ph.D

February, 1984

ABSTRACT

Four methods were studied for setting a standard on a written examination containing several clinical discipline subtests. The NBME method yielded the most consistent estimate. The Angoff and Ebel methods yielded slightly less consistent estimates. All estimates were more consistent when computed as the average of discipline standards rather than the judges' personal standards. All but the Essential Content method yielded similar and practical estimates of the standard. The Ebel method was not found to be feasible for use with the examination studied; however, incorporating clusters of equally difficult, relevant items and feedback based on the judges' previous judgments into the NBME method yielded subsequent estimates of the standard that were more consistent than, but not significantly different from, the first estimate. The findings suggest the importance of enabling judges to select the items which will be most meaningful to them in expressing their judgments, the benefit of presenting clusters of equally difficult and relevant items and the utility of providing the judges feedback about their judgments and the opportunity to reconsider them. The study also showed that Rasch calibration and equating procedures provide a feasible methodology for expressing judgments based on different item samples on a common measurement scale.

ESTIMATING CRITERION-REFERENCED STANDARDS FOR MULTIPLE-CHOICE EXAMINATIONS

Francis P. Hughes, Ph.D.
Charles F. Schumacher, Ph.D.
Benjamin D. Wright, Ph.D.

INTRODUCTION

Measurements are used to make decisions about individuals. In education they are used to determine whether an individual has achieved instructional goals; in professional certification, they are used to determine whether an individual has acquired the knowledge judged necessary for awarding a certificate. In either case, setting an examination standard requires that judgments be made about minimally acceptable knowledge. Minimally acceptable knowledge is usually defined either as a relative position among some specified group of "standard" individuals (norm-referencing) or in terms of some specified amount of "basic" knowledge that an individual should have acquired (criterion-referencing).

Setting a standard by norm-referencing requires judgments that are different from those needed to set a criterion-referenced standard. Norm-referenced standards require the definition of a reference group and a decision as to what percentage of examinees in that group are likely to be minimally acceptable. Criterion-referenced standards require judgments about what minimally acceptable examinees actually know of the content on which they are being examined.

The National Board of Medical Examiners (NBME) has used both approaches in setting standards for certifying examinations. When NBME examinations were essay tests, criterion standards were set by examiners as they graded the essays. Individual examiners applied their personal standards for minimally acceptable knowledge and the group's standard was a consensus of these personal judgments. When the NBME introduced multiple-choice examinations norm-referenced standards were set by the test performance of reference groups to yield failure rates comparable to the rates observed in similar groups when individual examiners applied their personal standards to essay tests.

In 1981 (National Board of Medical Examiners, EXAMINER), the NBME introduced a "criterion group" standard-setting method which uses the predicted test performance of several previously-tested examinee groups to set the standard for each new examination. This method produces a standard for which there is no predetermined failure rate for any group taking the current test (Shimberg, 1981; Livingston and Zieky, 1982). It does not, however, consider test content and, in that sense, is norm- rather than criterion-referenced.

In developing a standard-setting procedure it is important to distinguish between estimating minimally acceptable knowledge and selecting an examination cutting score. To do the former, one needs to judge only how well minimally knowledgeable examinees ("MKEs") will perform on the examination. To do the latter, one must not only estimate minimally acceptable knowledge but must also consider the educational and societal impact of the cutting score, and the consequences of erroneous decisions for examinees and society (Millman, 1973).

This is a study of the process by which minimally acceptable knowledge is estimated by judges. The study does not address the more complex problem of

setting cutting scores for NBME examinations.

In order for estimates of minimally acceptable knowledge to be useful in making standard-setting decisions, judges must have a method for translating their judgments into test scores. Different methods for accomplishing this have been proposed. Comparative studies of these methods have shown that the standard selected can depend upon which method is used. This study was undertaken to determine how useful four of these methods would be for obtaining standard-setting judgments for NBME certifying examinations.

REVIEW OF NBME RESEARCH

Andrew and Hecht (1976) found that the method described by Nedelsky (1954) yielded a lower standard for a nationally administered certifying examination in the health professions than the method described by Ebel (1972). Guerin, Burg, and Vaughan (1978) found that the standards for two recertifying examinations obtained with a modified Nedelsky technique were similar to the norm-referenced standards set for those examinations. Guerin, Butzin and Schumacher (1982) investigated a procedure that yielded a criterion-referenced standard for a recertifying examination not too different from the standard that would have been obtained had the modified Nedelsky method been used. They also found that different groups of standard-setters could make similar judgments, a finding reported by Andrew and Hecht. Hughes (1981) described another method that used feedback to foster agreement among standard-setters.

