The Effect of Rating Unfamiliar Items on Angoff Passing Scores

Educational and Psychological
Measurement

2017, Vol. 77(6) 901–916

© The Author(s) 2016

Reprints and permissions:

sagepub.com/journalsPermissions.nav

DOI: 10.1177/0013164416670983

journals.sagepub.com/home/epm

**Jerome C. Clauser¹, Ronald K. Hambleton²,
and Peter Baldwin³**

Abstract

The Angoff standard setting method relies on content experts to review exam items and make judgments about the performance of the minimally proficient examinee. Unfortunately, at times content experts may have gaps in their understanding of specific exam content. These gaps are particularly likely to occur when the content domain is broad and/or highly technical, or when non-expert stakeholders are included in a standard setting panel (e.g., parents, administrators, or union representatives). When judges lack expertise regarding specific exam content, the ratings associated with those items may be biased. This study attempts to illustrate the impact of rating unfamiliar items on Angoff passing scores. The study presents a comparison of Angoff ratings for typical items with those identified by judges as containing unfamiliar content. The results indicate that judges tend to perceive unfamiliar items as being artificially difficult resulting in systematically lower Angoff ratings. The results suggest that when judges are forced to rate unfamiliar items, the validity of the resulting classification decision may be jeopardized.

Keywords

standard setting, Angoff method, item response theory

¹American Board of Internal Medicine, Philadelphia, PA, USA

²University of Massachusetts Amherst, Amherst, MA, USA

³National Board of Medical Examiners, Philadelphia, PA, USA

Corresponding Author:

Jerome C. Clauser, American Board of Internal Medicine, 510 Walnut Street, Suite 1700, Philadelphia, PA 19106, USA.

Email: jclauser@abim.org

When examinations are used to classify examinees, the validity of the inferences based on these classifications depends on the properties of both the scores and the standards (e.g., passing scores, cut-scores; Kane, 2001). Typically, decisions about the requisite performance standards are informed by the opinions of invited panelists who participate in various standard setting exercises. Often panelists are recruited because of exemplary performance in their respective field and when this is so, it may be reasonable to assume they possess sufficient mastery of the content domain to be considered *content experts*. This content expertise is critical to the integrity of the standard setting process as it is widely considered a prerequisite for panelists to effectively review exam content and make associated judgments (Loomis, 2012; Raymond & Reid, 2001). Having selected these content experts, it is common to assume that they possess content expertise across the breadth of the tested material. For example, if panelists were selected for a middle school math assessment from among a pool of exemplary middle school math instructors, this assumption is likely to be satisfied.

In other cases, panelists may be selected based on traits less relevant than their content expertise. For example, testing programs like National Assessment of Educational Progress (NAEP) require that members of the public be included in the standard setting panels (National Research Council, 1999). These judges may provide a valuable policy perspective, but they cannot necessarily be assumed to possess mastery of all exam content. In still other cases, even the most exemplary practitioners may have deficiencies in their content expertise. For an exam assessing a broad or highly technical domain of content, gaps in the judges' content knowledge may be unavoidable. This is particularly common for credentialing exams where specialization among practitioners can result in very few experts with mastery of the complete content domain. For example, even an excellent cardiologist may not have expertise in the complete domain of medicine covered in a physician licensure exam that includes items on gastroenterology, obstetrics, psychiatry, and so on. This misalignment between exam content and judge expertise may introduce error into the standard setting process. When these errors affect examinee classification decisions, such as pass/fail status, the validity of inferences based on these classifications is jeopardized.

The most widely used and well-researched standard setting procedures employed on credentialing exams are based on the method originally described by Angoff in 1971 (Angoff, 1971; Meara, Hambleton, & Sireci, 2001). In the Angoff method, content experts begin by identifying the characteristics of the minimally proficient member of a given performance category. This hypothetical examinee is one whose knowledge, skills, and abilities are just barely sufficient to qualify for membership in a given performance category. The role of the subject matter expert is then to estimate the probability that this minimally proficient examinee would respond correctly to various test questions. Experts rely on their knowledge of the content domain vis-à-vis the hypothesized minimally proficient candidate to inform these estimates (Livingston & Zieky, 1982). The resultant probabilities are later summed across

items and averaged across judges within a given panel to arrive at a panel-specific recommended passing score expressed on the raw score scale. Although a variety of Angoff modifications are used in practice (Plake, Melican, & Mills, 1991), this straightforward procedure is always the core of the Angoff method.

