DOCUMENT RESUME

ED 230 617 TM 830 457

AUTHOR Hughes, Francis P.
TITLE Comparing Four Estimates of the Criterion-Referenced Standard for a Written Test.
PUB DATE Apr 83
NOTE 32p.; Paper presented at the Annual Meeting of the American Educational Research Association (67th, Montreal, Quebec, April 11-15, 1983); Some figures may be marginally legible due to small print.
PUB TYPE Speeches/Conference Papers (150) -- Reports - Evaluative/Feasibility (142)

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS *Criterion Referenced Tests; Cutting Scores; Higher Education; *Measurement Techniques; Minimum Competencies; Multiple Choice Tests; *Psychometrics; *Standards
IDENTIFIERS Angoff Methods; Ebel Method; Guerin Method; *National Board of Medical Examiners; *Standard Setting; Standards for Educational and Psychological Tests

ABSTRACT

Four procedures were used to estimate a criterion-referenced standard for a multiple-choice examination developed by the National Board of Medical Examiners (NBME). Two experimental procedures, the NBME method and a modification of the Guerin method, and the Angoff and Ebel procedures were evaluated on the consistency of the estimates they yielded, the plausibility of the failure rates, and the standard-setters' confidence in their judgments. The NBME and modified Guerin procedures yielded the most consistent and least consistent estimates, respectively. The failure rates associated with the standards obtained using these procedures were higher than the failure rate associated with the test's norm-referenced standard, but only the failure rate associated with the modified Guerin procedure was obviously unacceptable. The standard-setters said it was difficult to judge the success rate of "minimally knowledgeable examinees" with the test questions, but even more difficult to make those judgments for the hypothetical classifications of items used with the Ebel procedure. The estimates obtained using three of the procedures were relatively consistent and the failure rates associated with them, although higher than the rate experienced with a norm referenced standard, were plausible. (Author/PN)

COMPARING FOUR ESTIMATES OF THE CRITERION-REFERENCED STANDARD FOR A WRITTEN TEST

BY

FRANCIS P. HUGHES, PH.D.

THE AMERICAN BOARD OF ANESTHESIOLOGY, INC.

PRESENTED AT THE 1983 AERA ANNUAL MEETING (SESSION 26.11)

MONTREAL, QUEBEC

ABSTRACT

The Angoff and Ebel procedures were used to estimate a criterion-referenced standard on a written test. Two experimental procedures, the NBME method and a modification of the Guerin method, also were used to estimate a standard for the test. The four procedures were compared in terms of the consistency of the estimates they yielded, the plausibility of the failure rates, and the standard-setters' confidence in their judgments.

The NBME and modified Guerin procedures yielded the most consistent and least consistent estimates, respectively. The failure rates associated with the standards obtained using these procedures were higher than the failure rate associated with the test's norm-referenced standard, but only the failure rate associated with the modified Guerin procedure was obviously unacceptable. The standard-setters said it was difficult to judge the success rate of "minimally knowledgeable examinees" with the test questions, but even more difficult to make those judgments for the hypothetical classifications of items used with the Ebel procedure.

The findings were encouraging in several respects. The estimates obtained using three of the procedures were relatively consistent and the failure rates associated with them, although higher than the rate experienced with a norm referenced standard, were plausible. Only the modified Guerin technique yielded an inconsistent estimate with an obviously unacceptable failure rate, and those findings may say more about the modifications made to Guerin's procedure than about the unaltered procedure.

Comparing Four Estimates of the Criterion-Referenced Standard for a Written Test

by
Francis P. Hughes, Ph.D.1

INTRODUCTION

Measurements are used to make decisions about individuals, and in an educational setting they often are used to determine whether an individual has achieved desired instructional goals. If such decisions are to be valid, the standard that is used must be appropriate and acceptable as an indicator of "minimally acceptable achievement."

Prior to the date when the National Board replaced its written essay examinations with lengthy multiple choice tests, an examination standard was set collectively by the examiners as they graded the essay responses. The examiners applied their personal standards for "minimally acceptable achievement", and the group's standard represented a consensus of their personal judgments. With the introduction of its multiple choice examinations, the National Board formally adopted a norm-referenced standard that resulted in a failure rate similar to the one experienced when individual examiners applied their criterion-referenced standards to the grading of essay responses.

Setting a norm-referenced standard at a specified level in the distribution of test scores for a well-defined group of examinees requires judgments that are different from those needed to set a criterion-referenced standard. Norm-referenced standards require judgments about the definition of an appropriate reference group and the percentage of examinees in that group whose achievement is not likely to be "minimally acceptable." The emphasis is on the reference group and only indirectly on the knowledge that represents "minimally acceptable achievement."

Criterion-referenced standards, on the other hand, are based on judgments about what examinees whose achievement is "minimally acceptable" actually know of the content domain in which they are being examined. These judgments are expressed with respect to the content of items developed to assess that domain.