These studies suggest that useful criterion-referenced standards can be set. They also show that the method for setting a criterion-referenced standard affects the standard. Setting a criterion-referenced standard

requires not only a decision as to who the standard-setters will be, but also a choice of method to translate judgments about minimally acceptable knowledge into a test score. This study reports the consistency with which a standard was estimated using each of four methods and the consequence of these standards on failure rates.

STANDARD-SETTING METHODS

Four methods for setting a criterion-referenced standard are investigated: Angoff, Ebel, Essential Content and NBME. Each of these methods uses standard-setters' judgments about test item content to define minimally acceptable knowledge. The Angoff (1971) and Ebel (1972) methods use only judgmental data to estimate the standard. The NBME (Hughes, 1981) and Essential Content (Guerin, et al., 1982) methods also use the difficulty of the test items observed among examinees.

Because NB examinations are very long, it is impractical to use standard-setters' judgments about all test items to estimate their standards (the NBME Part II test used in this study contained 862 items distributed among six clinical sciences). Since standard-setters were selected for their subject-matter expertise, their judgments are restricted to items in their own discipline. Combining judgments about subtests is facilitated by Rasch item calibrations which use examinee performance to estimate the difficulty of all items on a common scale (Rasch, 1960; Wright, 1968 and 1977; Wright and Stone, 1979). The NBME and Essential Content methods yield estimates of standard-setters' personal standards on the total test because the item difficulties are calibrated to the total test scale. Since the Angoff and Ebel methods are limited to estimates of standards on a discipline subtest, Rasch calibrations

are used to equate these subtest scores to a score on the total test.

The Angoff method uses each standard-setter's judgment of MKE success rate with every item in his clinical science subtest to estimate his personal standard. The success rate is interpreted as the probability that the MKE will answer the item correctly. The sum of these probabilities over the subtest items is the MKE subtest score. The equivalent score on the total test is the estimate of the standard-setter's personal standard.

The Ebel procedure uses judgments about MKE success rate with hypothetical relevance and difficulty item-types to estimate each standard-setter's personal standard. Four categories of item relevance (Essential, Important, Acceptable and Questionable) and four categories of item difficulty (Easy, On The Easy Side, On The Hard Side and Hard) are specified. First, the standard-setter judges the MKE success rate for each of these 16 hypothetical item-types. Then he classifies every item in the clinical science subtest into one of the 16 item-types according to his perception of its relevance and difficulty. The success rate for any selection of item categories is used as the MKE probability of success with test items classified in those categories. The MKE score in these categories is estimated by summing these probabilities over all items. The score on the total test equivalent to this estimate of MKE knowledge on the subtest is the standard-setter's personal standard.

The NBME method combines each standard-setter's judgments about the MKE success rate with the calibrated difficulty of the items to estimate the judge's personal standard. The Rasch model expects that the probability of a correct response to a test item (P) is dominated by the examinee's knowledge (b) and the item's difficulty (d) on a linear scale which defines the latent variable underlying the examination. The difference between examinee

knowledge and item difficulty is modelled to be equal to the log odds (logit) of a correct response, i.e.,

$$(b-d) = \log(P/(1-P))$$

Each standard-setter's judgments about MKE item-success rates are used to calculate these item-success logits. The line obtained by regressing the observed difficulty (d) of the items on these MKE success logits intercepts the difficulty axis at the point where b = d. This intercept estimates the MKE knowledge level (b) on the latent scale. The total test score equivalent to this measurement (b) on the latent scale is the standard-setter's personal standard in test score units.

The Essential Content method uses the observed difficulty of the test items and each standard-setter's judgments about the relevance of item content to define minimally acceptable knowledge. Each standard-setter rates the relevance of every item in his clinical discipline as Essential, Important, Acceptable, or Questionable. The personal standard is set at the point on the latent scale corresponding to the difficulty (d) of the most difficult Essential items. The total test score equivalent to that point on the latent scale is the standard-setter's personal standard, and the average of these personal standards is the group's examination standard.

The rationale for this approach is that "essential content" must be mastered by examinees for their knowledge to be judged acceptable. Setting the standard at the most difficult of the Essential items specifies that the MKE is expected to have a 0.50 probability of answering that item correctly and that his success rate with other Essential items will be greater than 0.50. This is the principle underlying the procedure reported by Guerin, et al., (1982); their implementation, however, differs from the implementation in this study. Guerin, et al., used the average of their judges' relevance

ratings to assign a single relevance rating to each item. Their "Essential" items were identified as those with average relevance more than one standard deviation above the group mean. They also made the assumption that highly relevant subject matter must be well-known so that items assessing such content must be easy. To obtain this they deleted all "Essential" items which were nevertheless difficult, citing ambiguities in content as the source of these difficulties. The result was a small set of "especially essential" items chosen because they were also easy. In the Essential Content method used here difficult essential items are not deleted.