One limitation of this common implementation of the Angoff method is that judges must rate a common set of items without omission because panel-level passing scores are calculated on the raw score scale. This requirement may be avoided using item response theory (IRT; as we explain below) but more often raw scores are averaged across judges before being put on the IRT proficiency scale (Davey, Fan, & Reckase, 1996; Kane, 1987; Reckase, 1998). For this reason, judges are generally discouraged from skipping items. Unfortunately, when judges lack the necessary expertise to make reasonable judgments for all items, this requirement will force them to provide ratings for unfamiliar content. Some implementations of the Angoff method allow judges to identify unfamiliar test items that can be omitted from the calculation of the recommended passing score. More commonly, however, the potential influence of unfamiliar test content is viewed as unavoidable and standard setting facilitators dutifully request that judges provide a rating for each item.

Requiring subject matter experts to rate all items may pose no problem when their expertise extends to the complete content domain; however, when judges lack expertise for aspects of the tested material, it is difficult to argue that those judgments describe their conceptualization of the capabilities of the minimally proficient examinee. When judgments do suffer, the effect on the final passing score still may be inconsequential if the errors are random; however, the literature provides no evidence to support the assumption of random errors in this case. Moreover, when presented with content outside their expertise, we might expect judges to confound *their* unfamiliarity with the test material with the difficulty of the test material for examinees. Indeed, as we shall see, there appears to be a tendency among judges to perceive unfamiliar content as overly difficult. Unlike random error, systematic errors of this kind do *not* cancel one another out—they aggregate. Therefore, the purpose of this study is to consider the consequences of requiring judges to rate items they believe fall outside of their area of expertise and to suggest a strategy for excluding their associated ratings.

Literature Review

Several authors have examined the effect of panelists' expertise on the results of standard setting exercises. Research in this area has tended to focus on comparisons of aggregate passing scores for different groups of judges. Early work on this topic found that the impact of judge expertise on the recommended standard to be extremely modest (Busch & Jaeger, 1990; Norcini, Shea, & Kanya, 1988; Plake, Impara, & Potenza, 1994). For example, on a medical certification exam, Norcini et al. (1988) compared Angoff-based standards from two groups of panelists with two different medical certifications. Results showed that although the panels differed in

their recommended standard, there were no statistically significant interactions between judge certification and item content. This finding was replicated by Busch and Jaeger (1990), who analyzed standard setting data from a teacher certification exam. Passing standards were determined by two panels—one composed of public school teachers and the other of university professors—and then compared for seven different content areas. In all but one content area, these researchers failed to find statistically significant differences in passing scores. Similarly, Plake et al. (1994) compared the standards associated with panelists drawn from four different K-12 subject areas. Once again, only minor variations in the passing scores set by in-content and out-of-content experts were observed. Moreover, for the application they were investigating, the authors found that the observed differences were unimportant.

In contrast, more recent research suggests that occasions do arise when variations in panelists' expertise can have nontrivial effects on the resultant standards. Using data from a medical certification exam, Verhoeven, Verwijnen, Muijtjens, Scherpbier, and van der Vleuten (2002) compared Angoff-based standards across two panels: one comprising item writers and one comprising recent graduates from a medical training program. The researchers found that recent graduates provided significantly lower Angoff ratings on average than item writers: the former passing 92.8% of examinees and the latter less than half that number, 43.5%. Consistent with these findings, Wayne, Cohen, Makoul, and McGaghie (2008) found that differences in recommended Angoff-based standards for an assessment of physician communication skills differed widely across five panels when panel assignment was based on panelists' content expertise. In the most extreme case, the passing rate differed between panels by 27%.

Although these results indicate that differences may exist between groups of content experts, they provide relatively little insight as to whether those differences are due to difference in perceived item difficulty or differences in judges' conceptualization of the minimally proficient examinee. Saunders, Ryan, and Huynh (1981) examined the extent to which judges' content mastery affects their ratings in predictable ways. The authors observed a correlation of 0.30 between judges' performance on the exam and their recommended passing score. Although this relationship accounted for only 9% of the observed variation in passing scores across judges, it suggests that, in part, panelists confounded their unfamiliarity with a given item's content with its difficulty.