Regardless of the judgments on which a standard is based, it is useful to distinguish between the tasks of estimating the standard for a content domain and selecting the cutting score for an examination developed to measure knowledge of that domain. In this author's opinion, estimates of an examination standard

---

1Charles F. Schumacher and Benjamin D. Wright provided invaluable assistance, guidance and insight throughout the project and are co-authors of the report submitted to the NBME.

should be based on one consideration only: the level of achievement that is "minimally acceptable" or stated differently, what "minimally knowledgeable examinees" or "MKEs" actually know of the content domain. Selecting a cutting score, however, involves consideration of the estimated standard and its plausibility as well as the educational and societal impact of the cutting score and the likelihood of erroneous decisions and their consequences for the examinees and for society. (Millman, 1973)

Estimating a criterion-referenced standard, from a psychometric perspective, requires a procedure that will translate the standard-setters' judgments about what "MKEs" know into test scores. Since many procedures for estimating such a standard have been described in the psychometric literature, the choice of procedure is a decision that may influence the estimate of the standard. This investigation was conducted to provide information that could help an examining agency choose among four standard setting procedures involving judgments about test questions.

The four procedures were evaluated using the estimates they yielded of the examination standard. The evaluation focused on the consistency of the estimated standards as indicated by their standard errors, the plausibility of the failure rates associated with the estimates, the accuracy of the judgments on which the estimates were based, and the confidence and ease with which standard-setters and psychometricians could use the procedures with on-going examination programs.

REVIEW OF NBME RESEARCH

The National Board is very much aware of the continuing discussion in the psychometric literature regarding the issue of standard setting in general and the merits of normative and criterion standards in particular. During the past decade it has supported, either by itself or in cooperation with its client organizations, numerous research studies comparing the use of norm- and criterion-referenced standard setting procedures.

Andrew and Hecht (1976) found that the method described by Nedelsky (1954) yielded a much lower standard for a nationally administered certifying examination in the health professions than the method described by Ebel (1972). Guerin, Burg, and Vaughan (1978) reported that the standards for two recertifying examinations obtained using a modified Nedelsky technique were similar to the norm-referenced standards set for those examinations. Guerin, Butzin and Schumacher (1982) investigated a new procedure and found that it yielded an acceptable criterion-referenced standard for a recertifying examination that was not too different from the standard that would have been obtained had the modified Nedelsky method been used. They also

reported that different groups of standard-setters made similar judgments, a finding reported by Andrew and Hecht too. Hughes (1981) described another method that produced increased agreement among the standard-setters about the choice of a criterion-referenced standard each time they revised their previous judgments on the basis of new feedback information. However, the estimate of the standard tended to fluctuate, rather than converge, after three iterations with the procedure and it differed from the normative standard set for tests similar to the prototype examination used in the study.

The results of these studies are encouraging since they suggest that appropriate and acceptable criterion-referenced standards can be set. However, they also suggest that the choice of method may affect the criterion-referenced standard that is set. Therefore, setting a criterion-referenced standard not only requires a decision by the examining agency about who the standard-setters will be, it also involves the choice of a psychometric procedure to translate their judgments about "minimally acceptable achievement" into a test score. The current study was conducted to obtain information the National Board could use to guide its choice of a criterion-referenced standard setting procedure, should the Board decide to alter its present approach to standard setting.

STANDARD SETTING PROCEDURES

The methods described by Ebel (1972), Guerin et al. (1982) and Hughes (1981) were the subject of previous NBME studies and also were investigated in this study. The Angoff procedure (1971) was included in this study rather than the Nedelsky technique (1954) because it does not constrain the standard-setters' judgments by the number of choices examinees have when responding to an item.

The Angoff and Ebel methods use only the standard-setters' judgments about the content of the test items to estimate the standard. The Guerin procedure and the procedure described by Hughes, hereafter referred to as the NBME procedure, use the standard-setters' judgments and psychometric data obtained from a Rasch item calibration to estimate the standard (Rasch, 1960; Wright, 1968 and 1977; Wright and Stone, 1979). The calibrated item difficulties are independent of the examinees whose responses were used to conduct the calibration and of the sample of items drawn from the item pool to construct the particular examination. Therefore, it is not necessary to await the calibration of the current form of the examination before commencing the standard-setting activities, since judgments about pool items that have been previously calibrated to the examination scale can be used to estimate the standard. Neither is it necessary to use pool items that are included in the current form of the examination when estimating the standard,

although it may be desirable to do so.

The standard-setters were selected for their expertise in one of the six clinical science disciplines comprising the test; therefore, they only made judgments about the items in their clinical science subtest. Restricting the standard-setters' judgments to items in their clinical discipline was facilitated by the Rasch item calibration which estimated the difficulty of all items on a common scale. The NBME and Guerin procedures yielded an estimate of each standard-setter's personal standard on the total test because all item difficulties were calibrated to the same scale. The Angoff and Ebel procedures yielded an estimate of each personal standard on the discipline subtest, and Rasch procedures were used to equate the subtest score to a score on the total test.