DESIGN OF THE STUDY

The NBME examination used for this study is a Part II examination containing 862 multiple-choice items, distributed in equal numbers to Internal Medicine, Surgery, Obstetrics/Gynecology, Preventive Medicine/Public Health, Pediatrics, and Psychiatry subtests. Although the NBME scores candidates in each clinical science, it uses a total test score to determine whether an examinee's knowledge is acceptable. In this study, therefore, a criterion-referenced standard is estimated for the total examination by each of the four methods investigated.

A standard-setting panel of twelve medical educators with previous experience on NBME Part II Test Committees is used. These standard-setters were chosen for their expertise in a clinical science assessed by the examination and for their experience in writing items for multi-discipline examinations. There are two standard-setters for each clinical science.

The twelve standard-setters met for a two-day orientation. Before this meeting they were sent an overview of the study, information about the various

judgments they would be asked to make, and a sample of items in their clinical science. They were also asked to think about two hypothetical groups of examinees: MKEs, "minimally knowledgable examinees", and TUSMGs, "typical U.S. medical school graduates".

MKEs were defined as

> "... individuals who have just been awarded the MD degree by a US medical school and whose level of medical knowledge is the minimum acceptable for safe and effective medical practice, under supervision, at the beginning of residency training."

TUSMGs were defined as

> "typical U.S. medical school graduates who had attained a level of medical knowledge beyond the minimum acceptable for safe and effective medical practice under supervision."

Prior to the orientation meeting, standard-setters reviewed the sample of items in their clinical science and made the judgments needed to estimate MKE and TUSMG achievement for each standard-setting method. During the orientation meeting the standard-setters discussed the concepts of "minimally acceptable knowledge" and "minimally knowledgeable examinees" and asked questions about the judgements they were to make. In the course of this discussion the vagueness of the TUSMG definition became apparent, and the effort to clarify it by associating TUSMG's and typical examinees in the NB reference group was not altogether successful.

Following the meeting standard-setters reviewed the items in their clinical science and (1) specified the success rate they expected for MKEs and TUSMGs on each of the Ebel hypothetical relevance-by-difficulty item-types; (2) characterized each item according to their perception of its relevance (Essential, Important, Acceptable, Questionable) and difficulty (Easy, On The Easy Side, On The Hard Side, Hard); and (3) estimated the success rate for MKEs and TUSMGs on each item. The judgments appropriate to each of the four standard-setting methods are used to obtain each standard-setter's estimate of

MKE and TUSMG knowledge of the NBME Part II examination according to each method.

The average of the standard-setters' personal standards is the estimate of the group's examination standard. The consistency of the group's estimate is measured by a standard error, calculated by dividing the standard deviation of the personal standards by the square root of 12, the number of standard-setters. An estimate of the group's examination standard is obtained for each of the four methods. Because the same judges and, whenever possible, the same judgments are used with each procedure, differences among standards can be attributed to differences among the standard-setting methods.

PRESENTATION OF THE DATA

Estimates of standard-setters' personal standards and the group's examination standard are reported as percent scores on the examination. Table 1 shows the consistency and consequences of the examination standards obtained by each of the four methods. Estimates of the TUSMG average score are shown in Table 2. Table 3 reviews the accuracy of the judgments on which the estimates of the TUSMG average score are based. Estimates of MKE performance based on variations of the NBME method are shown in Tables 4-6.

Consistency of the Examination Standards

Consistency of examination standards is measured by the standard error of the average of the standard-setters' personal standards. The NBME method yields the most consistent estimate (Table 1) with a standard error of 1.8 percent score units. The Essential Content method yields the least consistent

estimate with a standard error of 3.3 score units. The Angoff and Ebel methods yield estimates with standard errors of 2.5 units.

The consistency of these estimates is improved by computing the examination standard as the average of the standards estimated for each clinical science (Table 1). A clinical science standard is the average of the personal standards of the two judges in that same science. When the examination standard is computed this way, the standard errors are 1.2 for the NBME method and 1.8, 2.0, and 3.2 for the Angoff, Ebel, and Essential Content methods.

Consequences of the Estimated Examination Standards

The Norm reference method used by NBME in 1980 when this Part II examination was administered produced a failure rate of 2.4%. The NBME reference group contained only examinees in their final year in a US medical school taking the Part II examination for the first time. Table 1 gives the failure rates associated with each of the four methods. The Essential Content method produces a reference group failure rate of 85.8%. The Angoff and Ebel methods produce reference group failure rates of 8.3% and 6.0%. The NBME method, at 3.5%, produces the lowest failure rate of the four methods.

