RESEARCH ARTICLES

An Empirical Examination of the Impact of Group Discussion and Examinee Performance Information on Judgments Made in the Angoff Standard-Setting Procedure

Brian E. Clauser, Polina Harik, and Melissa J. Margolis
National Board of Medical Examiners

I. C. McManus
University College London

Jennifer Mollon, Liliana Chis, and Simon Williams
Federation of Royal Colleges of Physicians of the UK

Numerous studies have compared the Angoff standard-setting procedure to other standard-setting methods, but relatively few studies have evaluated the procedure based on internal criteria. This study uses a generalizability theory framework to evaluate the stability of the estimated cut score. To provide a measure of internal consistency, this study also compares the estimated proportion correct scores resulting from the Angoff exercise to empirical conditional proportion correct scores. In this research, judges made independent estimates of the proportion of minimally proficient candidates that would be expected to answer each item correctly; they then discussed discrepancies and revised their estimates. Discussion of discrepancies decreased the
variance components associated with the judge and judge-by-item effects, indicating increased agreement between judges, but it did not improve the correspondence between the judgments and the empirical proportion correct estimates. The judges then were given examinee performance information for a subset of the items. Subsequent ratings showed a substantial increase in correspondence with the empirical conditional proportion correct estimates. Particular attention is given to examining the discrepancy between the judgments and empirical proportion correct estimates as a function of item difficulty.

Standard setting is by its nature a policy decision; standards do not exist in nature waiting to be discovered. Although establishing examination standards is not a scientific process, in order to be defensible a standard must have procedural credibility; Kane describes procedural evidence as a primary validity criterion for standards (Kane, 1994). One important aspect of procedural evidence is a demonstration that the procedure used to establish the standard produces a result that is not unduly influenced by estimation error. When the procedure is based on expert judgments, it is reasonable to expect test developers to demonstrate that experts are effectively making the required judgments. The present article examines both types of evidence with respect to implementation of the Angoff procedure.

William Angoff first described the standard-setting procedure that bears his name in a footnote in the second edition of *Educational Measurement* (1971). He suggested a procedure in which each judge is asked to state

the probability that the “minimally acceptable person” would answer each item correctly. In effect, the judges would think of a number of minimally acceptable persons, instead of only one such person, and would estimate the proportion of minimally acceptable persons who would answer the item correctly. The sum of these probabilities, or proportions, would then represent the minimally acceptable score. (p. 515)

There have been numerous variations on this procedure. The most widely used variation makes the process iterative, with judges receiving feedback between repeated rating sessions. The feedback may be in the form of data which report how a defined group of examinees performed on the individual items, it may be in the form of impact data showing the proportion of examinees in a reference group that would be expected to fail based on the judges’ tentative standard, or it could be information about the extent of agreement between judges in their individual item ratings (Reckase, 2001).

The Angoff procedure is one of the most widely used standard-setting procedures for high-stakes standardized tests (Plake, 1998). Although there have been numerous papers comparing the Angoff procedure to other standard-setting procedures or describing variations on the Angoff procedure, there have been relatively few
papers examining its psychometric properties. (See Brandon [2004] and Hambleton and Pitoniak [2006] for a more complete review of this literature.)

Brennan (1995) has recommended generalizability theory as a framework for examining the reproducibility of cut scores. On the proportion correct scale, the standard estimated by the Angoff procedure typically is produced by averaging across both the items and the raters. The reproducibility of the resulting mean score therefore may be influenced by error attributable to the selection of specific items used in the exercise as well as the specific selection of raters. In situations where the selection of items represents a random sample from a domain of similar items and the selection of raters represents a random sample from a universe of similar raters, the reproducibility of the estimated standard will increase as the numbers of items and raters increases. Generalizability theory provides a convenient framework for examining the impact of sampling error in this context; nonetheless, there have been relatively few studies applying this approach to data from standard-setting exercises.

In addition to examining the reproducibility of the estimated standard, it is reasonable to collect evidence that allows for an evaluation of the extent to which the expert judges are actually able to make the required judgments. For credentialing examinations, standard-setting judges typically are selected because they have expertise in the content area evaluated by the assessment. Involvement in education is also common, and particularly important in the context of credentialing examinations is familiarity both with the examinee population of interest and with the requirements for professional practice (Raymond & Reid, 2001). Although this background may make the judges an appropriate group to conceptualize and define minimal competence, it does not necessarily follow that expertise in these areas will make them able to accurately judge the proportion of examinees (from this or any other group) that will answer an item correctly. Previous research has shown that experts may have a difficult time with this task in the absence of examinee performance data (e.g., Busch & Jaeger, 1990; Clauser, Swanson, & Harik, 2002; Impara & Plake, 1998).