The Angoff method requires each standard-setter to judge the MKEs' success rate with every item in his or her clinical science subtest. The success rate is the judge's estimate of the proportion of MKEs answering the item correctly. The sum of these success rates over all items in the subtest is the standard-setters' estimate of the MKEs' subtest score. The equivalent score on the total test is an estimate of the standard-setter's personal standard.

The NBME procedure uses the same judgments about the MKEs' success rate in conjunction with the calibrated difficulty of the items to estimate the standard-setter's personal standard. This is done using the Rasch model which postulates that the probability of a correct response to a test item (P) is a function of the examinee's knowledge (b) and the item's difficulty (d). The model hypothesizes that for every test item $(b-d) = \log(P/(1-P))$. The NBME procedure regresses the calibrated item difficulty on the logarithmic transformation of the MKEs' success rate. The regression line intercepts the difficulty axis at the point on the log-odds axis where $(b-d) = 0$. Since $b = d$ at this point, the intercept (d) is an estimate of the MKEs' knowledge (b) as measured on the calibration scale. The total test score equivalent to this measurement is the estimate of the standard-setter's personal standard. (See Figure 1 for an illustration of this method.)

The Guerin procedure uses the calibrated difficulty of the test items and the standard-setters' judgments about the relevance of the items' content to estimate their personal standards. Each standard-setter rates the items in his or her clinical science subtest as Essential, Important, Acceptable or Questionable, but only the items judged to be Essential are used to estimate the personal standard. The Rasch model estimates item difficulty and examinee achievement on the same measurement scale, and the Guerin procedure defines the point on that calibration scale occupied by the most difficult of the Essential items as the MKEs' achievement level. The total test score equivalent to that measurement is the standard-setter's personal

standard. (See Figure 2 for an illustration of this procedure.)

The Ebel method uses judgments about the MKEs' success rate with hypothetical item-types characterized by relevance and difficulty to estimate the standard-setters' personal standards. Four categories of item relevance (Essential, Important, Acceptable and Questionable) and item difficulty (Easy, On The Easy Side, On The Hard Side and Hard) are defined. The standard-setter first judges the MKEs' success rate with each of the 16 hypothetical item-types, then classifies every item in the clinical science subtest according to his or her perception of its relevance and difficulty. The success rate for a category is used as the MKEs' probability of success with every item classified in the category, and "minimally acceptable achievement" on the subtest is estimated by summing the probabilities over all the items. The equivalent score on the total test is the estimate of the standard-setter's personal standard.

COLLECTING THE STANDARD-SETTERS' JUDGMENTS

A recent National Board Part II Examination was used to estimate a criterion-referenced standard for the content domain the test was developed to assess. It contained 862 multiple choice items that were used to obtain a total test score. These items were distributed in roughly equal numbers to the Internal Medicine, Surgery, Obstetric/Gynecology, Preventive Medicine/Public Health, Pediatrics, and Psychiatry subtests. The National Board evaluates its candidates in each clinical science discipline, but it uses a single score for the total test to determine whether an examinee's achievement is "minimally acceptable." Therefore, a criterion-referenced standard was estimated for the total test using each of the psychometric procedures being investigated.

To do this a panel of standard-setters was formed consisting of twelve medical educators with previous experience as members of the National Board's Part II Test Committees. The standard setters were chosen for their recognized expertise in one of the clinical science disciplines assessed by the examination and for their experience in writing items for, and in constructing, Part II Examinations. There were two standard-setters for each clinical science discipline.

The twelve standard-setters met in Philadelphia for a two-day orientation meeting. Before the meeting they were sent an overview of the study, information about the various judgments they would be asked to make, and a small sample of items in their clinical science discipline. Two groups of examinees were defined for them: MKEs and TUSMGs whose knowledge is "typical of graduates of US medical schools". 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 described as typical graduates of US medical schools who have attained

> "... a level of medical knowledge beyond the minimum acceptable for safe and effective medical practice under supervision."

The standard-setters were instructed to review the sample of items sent to them and to make the judgments needed to estimate the MKEs' and the TUSMGs' achievement on the item sample using each procedure being investigated. They completed this "instructional exercise" before coming to the meeting and it served as the basis for a brief training session during the meeting.

Following the meeting and acting independently of one another, the standard-setters were asked to review every item in their clinical science discipline and make the following judgments in the order indicated: (1) specify the success rate for MKEs and for TUSMGs with each of the hypothetical item-types characterized by relevance and difficulty, (2) classify each item according to their perception of its relevance and difficulty, (3) specify the success rate for MKEs and for TUSMGs with each of the items. The appropriate judgments were used with each procedure to obtain the standard-setter's estimate of the MKEs' and the TUSMGs' score on the total test.

