ACT Research Report Series 89-2

A Comparison of the Angoff, Beuk, and Hofstee Methods for Setting a Passing Score

John J. Bowers, PhD
Russelyn Roby Shindoll

May 1989

ACT

For additional copies write:
ACT Research Report Series
P.O. Box 168
Iowa City, Iowa 52243

A COMPARISON OF THE ANGOFF, BEUK, AND HOFSTEE METHODS FOR SETTING A PASSING SCORE

John J. Bowers, PhD
Russelyn Roby Shindoll

TABLE OF CONTENTS

ABSTRACT

This study compared four methods for setting a passing score on a 200-item nationally administered professional certification examination. Angoff ratings made without item difficulty available to raters had a moderately low positive correlation with item difficulties and produced a high passing score. Angoff ratings made with item difficulty data available correlated highly with those difficulties and produced a lower passing score. The Beuk and Hofstee approaches seek to produce a passing score that compromises between differences in ratings between raters who focus primarily on an absolute standard (i.e., percent of items correct) and those who focus on a relative standard (i.e., percent of examinees passing). The Beuk and Hofstee methods produced identical passing scores within the parameters set by the two sets of Angoff ratings. Further research on the Beuk and Hofstee methods is recommended.

A COMPARISON OF THE ANGOFF, BEUK, AND HOFSTEE METHODS FOR SETTING A PASSING SCORE

Purpose

This study compared four methods for determining a passing score (or cut-off score) for an examination. Two versions of the Angoff method (1971) were studied, differing in whether or not the judges were provided with item difficulty data. Two newer "compromise" methods proposed by Beuk (1984) and Hofstee (1983) were also studied.

Background

A passing score on a professional licensing or a certification examination distinguishes examinees who are at least minimally competent to practice from those who are not. It is critical that the passing score be established fairly and that the basis on which it is established be defensible.

All three methods in this study attempt to set a standard, expressed as the percent of items answered correctly, that an examinee must achieve in order to demonstrate at least minimal competence. Because the focus of the examination is on "competence," not on advanced or expert levels of skills and knowledge, the purpose of the passing score study was to determine the performance of the "borderline" test taker, i.e., the performance of an individual who is just above the borderline that separates competent from incompetent performance.

The Angoff procedure (Angoff, 1971; Livingston & Zieky, 1982) specifically addresses the issue of "borderline" performance by requiring judges to estimate the performance of minimally competent candidates on each item on a particular examination form. The estimates for each judge are added and the sums averaged across judges to obtain an estimate of the minimally acceptable score. This score could become the passing score on the examination. The Angoff procedure is one of the most widely used methods for setting passing scores because it is simple to perform and deals directly with estimates of minimal competence for test content. Nonetheless, some problems are associated with the procedure.

One issue is that content experts who qualify as judges tend to think of examinees who are average or above average, rather than those who are only minimally competent. If the concept of "minimal competence" is not clear to judges, estimates of performance would be likely to reflect that for average or above average examinees and result in an unrealistically high standard. To avoid this problem, it is important to train judges well in the concept of the minimally competent examinee both before and during the rating process (Francis & Holmes, 1983).

Another issue is whether the difficulty of each item, i.e., the percent of examinees who answered the item correctly in its last use, should be provided to judges. One position holds that the standard on which the passing score is established is supposed to be criterion-referenced, representing a level of competence in a specific domain of knowledge. As such, it should not be influenced by the number of persons passing or failing. However, when performance data are considered, an element of norm-referencing is introduced, causing the new passing score to be closer to the current passing score and performance of the examinee group.

Another position holds that providing item difficulty data allows judges to temper their ratings. If their expectation of what constitutes a minimally competent candidate is unrealistically high, they will produce a passing score that has an unacceptably high failure rate. Providing item difficulty data would allow judges to focus on what a minimally competent candidate will do, not what he or she ought to do. One way to determine if the availability of item performance information influences judges' ratings is to have judges rate each item twice--first without item difficulty data available and then with the item difficulty data provided.

One would expect that ratings made with item difficulty data would (1) correlate more highly with the item difficulties and (2) result in lower passing scores, than ratings made without item performance data. The first would happen because judges who make use of the item difficulty data would tend to adjust their ratings to make them more consistent with the item difficulties. The second would occur because the raters often set high standards when they do not use the data. Providing item data would be likely to draw the judges' ratings closer to actual examinee performance and result in a lower passing score.

Passing score studies reported in the literature generally confirm the first effect but provide mixed results on the second. Several studies (Harker & Cope, 1988; Cope, 1987; Garrido & Payne, 1987; Cross, et al, 1984) all report higher correlations between ratings and item difficulty data when difficulty data were made available to judges than when the data were not made available. Cross, et al (1984) reported that passing scores dropped dramatically when judges were provided with normative data on test items. However, Harker and Cope (1988) found that providing item difficulty data resulted in judges' ratings being pulled downward for difficult items and upward for easy items but that "these adjustments tended to balance and to result in relatively small mean changes" in the resulting passing score. Other studies (Cope, 1987; Garrido & Payne, 1987) found that providing item difficulty data actually resulted in higher passing scores.

More recently, other methods for setting passing scores have been proposed that focus on the differences between judges. The procedures proposed by Beuk (1984) and Hofstee (1983) are usually called "compromise methods" because they seek to resolve differences in ratings between judges who focus primarily on an absolute standard (i.e., percent of items correct) and those who focus primarily on a relative standard (i.e., percent of examinees passing). The methods are somewhat similar in their operation. Both require judges to set a passing score directly without consideration of individual item data, and incorporate judges' estimates of the performance of the examinee population. Both methods need the actual distribution of test scores and so cannot arrive at the passing score until after the test has been administered and scored. However, the information from the judges can be gathered prior to the test administration.