In a study by Impara and Plake, high-school science teachers were asked to make judgments with reference to a borderline group defined in terms of the final grade received in the teacher’s class. The results did not support the argument that the teachers were able to make the required judgments. Clauser and colleagues (2002) examined results from a standard-setting study for a medical licensing examination. Their findings indicated that, in the absence of examinee performance data, there was considerable variability between groups of standard setters as well as between judges in the same group. The results also showed that the judges had a limited ability to evaluate the relative difficulty of test items. Clauser and colleagues provided correlations between the conditional probability of success on an item for a person whose total score was at the judged standard and the judged probability of success provided by the experts. The results indicated that the relationship
between the empirical item difficulties and the judged difficulties was modest in the absence of examinee performance data. That is, in the absence of examinee performance data, the judges had only a modest ability to distinguish between more and less difficult items. When performance information was provided, ratings from independent standard-setting groups converged and the relationship between empirical and judged item difficulties increased dramatically. This latter relationship has been cited as an important measure of internal consistency by previous researchers (Goodwin, 1999; Kane, 1994; Plake, Impara, & Irwin, 1999; Smith & Smith, 1988).

The present research extends the work presented by Clauser et al. (2002). In that study, judges were trained as a group and provided preliminary estimates on a sample of test items. They then were given data describing examinee performance on the items and were asked to revise their ratings as they believed appropriate. Following the revision process, the judges met as a group, compared and discussed their individual ratings, and reviewed information about the impact that a standard based on their judgments would have on a defined reference group of examinees. They then received an additional sample of items to judge and finally, after those judgments were complete, they were given performance data on those items and were permitted to revise their judgments. The previous paper provides evidence about the impact of performance data in the absence of discussion and the combined effect of discussion and examinee performance data. That paper provided no evidence about the extent to which discussion of disparities within a group of raters acts on either the reproducibility of the final standard or the relationship between the judged conditional item difficulties and those estimated from examinee responses. Brandon (2004) similarly cites several papers describing the combined effects of performance information and group discussion, but notes no study examining discussion only. Consideration of the impact of group discussion in the absence of performance data is important because standards are sometimes established for credentialing examinations before the examination is administered, and so, before performance data are available. In the exercise examined in the present study, judges first made independent judgments; they then met as a group to discuss disparities. At this point, the judges had the opportunity to revise their judgments. Finally, they received examinee performance data (on a subset of the items). This arrangement allowed for examination of the impact of group discussion in the absence of performance data. It also allowed for a comparison of these results with those produced after judges were provided with performance data.

The present article also extends previous research by examining the relationship between the conditional item difficulty and the extent to which judges over- or underestimate the probability of a correct response for the borderline candidate. Previous papers have provided evidence that there may be a relationship between item difficulty and the accuracy of the ratings. Plake, Impara, and Irwin (1999)
presented evidence that judgments were less stable across occasions for relatively easy items than for more difficult items. Impara and Plake (1998) reported that teachers underestimated actual performance for most items when instructed to think about borderline examinees as those who had received a defined grade in class. In that study, an underestimate was defined as one for which the estimated probability of success was more than .10 less than the actual performance; an overestimate was one for which the estimated value was more than .10 above the actual performance. The proportion of underestimated items was reported to be greater for difficult and moderate items and lower for relatively easy items. In contrast to these results, Goodwin (1999) reported that raters were more likely to overestimate the expected performance for borderline examinees than to underestimate performance. She also reported that easy items tended to be underestimated and difficult items to be overestimated. Shepard (1995) reported that judges underestimated the probability of success for easy items and overestimated that probability for difficult items. Given that Shepard viewed this outcome as a fatal flaw in the Angoff procedure, further evaluation of this relationship is warranted; the present article provides such an evaluation.

The purpose of the present article is to extend the existing literature by further investigating the application of empirical procedures to evaluation of the performance of expert judges participating in an Angoff exercise. The article first provides an examination of the impact of (1) group discussion and (2) examinee performance information on the judgments made by experts. Finally, the article provides a detailed analysis of the relationship between item difficulty and the magnitude of the changes experts make in their judgments after group discussion of discrepancies.