The average of the standard-setters' personal standards was the estimate of the group's examination standard. The consistency of the group's estimate was expressed as a standard error, calculated by dividing the standard deviation of the personal standards by the square root of the number of standard-setters. An estimate of the group's examination standard was obtained in this manner using the personal standards estimated with each of the procedures being studied.

PRESENTATION OF THE DATA

Estimates of the standard-setters' personal standards and the group's examination standard are reported as percent scores rather than in the standard score metric used by the National Board. This type of score often is used when criterion-referenced standards are being considered because by implication it associates the standard with a mastery level of the content domain.

Consistency of the Estimated Examination Standards

The NBME procedure yielded the most consistent of the estimates (see Table 1) with a standard error of 1.8 percent score units. The Guerin procedure yielded the least consistent estimate, the standard error being 3.3 score units. Both the Angoff and the Ebel procedures yielded estimates with a standard error of 2.5 units.

The consistency of these estimates was improved by computing the examination standard as the average of the standards estimated for the clinical science disciplines. (See Table 2.) The average of the personal standards for judges with expertise in the same clinical science was used as the standard for that discipline. When the examination standard was computed in this manner, the standard error was 1.2 using the NBME procedure, and 2.0, 1.8, and 3.2 respectively using the Ebel, Angoff and Guerin procedures. This finding indicates that the variability within disciplines was greater than the variability between disciplines and suggests that differences among the standard-setters are individual differences unrelated to the clinical discipline in which they are expert.

Plausibility of the Estimated Examination Standards

Plausibility was assessed in two ways. One involved a comparison of the failure rates that would have occurred had each of the criterion-referenced standards been used with the failure rate that did occur when the norm-referenced standard was used. The other involved an evaluation of the accuracy of the standard-setters' judgments about success rates with individual items.

The National Board reference group only contains examinees who are in their final year at a US medical school, are taking the Part II Examination for the first time, and are candidates for NBME certification. Because medical educators have accepted a failure rate of 2.4% in this group for many years, the norm-referenced standard has been considered a sensible one. The standard estimated using the Guerin procedure (see Table 3) would have resulted in an 85.5% failure rate in the reference group, which clearly would be unacceptable. The standards estimated using the NBME, Ebel and Angoff procedures all had failure rates higher than 2.4%; however, their respective failure rates (3.5%, 6.0% and 8.3%) did not differ too greatly from the normative failure rate and might be considered acceptable.

It was not possible to assess the accuracy of the standard-setters' judgments about the MKEs' success rate with individual test items because p-values could not be determined for examinees whose achievement was "minimally acceptable." It was possible, however, to compare their judgments about the TUSMGs' success rate with item p-values based on the responses of the National Board reference group and to compare estimates of

the TUSMGs' average score with the average score for the reference group. These comparisons provided information (see Tables 4-5) that was used to evaluate the plausibility of the standard-setters' judgments.

In general, the difference between the standard-setters' TUSMG judgments and the reference group p-values did not exceed 10 for about 40% of the items in their clinical discipline. However, it exceeded 15 for about 45% of those items. (See Table 4.) The average difference for eight standard-setters was positive and indicated a tendency to overestimate the TUSMGs' success. For three of them the overestimate was relatively extreme as shown by an average difference exceeding +10. Only two standard-setters tended to underestimate the TUSMGs' success, but neither did by very much. Because almost one-half of the standard-setters' TUSMG judgments were inaccurate, there is reason to question the accuracy of their MKE judgments.

Two of the three standard-setters whose tendency to overestimate TUSMG success was relatively extreme (MED2 and PMPH2) also had personal standards that were much higher than the estimate for the group. Both judges who tended to underestimate the TUSMGs' success (PMPH1 and PED52) had personal standards that were much lower than the group estimate. Because judges whose TUSMG judgments tended to be inaccurate usually had personal standards that were extreme, there may be reason to question the accuracy of their personal standards as estimates of the examination standard.

The NBME, Ebel, and Angoff procedures were used to estimate the average score for the group of TUSMGs. (See Table 5.) The Guerin procedure was not used because it only defined a rule for determining the achievement level of MKEs. The average score achieved by the National Board reference group was 65.4%. The estimate of the TUSMGs' achievement was 60.1% using the NBME procedure, 65.1% using Ebel's procedure and 69.3% using the Angoff method. The accuracy of the estimate obtained using the Ebel procedure suggests that grouping items may help improve the accuracy with which the examination standard is estimated. The estimates of the TUSMGs' average score were more consistent than the estimates of the MKEs' score possibly because the judges are more familiar with typical medical students and their level of achievement.

Feasibility of Implementing the Procedures

Different procedures use different judgments to estimate the examination standard. Therefore, the opinions of the standard-setters regarding the comparative ease of making those judgments and their confidence in them were used to help evaluate the feasibility of using the standard-setting procedures with operational examination programs.