The studies described so far only investigated the impact of judges' expertise on panel-specific standards rather than focusing on the interactions of individual judges with specific items. These interactions were the topic of a study by Chang, Dziuban, Haynes, and Olson (1996) that compared the ratings for items a given judge answered correctly to the ratings for items he or she answered incorrectly. Knowledge of the correct answer may not be a *sufficient* condition to ensure a reasonable rating, but a persuasive case can be made that it is a *necessary* one. It is therefore perhaps unsurprising that the authors found statistically significant differences between the ratings provided for items the judges answered correctly and those they answered incorrectly.

While these authors conclude that an absence of expertise can artifactually affect Angoff judgments, their study design fails to adequately control for differences in item difficulty between those items that were correctly answered and those that were incorrectly answered. For example, the result observed by Chang et al. (1996) would arise when a judge struggles to correctly answer difficult items, and then provides those items with an appropriately low rating. In principle, controlling for item difficulty could allow the impact of expertise to be examined separately and the authors attempted to do this by comparing judges' ratings to item p -values; however, their p -values were based solely on the responses of the 17 panelists. Thus, it is difficult to know for certain whether judges provide low ratings for unfamiliar items or if unfamiliar items are simply more difficult.

More recently, Verheggen, Muijtjens, Van Os, and Schuwirth (2008) examined the effect of judge expertise on Angoff ratings. The study examined the results from a psychiatry exam with Angoff ratings provided by thirteen content experts. As with Chang et al. (1996), the judges were asked to answer items as they provided their ratings. These items contained between 1 and 3 questions that were scored independently. The authors used linear regression to predict Angoff ratings based on item level scores. These researchers found that a judge's performance on an item was a significant predictor of their Angoff rating. Although these results suggest that judge expertise influences judges' ratings it is difficult to interpret the magnitude of this effect in the context of these polytomous items.

As with some of the earlier literature on this topic, in this study we also investigated the extent to which ratings provided for unfamiliar items are systematically different from ratings associated with more familiar content. We sought to extend the current literature by fully controlling for item difficulty when evaluating judges' ratings. Furthermore, we attempted to quantify the magnitude of the errors that arose in the absence of expertise and consider the impact of such errors on standards and classification rates. A strategy for excluding ratings of unfamiliar content is offered. Finally, to increase the generalizability of our findings, we analyzed ratings data from two different standard setting exercises, each containing three panels of judges.

Method

Data Source

Data used in this study were collected during an operational standard setting exercise for Step 1 of the United States Medical Licensing Examination (USMLE). Successful completion of the USMLE is required for all physicians with an MD degree seeking a license to practice medicine in the United States. Step 1, which is part one of a three-part licensing sequence, measures mastery of biomedical sciences using dichotomously scored multiple-choice items. The data used in these analyses were collected in 2006 and 2013 and each year included three standard setting panels. Standard setting judges included practicing physicians as well as nonphysician PhDs working in medical education (e.g., professors of microbiology). The composition of each panel

was designed to be comparable and efforts were made to recruit panelists from across the United States and to balance panels with respect to physician specialty and practice setting. All standard setting activities were conducted by the same facilitator using identical procedures across panels and datasets.

Data Collection

At the start of each standard setting exercise, facilitators explained the purpose of the exam and the standard setting activities, introduced the Angoff methodology, and led a discussion on the characteristics of the minimally proficient examinee. This discussion was followed by a group activity in which panelists rated a practice set of 15 items. Ratings were made independently but then shared with the group. Discrepancies among panelists' ratings were then discussed with the goal of refining the concept of the minimally proficient examinee. This was followed by an additional training set of 75 test items. Judges were provided with the test items and the associated keys and asked to rate these items independently on a 0 to 100 scale. This process mirrored the operational procedures with respect to rating, reviewing, and revising. These 75 items were common across the three standard setting panels within a given year and were the focus of our analyses.