In Beuk's approach, a passing score set by a committee is adjusted according to the degree to which judges are candidate oriented or test oriented, i.e., whether they have a relative or an absolute standard. Each judge is asked to specify the knowledge level that an examinee should possess, expressed as the minimum percent of items answered correctly on the test, and the expected pass rate for that score, expressed as a percent of the examinees passing. The mean and standard deviation of both ratings are computed. These are denoted as $\bar{k}$ and s_k for the estimated passing score and $\bar{v}$ and s_v for the estimated passing rate. The point represented by these two means $(\bar{k}, \bar{v})$ is plotted on a graph of the actual passing rate as a function of the passing score. A line having a slope equal to s_v/s_k is drawn through this point. The point where the line intersects the curve representing the passing rate as a function of the passing scores represents the passing score.

In Hofstee's approach, each judge provides the minimum acceptable percentage of failing examinees (f_{min}), the maximum acceptable percentage of failing examinees (f_{max}), the minimum acceptable percentage of items that a minimally competent examinee should have to answer correctly (k_{min}), and the maximum acceptable percentage of items that a minimally competent examinee should have to answer correctly (k_{max}). The passing score is derived by averaging the judges' ratings and plotting two points: the intersection of k_{min} and f_{max} and the intersection of k_{max} and f_{min}. The point where the line intersects a curve representing the fail rate for each score point represents the passing score.

The Beuk method has an advantage over the Hofstee method in that it will always produce a passing score. The Hofstee method will fail to produce a passing score if the judges' estimates fall entirely above or below the score curve. If the judges' estimate of the highest acceptable failure rate is lower than the actual percent of the examinee group who would fail at the score representing judges' estimate of the lowest acceptable passing score, the line that is supposed to intersect the score curve will instead lie entirely under it. Likewise, if the judges' estimate of the lowest acceptable failure rate is higher than the actual percent of the examinee group who would fail at the score representing the judges' estimate of the highest acceptable passing score, the line that is supposed to intersect the score curve will instead lie entirely above it. In either case, a passing score can not be derived.

Specific details of these methods are discussed later in this paper.

Instrument

The examination was a nationally administered professional certification examination consisting of 200 four-option multiple-choice items. All but three of the items had been used on previous forms of the examination so that the difficulty of 197 items was known. The form studied here was administered in early 1988. Table 1 provides descriptive statistics for this test form.

Procedures

Five of the seven examination committee members (hereafter called judges or raters) met at ACT National Headquarters in Iowa City, Iowa, in the fall of 1987 to conduct the passing score study. ACT staff provided a brief orientation regarding the purpose and agenda of the meeting and presented some of the issues described above. Following an explanation of the Angoff procedure, judges were instructed how to make estimates about the proportion of minimally competent candidates who would answer a particular question correctly. Judges completed the two training exercises described below before making estimates on the 200 items that appeared on the examination.

Training the Judges

First, the judges were given materials to help them understand the objectives of the study and define a concept of minimal competence. These materials and the accompanying instruction helped judges to focus on the "minimally competent" examinee, rather than the average or above average examinee. In generating this image, the judges were asked to recall someone they might have known who was just barely competent. The judges shared perspectives and examples among themselves, but no attempt was made to reach consensus on what constitutes minimal competence. The judges were instructed to estimate what proportion of such a group of minimally competent examinees would (not should) answer each item correctly. They were also asked to consider the frequency and criticality of the knowledge or skill represented by each item. As noted in the instruction materials, it was probable that a greater proportion of minimally competent examinees would know the content of items having a higher frequency or criticality in practice.

Second, the judges completed an exercise replicating the Angoff procedure. A dozen items were selected from the examination item pool so as to represent roughly the content distribution and average difficulty level of the examination; however, these items did not appear on the examination under study. The judges were asked to make estimates for these sample items twice--once without the performance data available, and once with the performance data supplied.

The judges first rated the items without the performance data available, estimating for each item the percent of minimally competent examinees whom they judged would answer the item correctly. Although no upper limit was placed on their ratings, it was suggested that a logical lower limit of 25% was justified. This lower limit was established because there are four choices for each item and 25% of the minimally competent examinees would be expected to answer the item correctly by random guessing.

For their second set of ratings, judges were provided with item statistics from a previous administration of the 12 items. They also received instruction on how to read and interpret difficulty and discrimination indices. The data included the percent of examinees selecting each response for both the total group of examinees as well as three subgroups formed from those whose scores were in the upper 27 percent, middle 46 percent, and lower 27 percent of the examinees.

Before the judges made their second set of ratings, they were advised about the possible effects item difficulty may have on their ratings. Following Klein's (1984) suggestion, it was reasoned that for item performance data to be useful to judges, the item statistics needed to be put into the perspective of the overall program. Because item difficulty represents the performance of the average examinee, the performance represented by the item difficulty would likely be considerably higher than that of the minimally competent examinee. Thus, it seemed reasonable to advise the judges to use the performance of the middle 46 percent of examinees as an upper boundary of what could be expected of minimally competent examinees. However, the judges were also advised that their rating for a given item could exceed the item difficulty if the content represented a high level of criticality and frequency. Using the item performance data then, given the flexible upper and lower boundaries, helped temper unrealistic estimates of performance without actually predetermining the estimates, or the passing scores.

The judges recorded their estimates on separate ratings sheets for each exercise. In addition, they did not have access to the estimates from the first exercise while recording their estimates for the second exercise. The data were quickly analyzed, and the judges given the opportunity to discuss how the results obtained by the two methods differed. Additional questions about the rating process and about minimal competence were resolved before the actual passing score study was conducted.

Collecting Ratings

Once all questions about the procedure were resolved, ACT staff distributed to each judge a booklet containing the 200 items on the examination, without item difficulty data available, and a set of rating sheets. Each page of the booklet contained one item, with its text exactly as it would appear on the examination. Items were arranged in order by their content classification in the test blueprint. During the remainder of the first rating session, the judges recorded their estimates. When the judges completed making their ratings, ACT staff entered them in a microcomputer database analysis program that produced a complete listing of the judges' ratings including the mean and standard deviation of the ratings for each item. The standard deviation represents the degree of variability among the ratings--the greater the degree of disagreement, the higher the standard deviation. Judges were provided with a copy of this list. For the items that had the highest standard deviations, the judges discussed the reasons for their differences, and some revised their ratings based on this discussion. Of the relatively few revisions that were made, most consisted of one judge lowering, while another raised, his or her rating. The overall effect of these changes was deemed too small to analyze. The judges were not informed of the overall results of the first session, i.e., the passing score that would have resulted based on the first set of ratings.