It is not possible to assess the accuracy of standard-setters' individual judgments about MKE success rates because item p-values are not available for examinees whose achievement is "minimally knowledgeable". It is possible, however, to compare estimates of TUSMG score with the 65.4% average score of the NB reference group. The NBME, Ebel, and Angoff methods are used to estimate average TUSMG scores (Table 2). TUSMG achievement estimates are 60.1% with the NBME method, 65.1% with the Ebel method, and 69.3% with the Angoff method.

Considering the difficulty judges had in arriving at a clear definition of the competence of a typical U.S. medical graduate, the closeness of the average TUSMG scores to the NBME reference group average is encouraging. The details of the judges' TUSMG ratings in Table 3, however, bring out the trouble that even experts have when they try to predict the difficulty of multiple-choice items. While average scores for the set of items are within 5 percentage points of the reference group average, no judge gets more than 48% of his items within 10 percentage points of the reference group item p-values and one judge is more than 15 percentage points off on 51% of his items.

Feasibility of Implementing these Methods

The majority of standard-setters said it was easy to classify items according to the relevance and difficulty of their content, as required by the Ebel method. However, they said it was difficult to judge success rates with hypothetical items characterized by relevance and difficulty. Most were not sure how changes in relevance or the interaction between relevance and difficulty should affect their judgments about MKE success rates.

The majority of standard-setters said it was difficult to judge MKE success rates, whether with actual items as required by the Angoff and NBME methods or with hypothetical items. They thought it easier with actual items, however, because they were tangible and could be examined for content and format.

Standard-setters asked repeatedly for more information about the items. They wanted reference points to keep "...in touch with reality". NBME reference group p-values for the items were available but were not given to the standard-setters for fear this would bias their judgments about MKE success. A variation of the NBME procedure using reference group p-values to

select items (Table 4) is reported.

Estimating Standards Using Items Selected for TUSMG Accuracy

Standard-setters expressed judgments about items in their own clinical science discipline, concentrating on the subject matter they knew best. This should enhance the validity of their personal standards because the effect of extraneous factors on their judgments should be reduced.

To improve on this, the MKE knowledge level is estimated a second time using the NBME method and the initial MKE success rates the standard-setters specified, but only for items standard-setters characterized as relevant (Essential or Important) and for which their judgments about TUSMG success differed by less than 16 percentage points from the item's NBME reference group p-value (Table 4).

Since the items selected are those for which standard-setters' judgments about TUSMG success rate are near NBME reference group p-values, the new TUSMG estimate (64.2%) is closer to the reference group average (65.4%) than the estimate based on all items (60.1%). There is also a reduction in the standard error from 1.4 to 0.9.

The selected items produce a new estimate of the group's examination standard (51.0%) which is slightly lower than the original estimate (52.4%) and closer to the normative standard used with this Part II examination (50.5%). The standard error of the new estimate (1.8), however, is the same as the standard error of the original estimate.

Estimating the Standard Using Items Clustered by Difficulty and Feedback

A second item sampling strategy uses samples of items which

standard-setters characterized as Essential or Important placed in clusters according to their calibrated difficulty. Each cluster contains up to five items of comparable difficulty. The difficulties of adjacent clusters differ by at least 0.40 logits on the calibration scale. These item clusters are like Ebel's hypothetical item types without the distinction between Essential and Important and excluding Acceptable and Questionable items.

The personal standards reported in Table 5 at Time 1 are those reported in Table 1. They are obtained using judgments of MKE success with individual items. At Times 2 and 3 personal standards are obtained using judgments of MKE success with item clusters presented in ascending difficulty order. The regression method is used to estimate the standard-setters' personal standards.

At Times 2 and 3, standard-setters were asked to reconsider their previous judgments of MKE success. To help them refine their preceding estimates and to foster consensus, they were given two estimates of MKE success with each of their item clusters. One is based on the preceding estimate of their personal standard; the other on the preceding estimate of the group's standard. The implied success rate (P) for a cluster is calculated from:

$$P = \exp(b-d)/(1 + \exp(b-d))$$

where the preceding estimate of the standard is substituted for (b) and the difficulty of the cluster is substituted for (d).

The items in each cluster are comparable in difficulty; therefore, an examinee should have the same probability of answering them correctly. By examining a cluster of items when judging MKE success rate, standard-setters can focus on the common features of item content and format. Before

expressing their judgments about MKE success rate with each cluster, standard-setters studied the items in the cluster and deleted those they thought differed in difficulty from the others. If standard-setters' estimates of MKE knowledge level are stable, the success rates they specify will vary inversely with item cluster difficulty. With a few exceptions, that occurs.

The group's standard at Time 1 using judgments about individual items is 52.4%. The standard at Time 2 is 53.9%, and remains there at Time 3 (Table 5). The consistency of the estimates increases each time as evidenced by successively smaller standard errors of 1.8, 1.2, and 1.0.