**METHOD**

The MRCP(UK) [Membership in the Royal College of Physicians of the United Kingdom] is an examination sequence for credentialing physicians in the United Kingdom. The Part 1 Written Examination is comprised of five option, single-best-answer multiple-choice items and is the first component of the sequence; the data examined in this article resulted from two operational standard-setting exercises conducted for this examination in 2005. The judges reviewed items from different examination administrations during each exercise. During the two standard-setting exercises, judges reviewed and provided judgments for 200 and 195 items, respectively.

A subset of the items from each administration had been used in a previous administration. Data from these items (34 from the first administration and 43 from the second) were analyzed to produce item performance information that judges reviewed during the standard-setting process. This information included probabilities of success on the item for examinees in each of five score groups:
(1) a marginal group representing the ten percent of examinees above and below the cut score for the examination in which the item was previously used; (2) a clear pass group and a clear fail group representing the top and bottom ten percent of examinees; and (3) two intermediary groups. (The cut score is close enough to the center of the score distribution to make these category definitions workable.) Judges also were provided with data on the proportion of examinees in each of these five groups selecting each of the five options. It is important to note that the items for which performance information was available did not represent a random sample of the items on the test form; items with relatively low point-biserial correlations were excluded from re-use, as were items with very high or low $p$-values.

Six judges participated in each standard-setting exercise. Judges were practicing physicians with considerable involvement in the MRCP(UK) examination process. Because of this experience, they had a thorough understanding of the purpose of the examination and knowledge of the demands and expectations for practice for examinees completing the examination sequence. In addition, they all had prior experience with the Angoff procedure used in this exercise. Because of this significant background and experience, there was no additional formal training prior to beginning the exercise.

The judges first made independent decisions about each item and then met as a group to compare individual judgments with those of the other judges. A computer-display system showed the independent judgments of each judge. For each item, the individuals responsible for the most extreme judgments presented their opinions about the item; this was followed by a general discussion. At this point, the judges had the opportunity to modify their original judgments. Judges then were provided with the item performance information and again were permitted to modify their judgments. This procedure resulted in two sets of judgments (pre- and post-discussion) for each of the items and three sets of judgments for a subset of the items (pre- and post-discussion and post-provision of performance data).

Consistent with the previous discussion, analyses were implemented to examine two aspects of the judges’ performance: (1) generalizability analyses examined the reproducibility of the estimated standard; and (2) comparisons were made between the conditional item difficulties associated with the judgments and the empirical item difficulties estimated from examinee responses.

Generalizability Analysis

A complete description of the generalizability theory framework is beyond the scope of this article; the interested reader is referred to the comprehensive descriptions provided by Cronbach, Gleser, Nanda, and Rajaratnam (1972) and Brennan (2001a) for a detailed discussion of the theory and to Brennan (1995) for a discussion of the application of generalizability theory to standard setting. In generalizability
theory, analysis of variance procedures are used to estimate variance components associated with facets of measurement within the data collection design. In the present study, there are two facets in the design: one represents judges and the other represents items. The variance component for judges represents the variability across the mean ratings provided by the judges; the variance component for items represents the variability across the mean rating provided for each of the items. In addition to these two effects, there is a variance component for the judge-by-item interaction. This component represents the judge-by-item interaction as well as the residual variance reflecting sources of variance not accounted for in the design.

The impact of these potential sources of measurement error on the reproducibility of the estimated standard will be a function both of the size of the variance component and the number of times that the facet is sampled in estimating the standard; as the number of judges and/or items used in the estimate is increased, the precision of the estimate increases (i.e., the standard error decreases). Generalizability theory provides a means of examining how precision changes as a function of this sampling. The present application is, however, somewhat different than that which is typically described for examining the precision of test scores. In most applications, one of the facets of the design can be considered to be the object of measurement: most often this is the examinee. The remaining facets of the design (e.g., items) represent potential sources of measurement error. In the present application, both the item and rater facets are potential sources of measurement error. The analysis provides evidence about how the precision of estimation would vary as a function of the number of raters and items used in the standard-setting exercise. As with other applications of generalizability theory, the variance components that are viewed as contributing to error will be a function of the intended interpretation. The mGENOVA software (Brennan, 2001b) was used to complete parallel analyses for each of the five sets of judgments described previously.