In the second session, judges were given a second booklet, containing the same 200 items, except that item difficulty data were provided for each item (except for three items that had not been used previously). The judges were reminded that it was unlikely, though not impossible, that their estimates would exceed the performance of the middle group of examinees. Again, the judges discussed the reasons for their differences on items whose ratings had the highest standard deviations. However, the total number of revised ratings was even smaller than at the first session. As in the first session, judges were not informed of the passing score that would result from their ratings.

On the morning of the second day, the judges completed the questionnaires for the Beuk and Hofstee methods. The first two questions gathered data for use with the Beuk method and asked (1) the minimum percent of items that an examinee should answer to pass the examination and (2) what percent of the examinees would achieve this score or higher, assuming it was the passing score. The remaining four questions gathered data for use with the Hofstee method and asked (1) the lowest acceptable percentage of failing examinees, (2) the highest acceptable percentage of failing examinees, (3) the lowest acceptable passing score (expressed as percent correct), and (4) the highest acceptable passing score. The judges were instructed how to complete these questions, and they discussed the questions before answering them. However, they did not know each other's answers to the questions.

After completing the questionnaire, the judges were given a preliminary analysis of the data from the Angoff ratings. The judges briefly discussed the results and the meeting was adjourned.

Analysis and Results

For reference purposes, the Angoff ratings made without access to the item difficulties will be referred to as the Angoff #1 data, and the ratings made with access to item difficulties as the Angoff #2 data. All ratings and item difficulties in the tables and figures that follow are expressed in terms of percents, rather than proportions.

The judges' Angoff #1 and Angoff #2 ratings are summarized in Table 2. The analysis showed passing scores of 75.415 percent, and 68.374 percent, of the items correct, respectively, which convert to 150.830 and 136.748 items of the 200 items correct. These scores round to 151 and 137 items correct. The Angoff #2 passing score was significantly lower.

The standard deviation of their ratings increased from 5.785 to 10.942, the latter being closer to the actual standard deviation of the item difficulties (11.170) suggesting that the judges' Angoff #2 ratings conformed more closely to the item difficulties than their Angoff #1 ratings.

Table 3 presents intercorrelations of the judges' Angoff #1 and Angoff #2 ratings and the actual item difficulties. The Angoff #1 ratings correlated only +0.319 with item difficulties, and only +0.344 with the Angoff #2 ratings. However, the Angoff #2 ratings correlated +0.986 with item difficulties, again confirming the suggestion that providing item difficulty data caused the judges to align their ratings with that data. Correlations for individual judges' ratings showed a similar pattern.

Figure 1 presents an overlay plot of Angoff #1 and Angoff #2 ratings against item difficulties. Angoff #1 ratings tended to be pulled downward for difficult (low p-value) items and upward for easier (high p-value) items, though far more were lowered than raised. The figure also shows that the Angoff #2 ratings were somewhat farther below the actual item difficulties for easier items than for difficult items. Figure 2 presents a plot of differences in mean item ratings (Angoff #1 - Angoff #2) against item difficulties. This, too, makes clear the fact that Angoff #2 ratings were pulled downward more for harder items than for easier items.

A generalizability analysis of the Angoff ratings was conducted using GENOVA (Crick and Brennan, 1982; Brennan, 1983). Tables 4 and 5 report the ANOVA results, estimated random effects variance components, and estimates of mean cutting score variability for the two sets of ratings. To facilitate interpretation, the estimates of mean score variability are also reported in terms of standard deviations and number of items on the 200-item test.

Tables 4 and 5 show that the estimate of variance, over the population of raters, declined slightly from 7.370 to 6.205, but that the estimate of variance, over the population of items, rose sharply from 22.400 to 117.115. Note also that $\hat{\alpha}(\overline{X})$, the estimated standard deviation for generalizing over raters and items, and $\hat{\alpha}(\overline{X}|I)$, the estimated standard deviation of $\overline{X}$ for generalizing over samples of n_r raters, stay about the same from Angoff #1 to Angoff #2 and are very similar in magnitude. However, $\hat{\alpha}(\overline{X}|R)$, the standard deviation of $\overline{X}$ for generalizing over samples of n_i items is much smaller than either of these and nearly doubles from the Angoff #1 to the Angoff #2 ratings. These results are consistent with the data in Figure 1 which suggests that in making their ratings with the item difficulty data available, the judges varied slightly less among themselves and much more across items.

Table 6 summarizes the data obtained from the Beuk and Hofstee methods. Because both methods derive a passing score by graphing the data against a known distribution of scores under consideration for a particular test, the passing scores for the Beuk and Hofstee methods were determined after the test had been administered, about three months after the judges had made their ratings.

Figure 3 shows the Beuk method for deriving the passing score. A curve is drawn that shows the pass rate for all score points, i.e., the percent of candidates who would pass the examination at each score point, if that point were the passing score. A point is then plotted on the graph showing the mean pass score and mean expected pass rate, and a line with a slope equal to the ratio of the two standard deviations is drawn through this point until it intersects the passing score curve. The point of intersection with the curve is the passing score. The Beuk method produced a passing score of 71.91% of the items correct or 143.8 items, which rounds to 144 items correct.

Figure 4 shows the Hofstee method for deriving the passing score. A curve is drawn that shows the cumulative fail rate for all score points, i.e., the percent of candidates who would fail the examination at each score point, if that point were the passing score. Two points are then plotted: one point represents the means for the maximum acceptable value of the passing score and the minimum acceptable failure rate; the second point represents the means for the minimum acceptable value of the passing score and the maximum acceptable failure rate. A line is drawn between the two points and where it intersects the curve is the passing score. The Hofstee method produced a passing score of 71.95% of the items correct or 143.8 items, which rounds to 144 items correct. Thus, the Beuk and Hofstee methods produced essentially identical results.