This decrease in variability among standard-setters personal standards with the use of item clusters and feedback can be seen in Figure 1. Judges with extreme views about minimally acceptable knowledge tend to moderate those views but continue to maintain personal standards that are either more lenient or more stringent than those of the other standard-setters (B,MED2, G,PMPH1, H,PMPH2, and J,PEDS2 in Figure 1 and Table 5). The opinions of the other standard-setters change slightly at subsequent cycles, but these changes appear to be adjustments resulting from a more refined expression of their views rather than a modification.

The standard errors associated with the estimates of the standard-setters' personal standards either remain the same or increase from Time 1 to Time 2, but without exception they decrease at Time 3 (Figure 1). With only one exception, standard-setters' judgments at the third cycle are more consistent than their judgments at the first cycle. We attribute this increased consistency to their expressing judgments about a few clusters containing relevant items of comparable difficulty and to the feedback.

Independent Estimates Method

Instead of the regression approach described earlier, the difficulty of each item (d) and the standard-setter's judgment regarding MKE success (P) on that item can be used to obtain an independent estimate of MKE achievement (b) from:

$$b = \log (P/(1-P)) + d$$

These independent estimates obtained for separate items are averaged to estimate the standard-setter's personal standard. The standard error of the personal standard is the standard deviation of these independent estimates divided by the square root of their number. The group's standard and its standard error are calculated as before.

Estimates of personal standards and of the group's standard were obtained in this way at each cycle of the study. These estimates and their standard errors are reported in Table 6. The estimates are similar to the estimates obtained with the regression method. The group standard at Time 1 is not quite as consistent as the estimate obtained with the regression method but is about as consistent as the estimates obtained with the Angoff and Ebel methods. At Times 2 and 3, the consistency of the independent method estimate approaches the consistency of the regression method estimate.

DISCUSSION

The four standard-setting methods studied yield estimates of criterion-referenced standards that differ in consistency. Regardless of procedure, estimates are more consistent when computed as the average of clinical discipline standards.

The NBME regression estimate of the group standard is the most

consistent. It is least sensitive to aberrant judgments about individual items. The Angoff method gives equal weight to every judgment of the MKE success rate; the NBME procedure does not. By fitting a regression line through the mean difficulty of items judged to have the same MKE success rate, the NBME method minimizes the impact of aberrant judgments in the estimation of the standard.

Only the Essential Content method yields an examination standard that is impractical (85.5% failure rate in the NB reference group). This result is due to the way this method was implemented in the current study and should not be interpreted as indicative of results that might be obtained under the similar method described by Geurin, et al.

The data concerning the feasibility of implementing these standard-setting methods is mixed. Standard-setters found it easier to judge the relevance of individual test items than MKE success rate with those items. They were also less confident about judging MKE success rate with hypothetical groups of items than with actual test items. However, the Ebel method which requires judgments of item relevance and difficulty and of success rates with groups of hypothetical items yields the most accurate estimate of the average score for the NBME reference group. This occurs even though standard-setters said they were uncertain about the impact of changes in item relevance and difficulty on their judgments about success rates.

The standard-setting data and the standard-setters' opinions negate the feasibility of implementing the Ebel method with the NBME examination programs. The use of clusters of equally difficult, relevant items rather than individual items was incorporated into the NBME method, however, and used at the second and third cycles of this study. Relevance is based on standard-setters' judgments but item difficulty is based on examinee

performance. The intent of varying only difficulty and not relevance is to reduce the standard-setters' uncertainty when judging MKE success rates with item clusters. Judges were also guided in their judgments by feedback based on their previous judgments and by knowledge that the clusters were presented in difficulty order.

The use of item clusters and feedback do not yield standards significantly different from the first estimate obtained. Estimates do become more consistent with each cycle, indicating a movement toward consensus among the standard-setters. Increased consistency of estimates is present whether the regression or independent estimate method is used.

The regression and independent estimates methods yield statistically equivalent estimates of the examination standard at cycles 1, 2 and 3. The first estimate obtained with the regression method is more consistent than the first one obtained with the independent estimates method. By cycle 3, however, estimates are equally consistent.

It was not our objective to determine which criterion-referenced standard-setting procedures yielded estimates closest to the National Board norm-referenced standard, but the similarity between the normative standard and three of the criterion-based standards invites comment. The standard-setters used in this study were not involved in the process by which the National Board determined the standard for this test. The fact that the normative standard is close to the three estimates of criterion-based standards suggests that it may be possible to reconcile criterion and normative based standards in a way which facilitates their joint use.