One benefit of using USMLE datasets is that while judges were asked to rate all items, they were encouraged to mark items they felt were outside their scope of expertise. The facilitator explained that because of specialization, not all panelists are expected to be comfortable with the full breadth of the content domain. Items outside the judge's area of expertise for which they "had no frame of reference for making a judgment" were described as "unfamiliar" (M. Margolis, personal communication, March 3, 2013). While judges were instructed to provide ratings for such items, they were also told to check a box on their score sheet to indicate that they were unfamiliar with the tested content. Judges were informed that these items would not be included in the calculation of the recommended passing score. This design made it possible to compare ratings for familiar items with ratings for unfamiliar items. Table 1 presents the number of judges in each panel along with the number of items marked as unfamiliar.

Analytic Framework

Although, the Angoff method was developed for use with number correct scoring, analysis of Angoff ratings is frequently performed within an item response theory framework (e.g., B. E. Clauser, Swanson, & Harik, 2002; J. C. Clauser, Clauser, & Hambleton, 2011; Ferdous & Plake, 2008; Kane, 1987; Plake & Kane, 1991; van der Linden, 1982). The application of IRT to Angoff ratings is relatively straightforward. In IRT, the probability of a correct response on a given item is expressed as a function of various item parameters and (most often) a single examinee proficiency parameter—typically denoted θ . As observed above, standard setting concerns itself

Table 1. Characteristics of Standard Setting Data From Six Panels.

	2006		2013	
	No. of judges	No. of unfamiliar items (%)	No. of judges	No. of unfamiliar items (%)
Panel 1	9	97 (14)	10	24 (3)
Panel 2	10	25 (3)	10	23 (3)
Panel 3	10	8 (1)	8	13 (2)
Total	29	130 (6)	28	60 (3)

with but one such proficiency for each performance standard: the proficiency associated with the minimally proficient examinee. Let θ_{mpe} denote this particular proficiency. Note that when an IRT model for a given item is evaluated for θ_{mpe} , the probability it yields is the same probability that judges estimate when they rate an item: *the probability that a minimally proficient examinee will respond correctly to the item*. It follows that the mathematical expression for an IRT model, for example, the two-parameter logistic model,

$$P_i(\theta) = \frac{e^{a_i(\theta - b_i)}}{1 + e^{a_i(\theta - b_i)}}, \quad (1)$$

will have an analogous equation describing the relationship between a judge's rating, his or her conceptualization of θ_{mpe} and a given item's parameters in the two-parameter logistic model:

$$x_i = \frac{e^{a_i(\theta_{mpe} - b_i)}}{1 + e^{a_i(\theta_{mpe} - b_i)}} + \varepsilon_i \quad (2)$$

where a_i , and b_i , are the familiar discrimination, and difficulty parameters, respectively, for item i , x_i is the given Angoff rating for item i , and ε_i denotes the error associated with the rating x_i . Since our interest is in θ_{mpe} , not x_i (which is simply the given rating), Equation (2) may be solved for θ_{mpe} yielding

$$\theta_{mpe} = \frac{\ln\left(\frac{x_i - \varepsilon_i}{1 + \varepsilon_i - x_i}\right)}{a_i} + b_i. \quad (3)$$

In practice, ε_i will not be known and it is removed from the equation yielding the estimator:

$$\hat{\theta}_{mpe} = \frac{\ln\left(\frac{x_i}{1 - x_i}\right)}{a_i} + b_i \quad (4)$$

where the circumflex above $\hat{\theta}_{mpe}$ denotes that it is an estimate of the unknown θ_{mpe} .

Each transformed rating, $\hat{\theta}_{mpe}$, is itself a passing score estimate expressed on the IRT proficiency metric and in this sense differences across transformed ratings for a given rater follow only from inconsistencies in his or her judgments—not from differences in item difficulty. Put another way, if a judge were able to provide error-free judgments based on a consistent conceptualization of the minimally proficient examinee all ratings would yield the same recommended passing score on the IRT proficiency scale. This makes it possible to compare each rater's beliefs about the relative difficulty of familiar versus unfamiliar items. To the extent that panelists' transformed ratings for unfamiliar items differ from their transformed ratings for familiar items, it suggests that there is an interaction between panelists' familiarity with test content and their beliefs about the relative difficulty of that content for the minimally proficient examinee.