In general, the standard-setters said it was easy to

classify test items according to their perceptions of the relevance and difficulty of the content, as required by the Ebel procedure, but relatively difficult to judge success rates with hypothetical item-types 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 the MKEs' success rate. Therefore, they were uncertain of their judgments and, by inference, lacked confidence in the estimate of their personal standard based on those judgments.

The standard-setters also said it was difficult to judge the MKEs' success rate with actual items as required by the Angoff and NBME procedures. However, they thought it was easier with actual items than with hypothetical item-types because they were tangible and could be examined both for content and format. Only one said it was easier to judge success rates for hypothetical item-types and gave as a reason the conceptual standardization imposed by the relevance and difficulty characterizations.

The standard-setters expressed the desire for concrete information about the items on which they were basing their judgments. They wanted reference points to keep them "...in touch with reality." Although p-values for the items were available based on the responses of examinees in the National Board reference group, they were not made known to the standard-setters for fear that such information might bias their judgments about the MKEs' success rates.

The standard-setters' opinions suggest that it would not be feasible to use the Ebel procedure to estimate the standard for operational examination programs. Psychometricians are likely to concur in that opinion since the extra effort required of the standard-setters and psychometricians when using Ebel's procedure did not yield an estimate that was more consistent or plausible. The other procedures require only one judgment, not three, and presented no problems either to the standard-setters or psychometricians that would detract from the feasibility of using them with on-going examination programs.

DISCUSSION

The estimate of the examination standard obtained using the NBME procedure was the most consistent of the four. It was less sensitive to aberrant judgments about individual items -- low success rates for easy items or high success rates for hard ones -- because it fits a regression line through the mean difficulty of items judged to have the same success rate rather than giving equal weight to every judgment of an MKEs' success rate as is the case with the Angoff anb Ebel procedures. In this way it diminishes the impact of aberrant judgments on the estimation of the personal standard. Thus, the error inherent in the

estimation of the examination standard is reduced and, to a greater degree than with other procedures, that error reflects actual differences among the personal standards of the standard-setters rather than inconsistencies associated with their estimation.

Only the Guerin procedure yielded an estimate of the examination standard that obviously was not plausible since a failure rate of 85.5% in the National Board reference group clearly would not be acceptable. Although the present study used only single word definitions to distinguish among four degrees of relevance and Guerin provided his standard-setters with detailed descriptions of five degrees of relevance, it seems more reasonable to explain this finding in terms of the psychometric differences between the Part II Examination and the recertification examinations Guerin and his colleagues worked with. A typical recertification examination is likely to contain a larger percentage of relatively easy items than would be found in a National Board examination; therefore, the most difficult of the highly relevant items is more likely to be easier relative to the calibration scale of the recertification examination than the Part II Examination. This tendency probably was encouraged by Guerin's description of highly relevant items which suggested that they are likely to assess content most examinees know. Consequently, in Guerin's studies the most relevant items tended to be the easiest ones, resulting in the examination standards for the recertification examinations being set at a lower achievement level than those for the Part II Examination.

The data concerning the feasibility of implementing the standard setting procedures were a mixture of opinions. The standard-setters found it easier to judge the relevance of individual test items than the MKEs' success rate with those items, and they were less confident about judging the MKEs' success rate with hypothetical item-types than with actual test items. However, the Ebel procedure which requires judgments of item relevance and difficulty and of success rates with groups of hypothetical item-types yielded the most accurate estimate of the average score for the National Board reference group. This occurred even though the standard-setters said they were uncertain about the impact of changes in item relevance and difficulty on their judgments about success rates and suggests that it may be more efficacious for standard-setters to judge success rates for groups of items with common characteristics than for individual items.

It was not an objective of this study to determine which criterion-referenced procedures yielded estimates that closely approximated the norm-referenced standard used by the National Board, but the similarity between the normative standard and three of the estimates requires comment. The standard-setters used in this study were not involved in the process by which the National Board determined its norm-referenced standard and the judgments they expressed in this study were different from the

judgments used to determine that standard. Therefore, the fact that the normative standard is closely approximated by three estimates of the criterion-referenced standard suggests that, in the judgment of this group of standard-setters, it too may be a reasonable estimate of a criterion-referenced standard representing "minimally acceptable achievement."

SUMMARY, CONCLUSIONS, AND RECOMMENDATIONS

Four procedures were used to estimate a criterion-referenced standard for a multiple-choice examination developed by the National Board of Medical Examiners. The procedures were evaluated on the basis of their consistency, plausibility, and feasibility.

The NBME procedure yielded the most consistent, and the Guerin procedure the least consistent, estimate of the examination standard. The consistency of all the estimates increased when the examination standard was computed as the average of discipline standards rather than personal standards.