All but one of the methods used to obtain judgments from standard-setters about the expected performance of MKEs yield similar results. The results suggest that it is important for judges to be able to select the items which

will be most meaningful to them in expressing their judgments, that it is important to provide feedback to the judges which permits them to compare their judgments with those of their peers and modify their judgments on the basis of those comparisons and that it may be helpful to provide judges with clusters of items rather than indiviudal items on which to make their judgements.

If it is important to allow each judge to select a different set of items on which to make his judgments, then a methodology must be available by which judgments based on different samples of items can be expressed on a common scale so that the standard resulting from those judgments can be translated into performance on examinations. Without such a methodology none of the standard-setting methods studied here are practical for NBME because the number, content and length of examinations used by NBME make it impossible to perform the judgmental tasks required under any of the methods, for all items. This study shows that the Rasch calibration and equating procedures developed by NBME can provide such a methodology and, the process of establishing and monitoring a "criterion" based standard for an ongoing certifying examination is feasible if these procedures are utilized.

TABLE 1

PERSONAL STANDARDS, EXAMINATION STANDARDS, AND FAILURE RATES

Standard-Setters	Standard Setting Methods: NBME	Ebel	Angoff	Essential Content
A (MED1)	50.6	48.6	50.4	57.7
B (MED2)	61.0	54.9	66.1	84.1
C (SURG1)	52.6	59.3	60.6	78.0
D (SURG2)	51.7	56.6	56.7	83.1
E (OBGYN1)	58.1	66.1	65.3	85.6
F (OBGYN2)	53.2	57.9	53.5	85.5
G (PMPH1)	41.1	36.5	41.5	69.3
H (PMPH2)	65.0	70.1	69.3	56.1
I (PEDS1)	50.8	44.8	51.8	83.4
J (PEDS2)	43.3	47.2	40.4	63.6
K (PSYCH1)	50.1	52.6	52.6	76.1
L (PSYCH2)	50.8	56.0	56.6	55.3
GROUP BY INDIVIDUAL	NORM			
Standard	50.5	52.4	54.2	55.4	73.2
Std. Error		1.8	2.5	2.5	3.3
GROUP BY DISCIPLINE					
Standard		52.4	54.2	55.4	73.2
Std. Error		1.2	2.0	1.8	3.2
REFERENCE GROUP	NORM				
Failure	2.4% (n=113)	3.5% (n=169)	6.0% (n=289)	8.3% (n=400)	85.8% (n=4113)

TABLE 2

ESTIMATES OF NATIONAL BOARD REFERENCE GROUP
AVERAGE SCORES

Standard-Setters	------Estimation Methods------ NBME	Ebel	Angoff
A (MED1)	60.6	59.9	65.5
B (MED2)	65.7	72.3	76.2
C (SURG1)	54.0	70.5	65.5
D (SURG2)	61.0	75.1	78.4
E (OBGYN1)	60.4	68.2	71.7
F (OBGYN2)	60.8	70.1	72.3
G (PMPH1)	57.6	54.2	59.7
H (PMPH2)	71.5	(No Data)	77.4
I (PEDS1)	57.7	57.1	68.2
J (PEDS2)	54.2	58.3	58.9
K (PSYCH1)	62.5	69.4	70.0
L (PSYCH2)	54.9	61.5	67.3

	REF. GRP. AVERAGE			
GROUP Average	65.4	60.1	65.1	69.3
(%tile Rank)	(49th)	(21st)	(48th)	(70th)
Std. Error		1.4	2.0	1.8

TABLE 3

DIFFERENCES BETWEEN TUSMG JUDGMENTS AND REFERENCE GROUP P-VALUES

	---------Magnitude of Differences--------						
	Less than or Equal to 10		Between 11 and 15		Greater than 15		Mean Differences*
Standard-Setters	Items	%	Items	%	Items	%	TUSMG - P-Value
A (MED1)	56	39%	25	18%	62	43%	00
B (MED2)	53	37%	22	15%	69	48%	11
C (SURG1)	57	40%	20	14%	65	46%	00
D (SURG2)	57	40%	18	13%	67	47%	13
E (OBGYN1)	54	39%	21	15%	63	46%	06
F (OBGYN2)	55	41%	19	14%	61	45%	07
G (PMPH1)	52	35%	27	18%	69	47%	-06
H (PMPH2)	71	48%	22	15%	55	37%	11
I (PEDS1)	61	40%	19	14%	69	46%	03
J (PEDS2)	57	39%	16	10%	76	51%	-07
K (PSYCH1)	63	45%	26	19%	51	36%	04
L (PSYCH2)	60	43%	17	12%	63	45%	02

*A negative difference indicates that the standard-setter underestimated the TUSMG's success rate and a positive difference indicates that the standard-setter overestimated the TUSMG's success rate.