Analysis

For each judge, all Angoff ratings were transformed onto the IRT proficiency scale using Equation (4). The use of the two-parameter logistic model (2PL; Hambleton & Swaminathan, 1985) reflected in Equation (4) was chosen to mirror the IRT model used operationally for this exam. Unlike the original ratings, these transformed ratings account for differences in item difficulty and thus can be compared with one another directly—even when examinees rate unique sets of items. For each judge a distribution of these transformed ratings for familiar items was then constructed. Since only familiar item ratings were included, these null distributions represented the expected variability in judge's ratings for an exam where *all* content is familiar. Next, for each of the 29 judges with at least one flagged item the rank position for each *unfamiliar* item within their null distribution was calculated. This made it possible to compare the ratings for unfamiliar items with the expected ratings for familiar items. If judges' familiarity with test content does not affect their beliefs about relative item difficulty, ratings for unfamiliar items should be indistinguishable from those for familiar items. Conversely, if judges' perception of item difficulty is affected by their familiarity with the content, the unfamiliar items will systematically differ from familiar item ratings. Put another way, if item familiarity does not influence a judge's ratings, the transformed ratings for familiar and unfamiliar items will share a common distribution. When familiarity does influence item ratings, transformed ratings for unfamiliar items will appear to be outliers in the distribution of familiar test content.

For each judge, the percentile rank of each unfamiliar item rating within their null distribution was calculated. Percentile ranks were chosen because, excepting sampling error, they will be uniformly distributed when familiarity does not affect a judge's ratings. This is true because if there is no difference between ratings for familiar and unfamiliar items (e.g., the two set of items share a distribution) the unfamiliar items will appear to be randomly distributed within the distribution of familiar items, yielding uniformly random percentile ranks. When unfamiliar items differ

systematically from the distribution of familiar ratings the percentile ranks of unfamiliar items will fail to approximate a uniform distribution. The extent to which uniformity was observed is (a) shown visually by plotting frequency histograms of percentile ranks and (b) tested formally using the nonparametric Kolmogorov–Smirnov (K-S) test.

Additionally, unfamiliar item ratings were compared with familiar item ratings using linear regression. The untransformed Angoff ratings were regressed onto the familiarity variable along with dummy coded variables for both item and judge. When data are modeled in this way, the coefficient to the item familiarity predictor indicates the strength and significance of the relationship between familiarity and Angoff ratings after controlling for item difficulty and judge stringency. By using Angoff ratings in the linear regression rather than the transformed ratings the regression coefficients can be interpreted as the change in Angoff rating due to familiarity with the items content.

Finally, passing scores were calculated with unfamiliar items excluded. These passing scores were calculated by mapping each judge's raw passing score onto the IRT ability scale using a test characteristic curve based exclusively on familiar items. These passing scores were then averaged across judges within the panel. To provide context for these results, pass rates were calculated using these passing scores and operational examinee performance from the previous exam administration.

Results

When judges' perceptions of item difficulty are not affected by content familiarity, the rank positions of unfamiliar items will be random within the null distribution of familiar content. This random distribution of percentile ranks would yield uniformly distributed rating deciles for unfamiliar content. Nonuniform distributions of percentile ranks for unfamiliar items suggest that item familiarity affects panelists' beliefs about item difficulty. For the two datasets analyzed in this study, these distributions are shown in Figures 1 and 2. In both cases, distributions were *positively* skewed, suggesting that panelists perceive unfamiliar content as being more difficult than we would expect given their judgments about familiar content and the actual behaviors of examinees.

The failure of the rank positions to conform to the uniform expectation under the null hypothesis can be seen clearly in the figures. Furthermore, this finding was confirmed formally by the K-S test results, which rejected the null hypothesis that the rank positions were uniformly distributed. K-S statistics were 4.888 and 2.620 for 2006 and 2013, respectively, which corresponded to p -values $< .001$ in both cases.

These results provide evidence that when judges are asked to rate an unfamiliar item, they tend to provide lower ratings than for familiar items, indicating that they are perceived as being more difficult for the minimally proficient examinee. These findings were reiterated by the linear regression analyses. The regression analysis with untransformed ratings is beneficial because the magnitude of the effect can be

Figure 1. Distribution rank position of transformed Angoff ratings for unfamiliar items 2006 data.