Only the Guerin procedure yielded an estimate of the examination standard that clearly was not plausible since 85.5% of the NBME reference group would have failed had it been used. The outcomes associated with the other estimates could be considered plausible since they did not differ greatly from the 2.4% failure rate experienced with the normative standard.

There is reason to doubt the accuracy of the standard-setters' judgments about the MKEs' success rates with individual items since their judgments about the reference group's success rates were inaccurate for roughly 45% of the items. This raises questions about the validity of the standard estimated using those judgments, especially since standard-setters who made the least accurate judgments about the TUSMGs' success rate also had personal standards that were extreme. These findings and a desire for information about the items' performance suggest the need to provide some guidance to the standard-setters as they judge the MKEs' success rate with test questions.

The Angoff, Guerin and NBME procedures were considered feasible for use in estimating the standard with operational examination programs because they presented no unusual problems either to the standard-setters or the psychometricians. The Ebel procedure was not considered feasible, because the standard-setters' found it difficult and confusing to judge success rates for hypothetical item-types characterized by varying degrees of relevance and difficulty. However, there was some evidence to suggest that making judgments about groups of items rather than individual items may be advantageous.

Based on these findings it is recommended that the NBME procedure continue to be investigated. It appeared to be more promising than the other methods investigated, and the findings of this study suggest that several modifications may enhance its attractiveness as a procedure for estimating a criterion-referenced standard.

One modification would estimate the standard-setters' personal standards using only those items for which their judgment of the TUSMGs' success rate was a reasonably accurate estimate of the reference group's p-value. Another would have the standard-setters judge the MKEs' success rate with clusters of highly relevant items of varying difficulty rather than with individual items, thereby retaining the more desirable features of the Ebel procedure while eliminating the potential for confusion arising from the use of items that in the standard-setter's judgment are not relevant. A third modification would provide the standard-setters the opportunity to review their previous judgments in the light of feedback about the MKEs' success rate implied by the current estimate of their personal standard and the group's examination standard.

These modifications could yield more consistent estimates of the judges' personal standards and of the group's examination standard. Furthermore, the use of feedback could provide the standard-setters with a mechanism for refining the estimate of their personal standards and for approaching consensus about the examination standard in an atmosphere of reasoned and deliberative judgment devoid of heated argument and advocacy. Studies involving these modifications have been planned.

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.

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.

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. Chicago, University of Chicago Press, 1980.