Table 7 summarizes the results of the passing score methods in this study, including the percent of examinees who would have passed or failed for each standard. The highest passing score was produced by the Angoff #1 method and the lowest passing score and failing rate was produced by the Angoff #2 method. The Beuk and Hofstee methods produced the same passing score, falling in between the Angoff #1 and Angoff #2 methods. Although the highest and lowest passing scores were separated by only 13 items out of the 200 on the test, the passing rates were markedly different. Roughly 26 percent more of the examinees would have passed under the Angoff #2 standard than under the Angoff #1 standard.

Discussion

This study compared two variations on the Angoff method and the methods proposed by Beuk (1984) and Hofstee (1983) that seek a compromise between judges' absolute and relative standards for setting passing scores.

Angoff #1 ratings (made without knowledge of item difficulties) yielded the highest passing score of the four methods while Angoff #2 ratings (made with knowledge of item difficulties) yielded the lowest. Compared to the Angoff #1 ratings, the Angoff #2 ratings correlated much more highly with the item difficulties, the ratings being pulled downward more for difficult items than for easier items. Generalizability analyses showed relatively small changes in the observed variance of rater means and the estimates of mean cutting score variability, from Angoff #1 to Angoff #2. However, the analyses also showed large changes in the internal structure of the variance. The main effect for items rose dramatically, while the item-rater interaction dropped sharply. The main effect for raters dropped slightly.

The other two methods examined in this study seek to set a passing score by compromising between ratings made by judges when they focus on an absolute standard (i.e., percent of items correct) and ratings made by judges when they focus on a relative standard (i.e., percent of examinees passing). The Beuk and Hofstee methods produced identical passing scores within the parameters set by the Angoff #1 and Angoff #2 results, although somewhat closer to Angoff #1. The Beuk/Hofstee results were accepted as the new passing score by the examination committee. It was a "compromise" passing score that the committee viewed as appropriate with respect to both absolute (percent of items correct) and relative (percent of examinees passing) standards.

This still leaves unanswered the question of whether providing item performance information is a good or bad practice. Both sets of Angoff ratings are unclear with respect to their interpretability. Providing item difficulties appeared to make the judges' ratings more consistent, but it does not necessarily follow that the Angoff #2 ratings possess greater validity. It may be possible to have highly consistent ratings that yield an unrealistically high or low passing score. There appears to be no clear way to determine the validity of either approach.

In the Angoff #1 method, withholding item difficulty data may have encouraged judges to focus on an absolute standard that yielded a high passing score. In the Angoff #2 method, providing item difficulty data may have encouraged judges to focus on a relative standard that yielded a low passing score. The Beuk and Hofstee methods recognize that judges are sometimes oriented towards an absolute standard (i.e., percent of items correct) and sometimes oriented towards a relative standard (i.e., percent of examinees passing) and seek a compromise between these two positions. This study suggests that, for the Angoff method, the nature of the process itself may encourage an absolute or relative standard among the judges. A method for determining a compromise between the two Angoff results would have been useful.

In this study, the Beuk and Hofstee produced an acceptable passing score between the two Angoff passing scores. Although the Beuk and Hofstee methods produced identical passing scores, that may simply be due to the group of judges in this study. Other studies may find that judges produce different passing scores with the two methods.

One major advantage of the Beuk and Hofstee methods is that they involve judges' estimates of both the test performance required for minimal competence and also the performance of the examinee group. As Mills and Melican (1987) note, "these estimates are important collateral information because they are generally based on solid observation of the examinee population."

The Beuk and Hofstee methods also offer some practical advantages. They are easy for judges to understand, are simple and quick to perform, and require no complicated data analysis. The entire data gathering process for the two methods combined in this study took only about 40 minutes. If circumstances do not permit the use of the Angoff method, then these may be viable alternatives.

A major problem with the Beuk and Hofstee methods is that they do not require judges to consider carefully the content of each item in making their ratings. In this study, judges performed the Beuk and Hofstee ratings the data after carefully considering the content of each test item twice, once for the Angoff #1 ratings and again for the Angoff #2 ratings. Therefore, they had an intimate knowledge of the test before completing the Beuk and Hofstee questionnaire. Their ratings might have been very different if they had not been so familiar with the test content. Certainly, judges should be familiar with the content of the test and have read the items, but that may not substitute for the kind of careful consideration of each item required by the Angoff approach. Further studies comparing Beuk and Hofstee ratings made both before and after performing the Angoff method may clarify this issue.

The Beuk and Hofstee methods also have a few practical disadvantages. They both require that the examination be administered before the passing score can be determined, although this is unlikely to prevent their use in most testing situations. In the Hofstee method, it is possible for judges' ratings to produce a line that does not intersect the score curve and thus fail to produce a passing score. This is not a problem with the Beuk method, which will always produce a passing score.

The Beuk and Hofstee methods deserve further investigation and comparison with the Angoff method and others. One approach might be to modify Angoff ratings using the Beuk method. For example, after the judges have made their Angoff ratings, each judge would be told the passing score that would result from his or her ratings. Each judge would then estimate the percent of examinees who would pass if that were the passing score. The means and standard deviations of these ratings could then be used in the Beuk procedure to adjust the Angoff results.

REFERENCES

Angoff, W. H. (1971). Scales, norms and equivalent scores. In R. L. Thorndike, (Ed.), Educational Measurement (2nd ed.). Washington, DC: American Council on Education.

Berk, R. A. (1986). A consumer's guide to setting performance standards on criterion-referenced tests. Review of Educational Research, 56, 137-172.

Beuk, C. H. (1984). A method for reaching a compromise between absolute and relative standards in examinations. Journal of Educational Measurement, 21, 147-152.

Cope, R. T. (1987). A generalizability study of the Angoff method applied to setting cut-off scores of professional certification tests. . Paper presented at the annual meeting of the American Educational Research Association, Washington, DC.