Comparison of Judged Probabilities and Empirical Conditional Probabilities

When a rater provides a judgment as part of an Angoff standard-setting exercise, a conditional mean score for the item is being estimated (i.e., a mean score for examinees whose level of proficiency is at the cut score). To be logically consistent, these judgments must be related to the difficulties produced when examinee responses are analyzed. Given that judgments are intended to reflect how the minimally competent examinee would perform, it would be difficult to argue for the appropriateness of the procedure if judges concluded that minimally competent examinees would do relatively well on items that examinees found to be difficult and less well on items that examinees found to be relatively easy.

This criterion of correspondence between judged and empirical probabilities has been widely recommended as an appropriate evaluation of the internal consistency
of the procedure (Clauser, et al., 2002; Goodwin, 1999; Kane, 1994; Plake, Impara, & Irwin, 1999; Smith & Smith, 1988). To appropriately examine this aspect of the procedure, it is necessary to estimate the conditional proportion correct scores for the empirical item responses. The paper by Smith and Smith (1988) cites a personal communication with Angoff in which he recommends estimating the \( p \)-value achieved by examinees close to the cut score when making such an evaluation rather than using the \( p \)-value for the full group. Although this approach is superior to making comparisons based on \( p \)-values estimated from the full sample, basing the estimation of the conditional probability of success for an item on the actual responses of examinees near the cut score may be limited in two ways. First, if the group of examinees completing the test is relatively small, the sub-sample with scores close to the cut score may be so small that it produces an unstable estimate. Second, a problem exists in defining what is meant by “close to the cut score”; any definition will be somewhat arbitrary and as the range increases it becomes a less appropriate approximation of the probability at the cut score.

Item response theory models provide an alternative approach to estimating this conditional probability. These models represent a theoretical framework for estimating the expected probability of success on an item for an examinee with a specified level of proficiency (Hambleton, Swaminathan, & Rogers, 1991). In the present study, the estimates were made using the one-parameter logistic model commonly known as the Rasch model. With this model, the expected probability of a correct response for examinees at a given proficiency level (\( \theta \)) is represented as:

\[
P(\theta) = \frac{e^{(\theta-b)}}{1 + e^{(\theta-b)}}
\]

where \( \theta \) represents examinee proficiency and \( b \) represents item difficulty. Estimates were produced using the WINSTEPS software (Wright & Linacre, 1999).

In this model, for a fixed set of items each number correct score maps to a unique level of \( \theta \). To produce the empirical conditional probability of success on an item, the number correct score associated with the cut score for the group was established. The examinee proficiency level associated with this number correct score then was identified, and using the above equation and the estimated item difficulties, the conditional probability of success on each item was estimated. This process was repeated for each of the five sets of judgments (i.e., pre- and post-discussion for the full sample of items and pre- and post-discussion and post-provision of performance data for the subsample of items). Scatter plots and correlations representing the strength of relationship between judged and empirical conditional item difficulties then were produced.

To further examine the relationship between the predicted probabilities and the empirical probabilities, differences between the empirical and predicted probabilities
were calculated and plotted as a function of the empirical conditional $p$-value. These residual plots allowed for examination of the extent to which accuracy of the judgments varies as a function of item difficulty.

RESULTS

Descriptives

The purpose of this article was to provide a detailed examination of the effects of feedback on judges participating in an Angoff exercise. Understanding how judges use this type of information is important in structuring an effective exercise. Before moving to the more detailed analysis, it is useful to consider whether the feedback had an impact on the estimated cut scores.

For the primary data set (that comprising 200 items), the discussion of discrepancies changed the cut score on the percent correct scale from 62% to 60%. For the second data set, discussion changed the cut score from 62% to 61%. These differences in the cut score would have increased the percent of examinees passing by 3.7% and 3.8% for the primary and secondary data sets, respectively. Although the changes are not large, clearly they are nontrivial differences.

Generalizability Analysis

Table 1 presents the variance components based on analysis of the pre- and post-discussion judgments of the sample of 200 items. These variance components represent the sources of variability in the judges’ estimates of the conditional probability of success for a minimally proficient examinee. The results indicate that the discussion has a substantial impact on the variance components. Discussion results in a reduction in the difference between mean ratings across judges and a reduction in the judge-by-item interaction; at the same time, it produced an increase in the item variability.