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

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

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

FIGURE 1

An Illustration of the NBME Procedure

FIGURE 1

An Illustration of the NBME Procedure

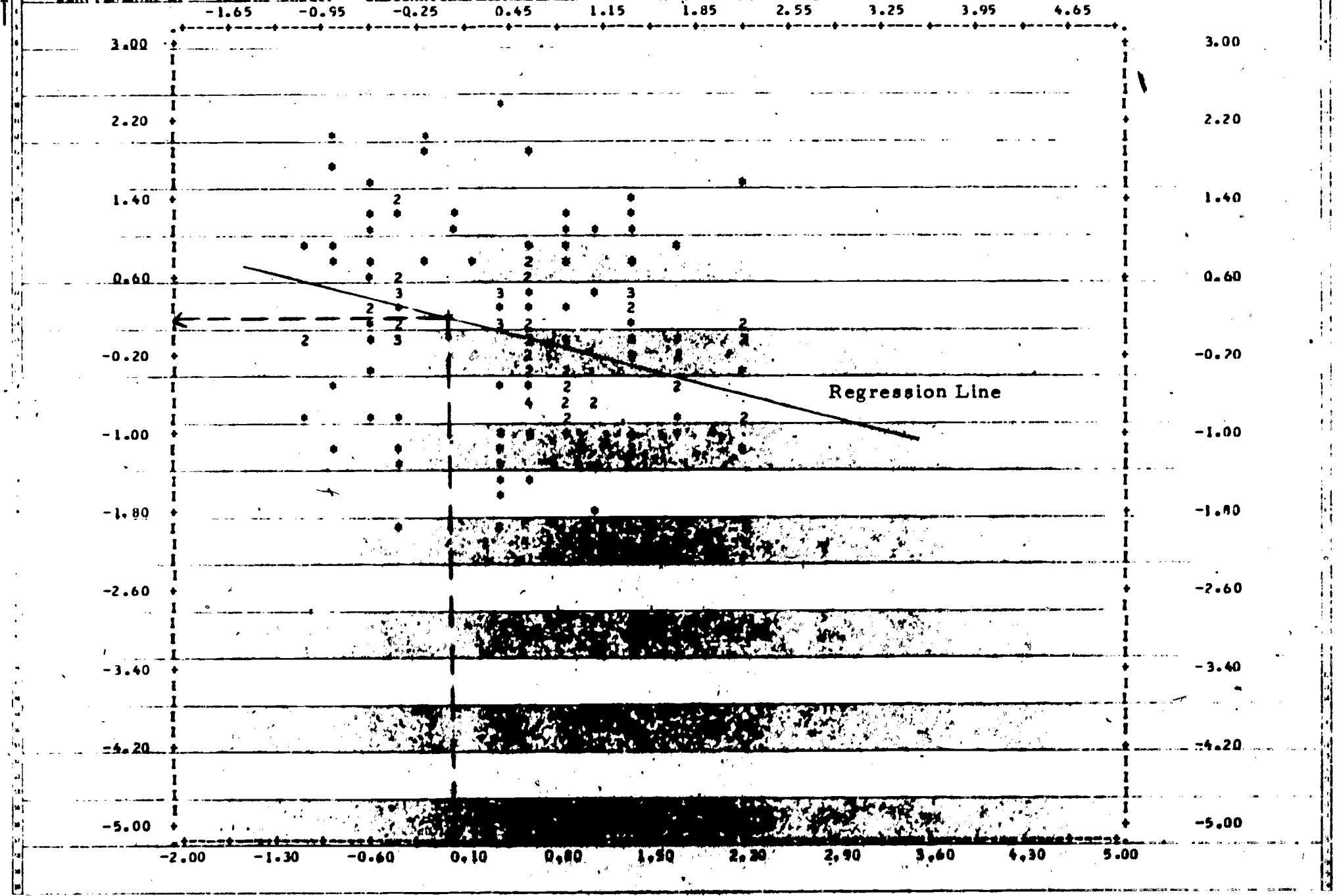
$$P = \exp(b-d)/[1 + \exp(b-d)]$$

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

Regress d on $\log [P/(1-P)]$

Then, the intercept occurs at $(b-d) = 0$ (i.e., where $b = d$)

FILE NCNAME (CREATION DATE = 06/01/82)
SCATTERGRAM OF (DOWN) CALDIF CALIBRATION DIFFICULTY (ACROSS) LOGMCE MOE LOG ODDS SUCCESS



STATISTICS..

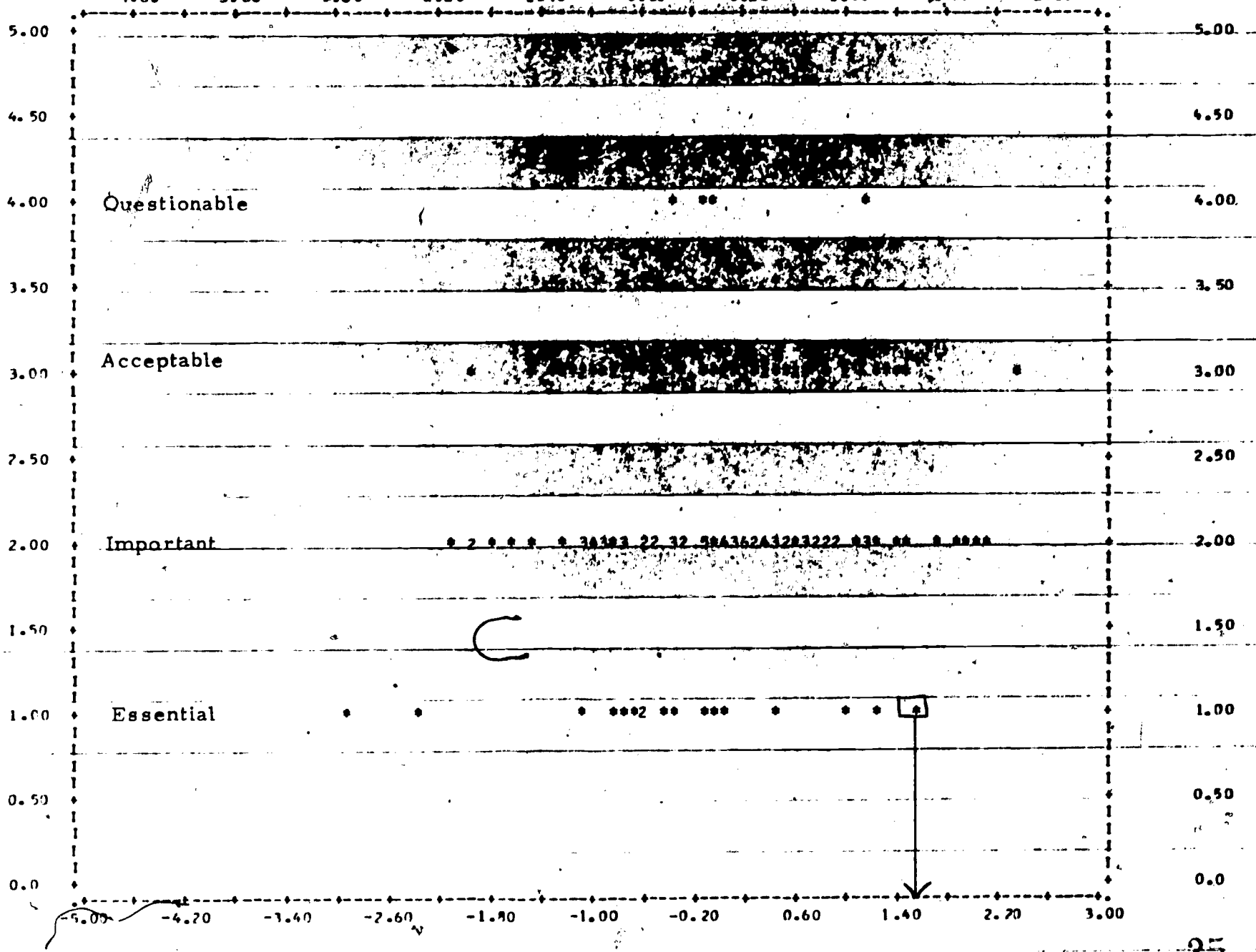
CORRELATION (R)-	-0.28685	R SQUARED -	0.08228	SIGNIFICANCE -	0.00054
STD ERR OF EST -	0.94237	INTERCEPT (A) -	0.13685	SLOPE (B) -	-0.31169
PLOTTED VALUES -	142	EXCLUDED VALUES-	10	MISSING VALUES -	10