Cross, L., Impara, J. C., Frary, R. B., & Jaeger, R. M. (1984). A comparison of three methods for establishing minimum standards on the National Teacher Examinations. Journal of Educational Measurement, 21, 113-129.

Francis, A. & Holmes, S. E. (1983). Criterion-referenced standard-setting in certification and licensure: Defining the minimally competent candidate. Paper presented at the annual meeting of the American Psychological Association, Anaheim, CA.

Garrido, M. & Payne, D. (1987). An experimental study of the effect of judges' knowledge of item data on two forms of the Angoff standard setting methods. Paper presented at the annual meeting of the National Council on Measurement in Education, Washington, DC.

Harker, J. K. & Cope, R. T. (1988). The effect of several variables on judgmentally-obtained cut scores. Paper presented at the annual meeting of the American Educational Research Association, New Orleans, LA.

Hofstee, W. K. B. (1983). The case for compromise in educational selection and grading. In S. B. Anderson and J. S. Helmick (Eds.) On educational testing (pp. 109-127). San Francisco: Jossey-Bass.

Klein, L. W. (1984). Practical considerations in the design of standard setting studies in health occupations. Paper presented at the annual meeting of the American Educational Research Association, New Orleans, La.

Livingston, S. A. & Zieky, M. J. (1982). Passing scores: A manual for setting standards of performance on educational and occupational tests. Princeton, NJ: Educational Testing Service.

Mills, C. N. and Melican, G. J. (1987). A preliminary investigation of three compromise methods for establishing cut-off scores. Research Report 87-14. Princeton, NJ: Educational Testing Service.

TABLE 1

Descriptive Information for the Test

Number of items		200
Number of examinees		1,621
Raw score*:	Mean	149.922
	SD	15.318
Skewness		-0.577
Reliability (KR-20)		.86
Standard error of measurement		5.731
Item difficulty**:	Mean	75.025
	SD	11.170

*Number of items correct

**Percent of examinees answering the item correctly

TABLE 2

Summary of Judges' Angoff #1 and Angoff #2 Ratings

Judge	Angoff #1 Mean	Angoff #1 SD*	Angoff #2 Mean	Angoff #2 SD*	Difference: ($Mean_1 - Mean_2$)
1	79.750	7.481	71.295	12.508	8.455
2	73.970	12.201	65.430	11.076	8.540
3	76.510	8.401	69.050	11.193	7.460
4	72.890	6.500	66.155	11.815	6.735
5	73.965	8.422	69.960	10.328	5.591
Group	75.415	5.785	68.374	10.942	7.041

NOTE: All means and standard deviations are over items. Ratings are expressed as percents. For a paired t-test for the difference between the group means: $t = 9.0992$, $df = 4$, $p < .001$.

TABLE 3

Correlations Among Judges' Ratings, Mean Ratings, and Item Difficulties for the Angoff #1 and Angoff #2 Methods

	Angoff #1						Angoff #2						Item
Judges-->	1	2	3	4	5	$Mean_1$	1	2	3	4	5	$Mean_2$	Diff*
Angoff 1													
Judge 1	1.000	.371	.373	.268	.216	.647	.217	.196	.216	.192	.218	.216	.196
Judge 2		1.000	.352	.268	.355	.784	.220	.210	.260	.184	.254	.234	.229
Judge 3			1.000	.186	.282	.660	.177	.127	.270	.146	.160	.183	.195
Judge 4				1.000	.294	.547	.139	.169	.105	.189	.139	.154	.127
Judge 5					1.000	.645	.288	.325	.255	.341	.479	.348	.291
$Mean_1$						1.000	.315	.309	.342	.311	.380	.344	.319
Angoff 2													
Judge 1							1.000	.936	.912	.938	.889	.975	.974
Judge 2								1.000	.896	.941	.881	.969	.950
Judge 3									1.000	.884	.851	.946	.944
Judge 4										1.000	.901	.972	.958
Judge 5											1.000	.939	.916
$Mean_2$												1.000	.986
Item Diff*													1.000

*Item diff = item difficulties. These correlations do not include 3 items for which difficulties were unavailable. For all other correlations in this table, n = 200.

TABLE 4

ANOVA, Variance Components, and the Variability of Mean Scores for the Angoff #1 Method

Effect (α)	df	Sum of Squares	Mean Squares	$\hat{\sigma}^2(\alpha)$
i	199	33,297.575	167.324	22.400
r	4	6,117.300	1,529.325	7.370
ir	796	44,039.900	55.327	55.327

Estimated variances			Estimated standard deviations*			
$\hat{\sigma}^2(\bar{X}_r)$	=	7.6466	$\hat{\sigma}(\bar{X}_r)$	=	2.7653	(5.53)
$\hat{\sigma}^2(\bar{X})$	=	1.6413	$\hat{\sigma}(\bar{X})$	=	1.2811	(2.56)
$\hat{\sigma}^2(\bar{X}\|R)$	=	0.1673	$\hat{\sigma}(\bar{X}\|R)$	=	0.4091	(0.82)
$\hat{\sigma}^2(\bar{X}\|I)$	=	1.5293	$\hat{\sigma}(\bar{X}\|I)$	=	1.2367	(2.47)

NOTE: This table is modeled after Brennan and Lockwood (1980). The terms $\hat{\sigma}^2(\alpha)$ are, more specifically, $\hat{\sigma}^2(r)$, and $\hat{\sigma}^2(i)$, and $\hat{\sigma}^2(ri)$. Results in the bottom portion of this table for the variability of mean scores assume that $n_r = 5$ and $n_i = 200$.

*Results within parentheses are expressed in terms of number of items.