TABLE 4

REFERENCE GROUP ESTIMATES AND PERSONAL STANDARDS USING A SELECTED SAMPLE* OF ITEMS

Standard-Setters	Reference Group Estimates		Personal Standards	
	All Items	Sample	All Items	Sample
A (MED1)	60.6	63.6	50.6	48.4
B (MED2)	65.7	69.8	61.0	60.0
C (SURG1)	54.0	62.9	52.6	57.5
D (SURG2)	61.0	63.9	51.7	42.9
E (OBGYN1)	60.4	61.4	58.1	55.7
F (OBGYN2)	60.8	60.0	53.2	47.2
G (PMPH1)	57.6	65.5	41.1	46.0
H (PMPH2)	71.5	70.6	65.0	62.4
I (PEDS1)	57.7	60.8	50.8	47.0
J (PEDS2)	54.2	64.0	43.3	43.7
K (PSYCH1)	62.5	65.5	50.1	47.2
L (PSYCH2)	54.9	62.9	50.8	53.8
GROUP				
Standard	60.1	64.2	52.4	51.0
(%tile Rank)	(21st)	(42nd)	--	--
Std. Error	1.4	0.9	1.8	1.8

Ref. Grp. = 65.4 Norm. Std. = 50.5

*Each sample contained only those items whose relevance the standard-setter rated as Essential or Important and for which the judged success rate in the reference group differed by no more than 15% from the reference group p-value.

TABLE 5

PERSONAL STANDARDS AND EXAMINATION STANDARDS USING THE REGRESSION METHOD

Standard-Setters	All Items Time 1* Standard	SE	Clusters of Relevant Items Time 2# Standard	SE	Time 3# Standard	SE
A (MED1)	50.6	1.6	49.2	1.8	50.6	1.0
B (MED2)	61.0	1.7	57.9	1.7	57.3	1.0
C (SURG1)	52.6	1.6	58.1	1.7	56.6	0.3
D (SURG2)	51.7	1.6	53.0	1.6	51.6	0.6
E (OBGYN1)	58.1	1.8	58.3	0.8	57.9	0.4
F (OBGYN2)	53.2	1.7	51.4	2.1	50.2	0.4
G (PMPH1)	41.1	1.7	48.8	2.2	49.4	0.5
H (PMPH2)	65.0	1.8	60.0	3.6	60.0	1.7
I (PEDS1)	50.8	1.9	56.0	1.5	56.0	0.6
J (PEDS2)	43.3	1.8	47.2	2.1	48.8	1.6
K (PSYCH1)	50.1	1.8	54.0	4.4	55.7	3.1
L (PSYCH2)	50.8	2.0	52.4	0.6	53.0	0.6
GROUP						
Standard	52.4		53.9		53.9	
Std. Error	1.8		1.2		1.0	

Norm Std. = 50.5

*Judgments of MKE success were made for individual items, regardless of their relevance. No feedback was provided.

#Judgments of MKE success were made for clusters containing only relevant items. Feedback based on previous judgements was provided.