'*********' IS PRINTED IF A COEFFICIENT CANNOT BE COMPUTED.

FIGURE 2

An Illustration of the Guerin Procedure

FILE NONAME (CREATION DATE = 06/01/82)
SCATTERGRAM OF (DOWN) RM RELEVANCE RATING (across) ITEM DIFFICULTY



STATISTICS..

CORRELATION (R)-	0.09486	R SQUARED -	0.00900	SIGNIFICANCE -	0.26144
STD ERR OF EST -	0.68176	INTERCEPT (A) -	2.21256	SLOPE (B) -	0.06604
PLOTTED VALUES -	142	EXCLUDED VALUES-	0	MISSING VALUES -	10

'*********' IS PRINTED IF A COEFFICIENT CANNOT BE COMPUTED.

TABLE 1

ESTIMATES OF PERSONAL STANDARDS AND THE EXAMINATION STANDARD

Standard-Setters	Standard Setting Procedure			
	NBME	Ebel	Angoff	Guerin
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				
Standard	52.4	54.2	55.4	73.2
SD	6.4	8.8	8.7	11.6
Std. Error	1.8	2.5	2.5	3.3
Standard - 2 SE	48.8	49.2	50.4	66.6
Standard + 2 SE	56.0	59.2	60.4	79.8

TABLE 2

ESTIMATES OF THE DISCIPLINE STANDARDS AND THE EXAMINATION STANDARD

Discipline	Standard Setting Procedure			
	NBME	Ebel	Angoff	Guerin
Med	55.8	51.8	58.3	70.9
Surg	52.2	58.0	58.7	80.6
Ob/Gyn	55.7	62.0	59.4	85.6
PMPH	53.1	53.3	55.4	62.7
Peds	47.1	46.0	46.1	73.5
Psych	50.5	54.3	54.6	65.7
GROUP				
Standard	52.4	54.2	55.4	73.2
SD	3.0	5.0	4.5	7.9
Std. Error	1.2	2.0	1.8	3.2
Standard - 2 SE	50.0	50.2	51.8	66.8
Standard + 2 SE	54.8	58.2	59.0	79.6

TABLE 3

FAILURE RATES ASSOCIATED WITH ESTIMATES OF THE EXAMINATION STANDARD

	Norm-Referenced Estimate	Criterion-Referenced Estimate			
		NBME	Ebel	Angoff	Guerin
Estimated Standard	50.5	52.4	54.2	55.4	73.2
Reference Group Failure Rate	2.4% (n=113)	3.5% (n=169)	6.0% (n=289)	8.3% (n=400)	85.8% (n=4113)

TABLE 4

DIFFERENCE BETWEEN TUSMG JUDGMENT AND REFERENCE GROUP P-VALUE

Standard-Setters	Magnitude of the Difference: Less than or Equal to 10 n	%	Between 11 and 15 n	%	Greater than 15 n	%	Mean Difference* 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 TUSMGs' success rate and a positive difference indicates that the standard-setter overestimated the TUSMGs' success rate.

TABLE 5

ESTIMATES OF THE AVERAGE SCORE
IN THE NATIONAL BOARD REFERENCE GROUP
(Average = 65.4%; Percentile Rank = 49th)

Standard-Setters	Estimation Procedure		
	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
Group Average	60.1	65.1	69.3
(Percentile Rank)	(21st)	(48th)	(70th)
SD	4.8	6.8	6.1
Stnd. Error	1.4	2.0	1.8
Average - 2 SE	57.3	61.1	65.7
Average + 2 SE	62.9	69.1	72.9