TABLE 5

ANOVA, Variance Components, and the Variability of Mean Scores for the Angoff #2 Method

Effect (α)	df	Sum of Squares	Mean Squares	$\hat{\sigma}^2(\alpha)$
i	199	119,121.724	598.602	117.115
r	4	5,016.374	1.254.094	6.205
ir	796	10,368.026	13.025	13.025

Estimated variances		Estimated Standard Deviations*		
$\hat{\sigma}^2(\bar{X}_r)$	= 6.2705	$\hat{\sigma}(\bar{X}_r)$	= 2.5041	(5.01)
$\hat{\sigma}^2(\bar{X})$	= 1.8397	$\hat{\sigma}(\bar{X})$	= 1.3563	(2.71)
$\hat{\sigma}^2(\bar{X}\|R)$	= 0.5986	$\hat{\sigma}(\bar{X}\|R)$	= 0.7737	(1.55)
$\hat{\sigma}^2(\bar{X}\|I)$	= 1.2541	$\hat{\sigma}(\bar{X}\|I)$	= 1.1199	(2.24)

NOTE: This table is modeled after Brennan and Lockwood (1980). The terms $\hat{\sigma}^2(\alpha)$ are, more specifically, $\hat{\sigma}^2(r)$, and $\hat{\sigma}^2(i)$, and $\hat{\sigma}^2(ri)$. Results in the bottom portion of this table for the variability of mean scores assume that $n_r = 5$ and $n_i = 200$.

*Results within parentheses are expressed in terms of number of items.

TABLE 6

Summary of Data from Beuk and Hofstee Methods

Beuk Method	Mean	Standard Deviation
1. Minimum passing score, i.e., minimum percent of items that an examinee should answer to pass the examination	78.80	3.42
2. Expected pass rate, i.e., percent of examinee who will achieve this score or higher	79.20	4.38

NOTE: For the Beuk method, the slope = 4.38/3.42 = 1.28

Hofstee Method	Mean
1. Failure rates, i.e., percent of examinees failing	
A. Lowest acceptable failure rate	18.60
B. Highest acceptable failure rate	32.80
2. Passing scores, i.e., percent of items correct	
A. Lowest acceptable passing score	69.60
B. Highest acceptable passing score	85.20

TABLE 7

Summary of Passing Scores and Pass/Fail Rates for the Angoff, Beuk and Hofstee Methods

Method	Passing Score		Estimated Percent of Examinees	
	Percent Correct	Number of Items	Passing	Failing
1. Angoff #1	75.41	151	55.1	44.9
2. Angoff #2	68.37	137	81.5	18.5
3. Beuk	71.91	144	69.1	30.9
4. Hofstee	71.95	144	69.1	30.9

NOTE: The number of items has been rounded to the nearest item.