<table>
<thead>
<tr>
<th>Effect</th>
<th>Pre-Discussion</th>
<th>Post-Discussion</th>
</tr>
</thead>
<tbody>
<tr>
<td>Judge</td>
<td>0.00038</td>
<td>0.00010</td>
</tr>
<tr>
<td>Item</td>
<td>0.00507</td>
<td>0.01036</td>
</tr>
<tr>
<td>Item $\times$ Judge</td>
<td>0.00713</td>
<td>0.00306</td>
</tr>
</tbody>
</table>
The following formula provides a practical basis for interpreting the variance components in Table 1:

\[
\sigma^2_e = \frac{2}{n_j} \frac{\sigma^2_j}{\sigma^2_{ji} + \frac{n_j}{n_{ji}}}
\]

In this formula, the error variance is a function of the sum of the variance component for judges divided by the number of judges and the residual term (or judge-by-item variance component) divided by the total number of observations.

This formula is appropriate for conditions in which the item variance would not contribute to error in estimating the cut score. This model is appropriate under the circumstances in which the standard is established for application to a single form of the test and is based on judgments of all items on that form. (This is the case for the examination under study.) With this design, items may be considered a fixed facet in the analysis and the item variance component would not contribute to error. When statistical adjustments are made to place the standard on the common score scale a standard established based on a single test form can be applied across test forms. The results produced by this formula would be appropriate for this condition as well, although these results will be overly optimistic to the extent that the statistical adjustments are less than perfect. The general conditions in which this formula is appropriate are common, although other designs are possible. For example, if the standard is being established for a domain, the precision of the estimated standard will depend on the sampling of items as well as judges.

It is reasonable to use the variance components from the pre-discussion judgments to estimate the expected measurement error associated with the standard-setting process. This interpretation is based on the assumption that a replication of the procedure would be based on random selection of another set of judges (and possibly items) from a defined universe of acceptable judges (and items). It is not sensible to make the same assumption about the post-discussion judgments. The reduction in the judge and judge-by-item variance components suggests that the discussion causes the judges within the group to revise their ratings so that they become more similar. This reduction does not provide evidence that within-group discussion would result in convergence between independent groups of standard setters. Clauser and colleagues (2002) reported a convergence of between-group estimates after group members reviewed performance information. The present data do not provide a similar replication across groups examining the same item sets and so the same comparison cannot be made.

Table 2 presents the variance components for the sub-sample of items for which performance data were available. Although the specific values differ from those reported in Table 1, the pattern of change is similar from the pre-discussion
to the post-discussion results. These items were not a random sample from the complete test form so it is not unexpected that the specific components differ from Table 1 to Table 2. Providing the performance data appears to have had little impact on the variance component for judges, although it increased the variance component for items and decreased the component for the judge-by-item interaction.

Tables 3 and 4 provide results comparable to those in Tables 1 and 2 for the second data set. Although the results differ in specifics, the pattern of results leads to the same general conclusions. In Table 4, the estimated judge variance component of 0.00000 most likely results from estimation error. It should be remembered that the item samples are small for the results in Tables 2 and 4, and this leads to unstable estimates for all of the components. Negative estimates are not uncommon and are by convention set to zero.

### Table 2
Variance Components Based on Analysis of the Pre- and Post-Discussion and Post-Provision of Performance Data Judgments for 34 Items

<table>
<thead>
<tr>
<th>Effect</th>
<th>Pre-Discussion</th>
<th>Post-Discussion</th>
<th>Post-Performance Data</th>
</tr>
</thead>
<tbody>
<tr>
<td>Judge</td>
<td>0.00012</td>
<td>0.00000</td>
<td>0.00002</td>
</tr>
<tr>
<td>Item</td>
<td>0.00341</td>
<td>0.00697</td>
<td>0.00825</td>
</tr>
<tr>
<td>Item × Judge</td>
<td>0.00557</td>
<td>0.00297</td>
<td>0.00200</td>
</tr>
</tbody>
</table>

### Table 3
Variance Components Based on Analysis of the Pre- and Post-Discussion Judgments for 195 Items

<table>
<thead>
<tr>
<th>Effect</th>
<th>Pre-Discussion</th>
<th>Post-Discussion</th>
</tr>
</thead>
<tbody>
<tr>
<td>Judge</td>
<td>0.00049</td>
<td>0.00027</td>
</tr>
<tr>
<td>Item</td>
<td>0.00564</td>
<td>0.01115</td>
</tr>
<tr>
<td>Item × Judge</td>
<td>0.00799</td>
<td>0.00351</td>
</tr>
</tbody>
</table>