Figure 2. Distribution rank position of transformed Angoff ratings for unfamiliar items 2013 data.

interpreted on the familiar p -value scale rather than on the IRT proficiency scale for this particular exam. For both datasets, unfamiliarity with an item was a significant predictor of Angoff ratings ($p < .001$), with regression coefficients of 0.138 in 2006 data and 0.107 in 2013 on the p -value scale. The R^2 values were 0.388 in 2006 and 0.507 in 2013. These results indicate that when holding judge and item constant, an unfamiliar item has an expected Angoff rating 0.138 lower than a familiar item for the 2006 data and 0.107 lower than a familiar item for the 2013 data. These results suggest that the recommended passing score could be profoundly affected by a judge's familiarity with specific exam content.

The results presented above indicate that judges' ratings are affected by their expertise, but they do not indicate the extent to which the final recommended passing score is affected. For this standard setting activity, a given panel's recommended passing score was based on between 600 and 750 unique ratings (i.e., number of panelists multiplied by 75 items). While a relatively small number of unfamiliar items is not likely to have a substantial impact on the final passing score, as the number of unfamiliar items increases, so should our concerns. In this study, all six panels saw their recommended passing score increase when unfamiliar items were omitted. For the two panels with the fewest items marked as unfamiliar (2006 Panel 3 and 2013 Panel 3) this increase on the IRT proficiency scale was modest and the impact on the raw recommended passing score was less than 1 point on the 308 item test. For panels with more unfamiliar items, the increase may be of greater practical consequence. The three panels with between 23 and 25 items omitted had their number-correct passing scores increase by 1.6 to 2.8 raw score points on the full-length exam. For the panel with the greatest number of marked items, Panel 1 from 2006, 97 items were marked as unfamiliar. Omitting these marked items increased the raw passing score by more than 10 items—from 167.0 to 177.2. Increasing the passing score by these magnitudes would be of practical significance for many tests. For example, the change in 2006 Panel 1's passing score would have decreased the pass rate from 98% to 96.5%. Note, however, that this exam's population is highly proficient; had a change of a similar magnitude occurred elsewhere in the proficiency distribution rather than in the lower tail, the passing rate could have declined by as much as 10%.

Discussion

When exams are used to make classification decisions, the range of examinee performance associated with each performance category must be defined. While it is generally believed that the process of defining these ranges can be productively informed by expert opinion, our findings suggest that these opinions may be contaminated by systematic error when judges are presented with unfamiliar test content. In the context of the Angoff method, misperceptions of this kind appeared to yield artifactually low ratings for our datasets. Moreover, because the Angoff method traditionally has no mechanism for excluding ratings, these errors would have lowered the resultant recommended passing score. In general, the magnitude of such effects will depend on the proportion of unfamiliar exam content but it is conceivable that they could have practical consequences for examinees and other stakeholders by undermining the validity of inferences based on examinee classification decisions (American Educational Research Association, American Psychological Association, National Council on Measurement in Education, & Joint Committee on Standards for Educational and Psychological Testing, 2014).

In principle, several statistical approaches exist to allow judges to omit unfamiliar content. For example, within the classical test theory framework rather than summing each judge's ratings to arrive at his or her recommended passing score, the recommended percent correct passing score could be calculated based exclusively on

familiar items. These judge-level percent correct passing scores would then be aggregated across all judges in the panel. This approach is limited; however, because it fails to account for differences in the difficulty of each judge's set of rated items. This creates the potential for bias, when a judge omits particularly difficult or particularly easy test items. The use of IRT provides another approach for judges to omit content, but properly accounts for differences in item difficulty.

Translating Angoff ratings to an IRT proficiency scale provides a mechanism for combining ratings when each judge has rated a unique subset of items. Rather than estimating each judge's recommended IRT passing score using the complete test characteristic curve, each judge has their passing score estimated using their own test characteristic curve based exclusively on rated items. These judge-specific passing scores on the IRT proficiency scale can then be aggregated to arrive at the final recommended passing score for the panel. For exams with established IRT model fit, this approach gives judges the flexibility to exclude ratings for unfamiliar items, allowing the effects of unfamiliar items to be avoided.