TABLE 6

PERSONAL STANDARDS AND EXAMINATION STANDARDS USING THE INDEPENDENT ESTIMATES METHOD

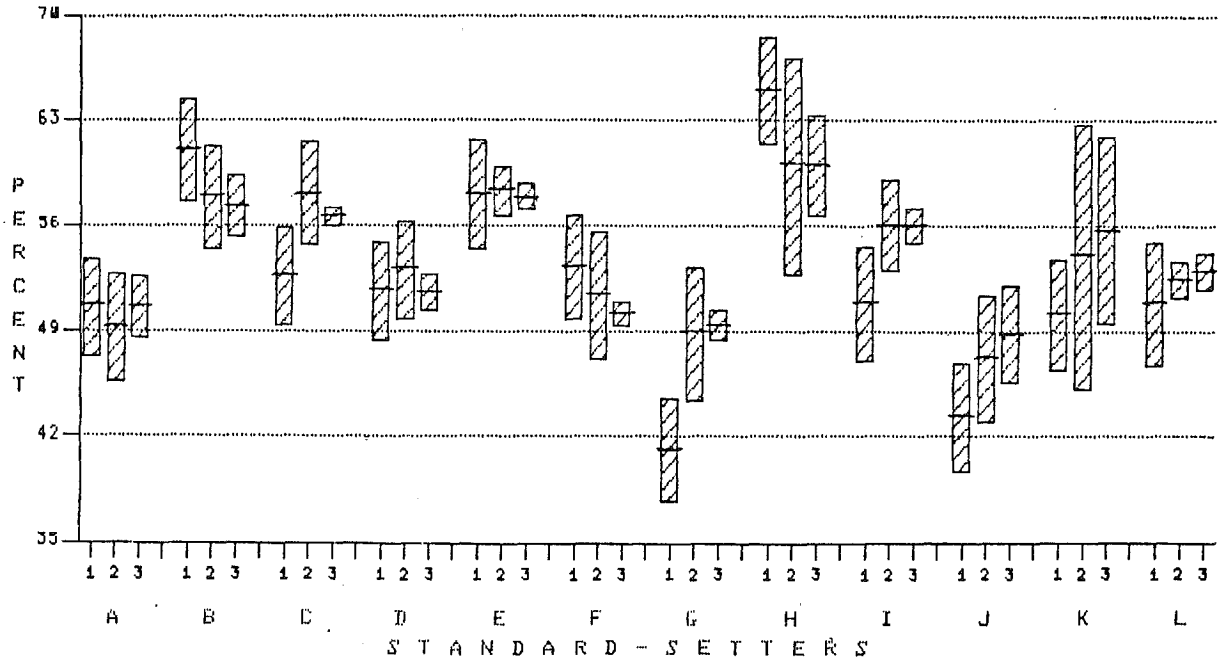
Standard-Setters	All Items Time 1* Standard	SE	Clusters of Relevant Items Time 2# Standard	SE	Time 3# Standard	SE
A (MED1)	50.4	2.0	46.5	3.1	48.8	1.7
B (MED2)	63.3	1.7	54.0	2.1	54.6	2.3
C (SURG1)	60.3	1.8	56.1	2.1	55.2	1.2
D (SURG2)	55.7	1.7	47.6	2.6	51.2	0.6
E (OBGYN1)	68.9	1.7	58.7	0.8	58.3	0.4
F (OBGYN2)	54.0	1.7	55.3	5.3	50.6	1.0
G (PMPH1)	41.1	1.7	46.0	2.6	49.2	0.4
H (PMPH2)	65.1	1.9	55.7	4.5	58.7	1.7
I (PEDS1)	51.7	2.0	58.9	2.1	55.7	0.6
J (PEDS2)	41.4	1.7	47.2	2.6	48.4	1.7
K (PSYCH1)	50.9	1.8	49.0	5.0	50.2	4.2
L (PSYCH2)	54.9	2.0	51.0	1.2	51.8	1.0
GROUP						
Standard	54.8		52.2		52.7	
Std. Error	2.4		1.3		1.0	

Norm Std. = 50.5

*Judgments of MKE success were made for individual items, regardless of their relevance. No feedback was provided.

#Judgments of MKE success were made for clusters containing only relevant items. Feedback based on previous judgments was provided.

FIGURE 1
Changes in Personal Standards Using the Regression Method
(Data from Table 5)



Top = Mean +2 Std. Errors
Bottom = Mean -2 Std. Errors
Middle Line = Mean

REFERENCES

Andrew, B.J. & Hecht, J.T. A preliminary investigation of two procedures for setting examination standards. Educational and Psychological Measurement, 1976, 36, 45-50.

Angoff, W.H. Scales, norms and equating. In R.L. Thorndike (ed.), Educational measurement (2nd ed.). Washington, DC; American Council on Education, 1971.

Ebel, R.L. Essentials of educational measurement. Englewood Cliffs, NJ; Prentice-Hall, 1972.

EXAMINER, National Board of Medical Examiners, Winter 1981.

Guerin, R.O., Burg, F.D. & Vaughan, V.C. Paper presented at ABMS Conference on Research in Evaluation Procedures. March, 1978.

Guerin, R.O., Butzin, D. & Schumacher, C.F. Paper presented at the annual meeting of the American Educational Research Association. New York, 1982.

Hughes, F.P. A procedure for estimating a criterion-referenced standard. Paper presented as part of a symposium on standard setting at the annual meeting of the Northeastern Educational Research Association. Ellenville, NY, 1981.

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

Millman, J. Passing scores and test lengths for domain-referenced measures. Revision of Education Research, 1973, 43, 205-216.

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

Rasch, G. Probabilistic Models for Some Intelligence and Attainment Tests. Copenhagen, Danmarks Paedogogiske Institut, 1960 (Chicago, University of Chicago Press, 1980).

Shimberg, B. Testing for licensure and certification. American Psychologist, Vol. 36, No. 10, pp 1138-1146, October 1981.

Wright, B.D. Sample Free Test Calibrations and Person Measurement. Proceedings of the 1967 Invitational Conference on Testing Problems. Princeton, NJ, ETS, 1968.

Wright, B.D. Solving Measurement Problems with the Rasch Model. Journal of Educational Measurement, 1977, 14, 97-114.

Wright, B.D. & Stone, M.H. Best Test Design. Chicago, MESA Press, 1979.