### Table 4
Variance Components Based on Analysis of the Pre- and Post-Discussion and Post-Provision of Performance Data Judgments for 43 Items

<table>
<thead>
<tr>
<th>Effect</th>
<th>Pre-Discussion</th>
<th>Post-Discussion</th>
<th>Post-Performance Data</th>
</tr>
</thead>
<tbody>
<tr>
<td>Judge</td>
<td>0.00000</td>
<td>0.00008</td>
<td>0.00017</td>
</tr>
<tr>
<td>Item</td>
<td>0.00371</td>
<td>0.00648</td>
<td>0.01083</td>
</tr>
<tr>
<td>Item × Judge</td>
<td>0.00529</td>
<td>0.00293</td>
<td>0.00192</td>
</tr>
</tbody>
</table>
Comparison of Judged Probabilities and Empirical Conditional Probabilities

Figures 1 and 2 present scatter plots of the mean ratings and the empirical conditional probabilities for the pre- and post-discussion judgments for items in the primary data set. The correlations (approximately .58) are essentially unchanged by discussion. The variability of the mean ratings has increased substantially, changing the appearance of the scatter plot, but the strength of the relationship is unchanged. Table 1 indicates that after discussion the spread of the group ratings increases and the level of agreement between judges increases; Figures 1 and 2 suggest that this increase is not associated with an increase in the relationship between the judgments and the empirical conditional item difficulties.

Figures 3 through 5 are scatter plots for the subset of 34 items for which performance data were available. Again, the discussion did not produce an increase in the correlation between the judgments and the empirical item difficulty; the estimated

![Figure 1](image-url)
correlation actually dropped modestly (.40 to .34). Providing performance data substantially increased this correlation (.66).

Examination of the scatter plots in Figures 1 and 2 makes it clear that the distribution of predicted probabilities is narrower than the distribution of empirical conditional probabilities. Figure 6 presents a plot of the difference between the predicted and empirical probabilities for the post-discussion judgments as a function of the empirical conditional probability; Figure 7 provides an analogous plot of the absolute values of the differences. In both plots, the values are closest to zero at the cut score. Figure 6 makes it clear that judges systematically overestimate probabilities for values below the cut score and underestimate them for values above the cut score. The asymmetrical shape of the plot in Figure 7 may seem to suggest that as the empirical probability moves away from the cut score, the accuracy of the judgments decreases more rapidly for relatively easy items than for relatively difficult items. However, interpretations based on proportion correct scales may be problematic. As an additional check on this relationship, the data were transformed from probabilities to the logit scale (ln(p/1-p)). The intention
of this transformation was to make the data more closely approximate an interval scale; this minimizes the possible ceiling effects that exist for probabilities at the upper end of the scale. The result is displayed in Figure 8. The asymmetry now has largely vanished.

Having examined the discrepancy between the predicted and empirical probabilities in some detail, it is worth noting that the mean of the absolute value of the discrepancies was essentially unchanged from the pre- to the post-discussion ratings (.129 for the pre-discussion values and .127 for the post-discussion ratings). When this value was examined for the relatively difficult items (those with a conditional p-value greater than the cut score), the mean was similarly unchanged (.145 for the pre-discussion values and .140 for the post-discussion ratings).

Scatter plots analogous to those presented in Figures 1 through 5 were produced based on the second data set. In general, these showed the same pattern of results as those for the primary data set. (Copies are available from the first author.)
This study was designed to examine two types of evidence about the performance of the Angoff procedure: the reproducibility of the estimated standard and the relationship between the group judgments and empirical item difficulties. The data used in this study came from an exercise in which experts independently made judgments and then modified those judgments after group discussion. Previous research has shown that both the generalizability of the results and the correlation with empirical conditional item difficulties increase after judges review performance data (Busch & Jaeger, 1990; Clauser et al., 2002). The present research examined the effect of group discussion in the absence of performance data. Results indicate that discussion substantially reduced discrepancies between judgments and increased the variability of group judgments across items, but it did not increase the strength of the relationship between the judgments and the empirical conditional item difficulties.
Examinee performance data were provided to the judges for a sub-sample of items. For these items, the pattern of results for the pre- and post-discussion judgments was similar to that for the larger data set. (The pattern is similar but the actual correlations are lower because these items were sampled to have less variability in item difficulty; extremely difficult or easy items were not reused and so statistical information was available for a more restricted sub-sample.) After performance data were provided, the relationship between the judgments and the empirical conditional item difficulties was substantially increased. These results argue against the utility of the group discussion in the absence of examinee performance data. They also are consistent with previous reports which suggest that, in the absence of performance data, content experts have a limited ability to make judgments about the relative difficulty of test items.