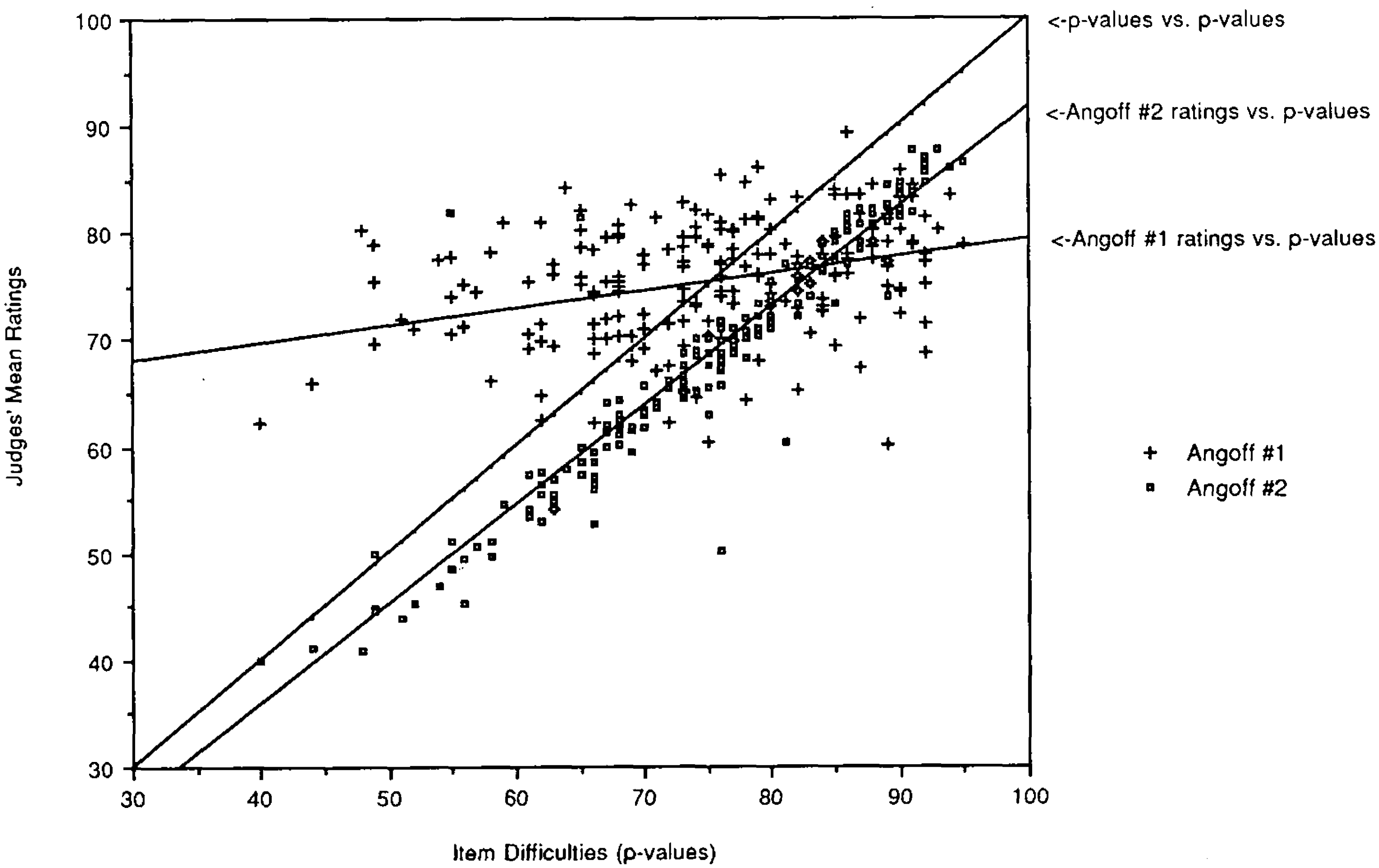

Figure 1. **Angoff #1 and Angoff #2 Ratings vs. Item Difficulties**

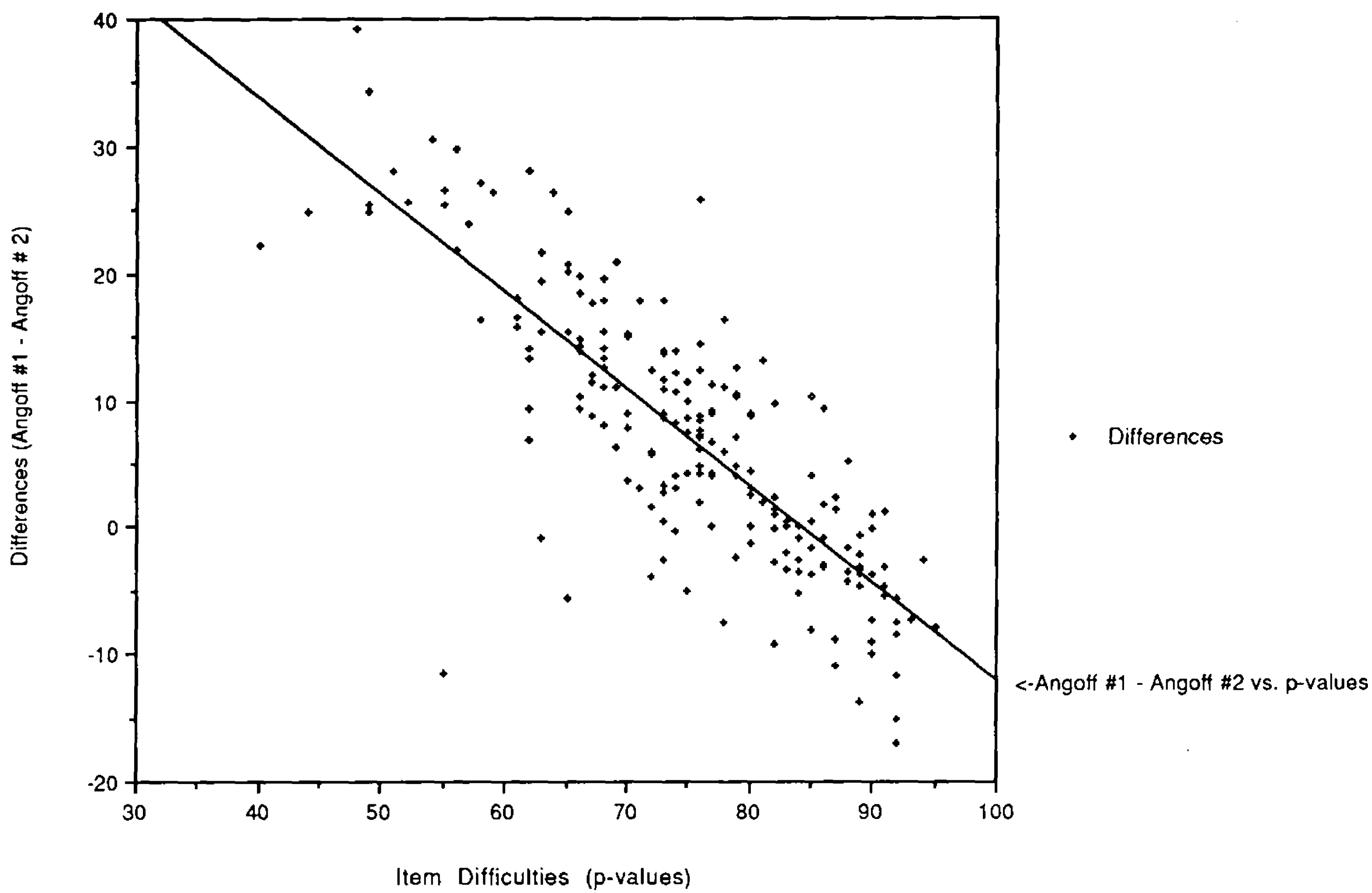

Figure 2. Differences in Mean Item Ratings vs. Item Difficulties