Because of the limited ability of content experts to accurately make the required judgments in the absence of performance data, previous authors have argued against the use of the Angoff procedure when such data are unavailable (e.g., Clauser et al., 2002; Impara & Plake, 1998). The present results are consistent with those presented.

FIGURE 5 The relationship between the empirical conditional probability of success on an item and the judged probability of success for the sub-sample of items after provision of performance data.
previously, and the results of this study go beyond previous research in showing that group discussion did nothing to improve the relationship between the judgments and empirical values.

The present study provided a more detailed examination of the relationship between judge accuracy and item difficulty than that available from previous literature. Specifically, the evidence indicates that both before and after discussion judges have difficulty with the required judgment across the continuum of item difficulty. Judges appear to systematically overestimate the probability of success on difficult items and underestimate the probability of success on easy items. In the present study, when the results are presented on the proportion correct scale (Figure 7) it appears that judges have a greater tendency to overestimate performance on difficult items than to underestimate that of easy items. Although this might be viewed as supportive of Goodwin’s (1999) conclusions, when the scale is transformed this apparent asymmetry is eliminated; this may be more consistent with Shepard’s conclusions (1995). The fact that rescaling results in manipulating this apparent effect suggests that it may vary as a function of the placement of the cut score on the proportion correct scale. Additional study is warranted in this area; it may be that Goodwin’s reported effect would have vanished if a similar transformation had been used.

The results reported in the present study are generally consistent with those reported by both Goodwin and Shepard, but they are at odds with those reported
by Impara and Plake (1998). Impara and Plake reported that: (1) judges were more likely to underestimate performance for the borderline group than to overestimate it; and (2) the proportion of items being underestimated was greater for relatively difficult items than for easy items. This may result from the fact that the definition of the borderline group was statistical in the present study and in the studies by Goodwin and Shepard. By contrast, Impara and Plake defined the borderline group in terms of the grade they received in science class.

Of course, interpretation of these results requires that the model used to estimate the conditional probabilities reasonably fits the data. Analysis of fit for the data (not presented) provided general support for the model. The one-to-one mapping between number correct scores and scaled scores (implicit in the Rasch model) is also consistent with the way the test is scored.

The results of this study contribute to the body of evidence about the limitations of the Angoff procedure when performance data are unavailable. These results provided no evidence that the discussion of discrepancies results in an improvement in the ability of judges to correctly identify conditional item difficulties. The correspondence between empirical item difficulties and judged difficulties is substantially increased after judges reviewed examinee performance data.
The limited usefulness of the Angoff procedure in the absence of examinee performance data may appear to call into question the extent to which the procedure provides a content-based or test-centered as opposed to a norm-referenced standard. However, to the extent that the judges have been selected because of their knowledge of both test content and the professional performance of the population of examinees, this feedback may allow judges to make a more fully informed and integrated decision. The provision of performance data in no way prompts judges to make a specific decision; it simply provides evidence about the extent to which their expectations for a group of conceptualized minimally proficient examinees is consistent with the performance of examinees from a group with which they are familiar (Clauser et al., 2002). That is to say, the provision of the normative data does not tell the judges whether 5%, 50%, or 95% of examinees should fail the test. It simply affords judges the opportunity to integrate their expectations for the group with their expectations based on the content of each item.

The generalizability of the pattern of results discussed in this article is supported by the results of a replication based on an independent standard-setting exercise for a different administration of the same examination. Although this article contributes to the literature on the performance of the Angoff procedure, this literature remains less than definitive. The procedure remains among the most widely
used standard-setting procedure for tests based on multiple-choice items, yet it also remains seriously limited by the fact that the required judgments are so difficult for subject-matter experts. At present, clear and definitive guidance on how and when to implement this procedure is unavailable. The results reported in this article make clear the limitations of implementing the procedure when performance data are unavailable. There do, however, remain instances in which it is necessary to establish standards in the absence of examinee performance data. In spite of its limitations, it may be that some variation on the Angoff procedure remains the method of choice.

**REFERENCES**