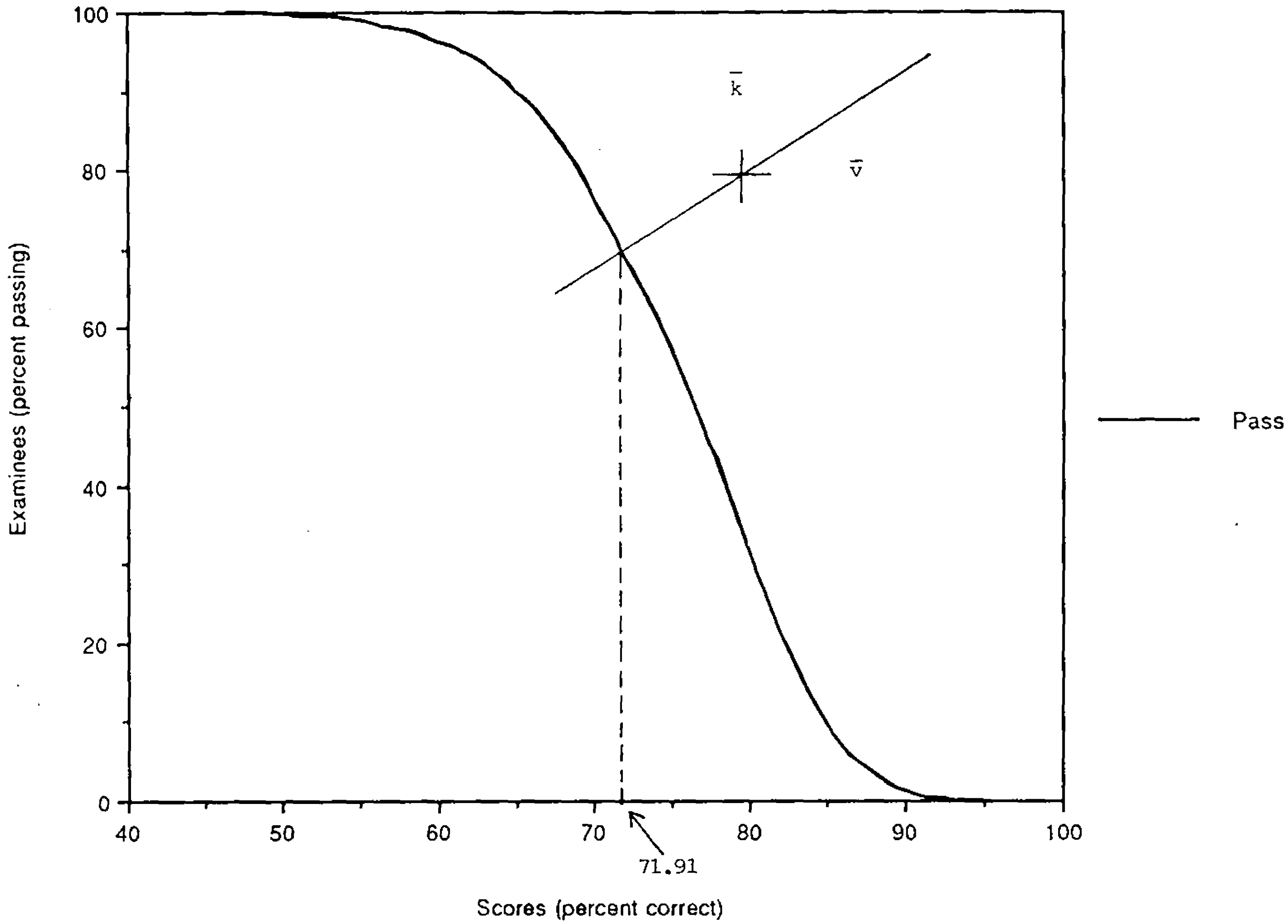

Figure 3. **Determining the Passing Score by the Beuk Method**

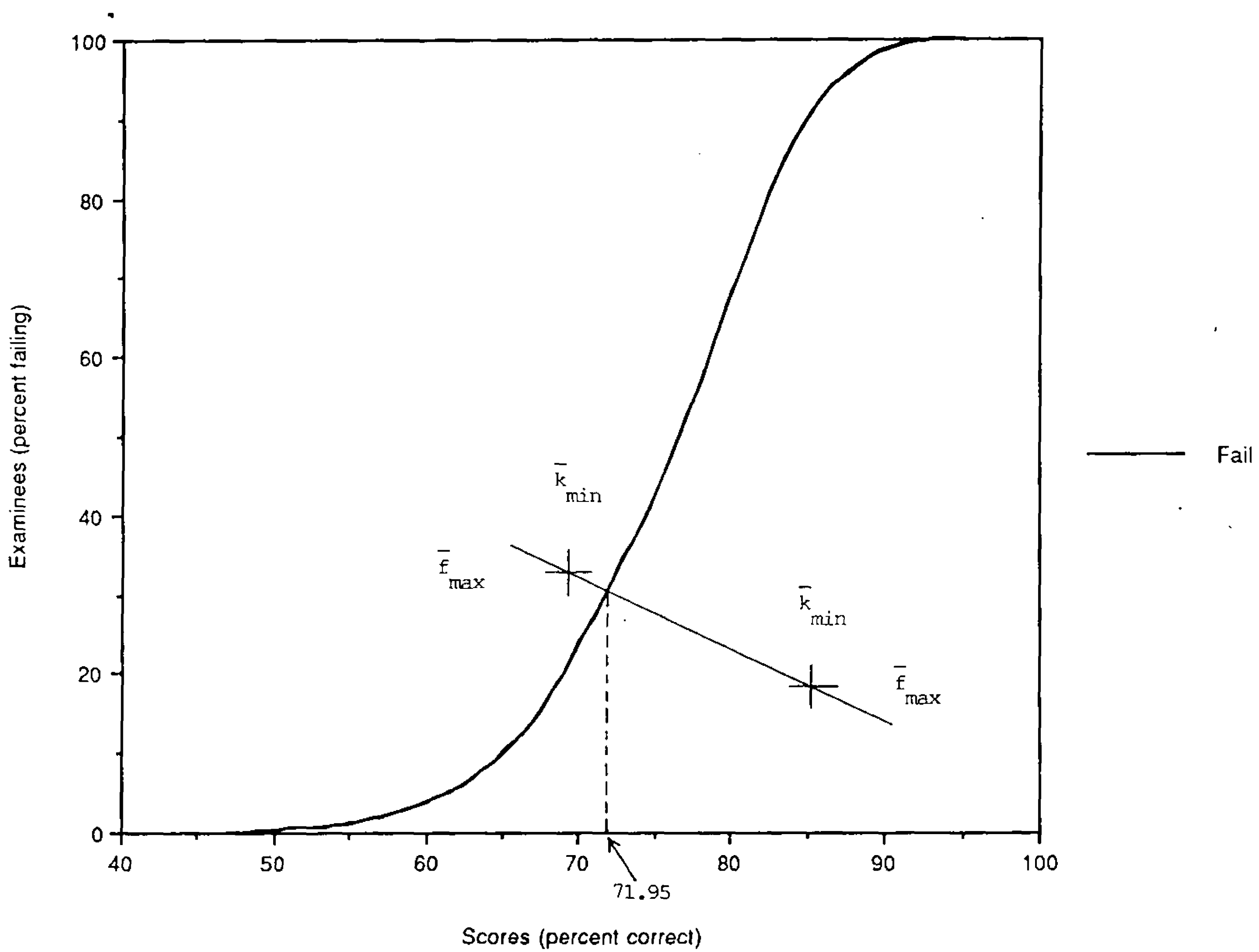

Figure 4. Determining the Passing Score by the Hofstee Method