Although our results suggest that excluding unfamiliar items may eliminate a source of systematic error, there are occasions when this approach can lead to unintended outcomes. For example, it is customary to balance the composition of a standard setting panel with respect to various rater characteristics such as race, gender, specialty, workplace setting, and other factors. However, in some cases, such characteristics may be related to a judge's propensity to mark an item as unfamiliar even after controlling for differences in content familiarity. If judges representing one of these groups are far more likely to indicate that items are unfamiliar, excluding these ratings may interfere with the desired balance of judges within the panel. As a practical matter it is certainly possible to weight all judges equally, even when they have rated different numbers of items, but each judge's recommended passing score will be based on a unique selection of content. Practitioners will have to determine if it is more defensible to base each judge's recommended passing score on different items, or to include ratings for content the judges have identified as unfamiliar.

Similarly, judges' propensity to omit content could be systematically influenced by content area. If a large number of judges are unfamiliar with the same content area, the influence of that area on the passing score will be diminished. In extreme cases, an individual test item could be omitted by all judges, effectively removing it from the standard setting process. Although, in principle this should not interfere with the estimation of the passing score it may call into question the composition of the standard setting panel, the distribution of content on the exam, or both. As with the composition of panels of content experts, policy makers will have to determine if it is more defensible to allow judge to alter the distribution of exam content or to insist that judges provide ratings to unfamiliar content.

When considering excluding ratings to unfamiliar content it is important to remember that we cannot know the judge's true internalized ability of the minimally proficient examinee. Therefore, we can never definitively know the distribution of errors associated with the judge's ratings. Empirical standard setting studies

frequently assume that error distributions are approximately normal (Plake & Kane, 1991). This assumption is in keeping with the common operational practice of averaging ratings across items, but little research has explored this normality of errors assertion. In the context of this study, the error distribution has profound implications for the treatment of ratings for unfamiliar content. If the errors for typical *familiar* ratings are approximately normal, exclusion of unfamiliar content will eliminate a source of systematic error and increase the accuracy of the estimated passing score. Since we do not know the judge's true internalized ability for the minimally proficient examinee, however, other interpretations are possible.

If the errors for judges' ratings are approximately normal across *all* the items, they are presented, the exclusion of unfamiliar content would serve to bias the resultant passing score. In the most extreme case, it could be argued that content experts have inherently inflated expectations for the minimally proficient examinee. From this perspective, the familiar ratings may be *more* prone to error than the unfamiliar content. Although, we cannot know definitively which of these interpretations is correct, reason suggests that when items are explicitly identified as unfamiliar their ratings should be treated with suspicion. Practitioners will have to determine what is appropriate in the context of their particular exam.

Limitations

The results of this study indicate that in both of our datasets judges provide systematically lower ratings for unfamiliar exam content. Although these results are interesting they are based exclusively on standard setting data from Step 1 of the USMLE and therefore these findings may not generalize to other testing contexts. Step 1 of the USMLE covers a broad content domain including both medicine, and biomedical sciences. Because of the nature of the domain, judges are often many years removed from significant work with the tested material and therefore may have large deficits in their content mastery. The findings within this context may not generalize to other high-stakes testing environments where content experts have more consistent exposure to the subject matter. Further research is needed to determine if these findings have implications for other testing programs.

In addition, because these results were collected as part of an operational standard setting meeting, some features of the design could not easily be manipulated. For example, judges were provided with the item key for each item despite the potential for this to influence their propensity to classify an item as unfamiliar. The research is split as to whether this hindsight bias affects ratings (Hudson & Campion, 1994; Margolis, Mee, Clauser, Winward, & Clauser, 2016), but the presence of the key may influence a judge's willingness to mark content as unfamiliar. In addition, because their ratings were used operationally, judges were notified that unfamiliar items would be excluded from their recommended passing score. This may have resulted in judges taking these ratings less seriously than ratings for familiar items. Additional research is needed to explore how these factors influence item classification and ratings.

Finally, the results of this study are based exclusively on Angoff standard setting data. Thus, while it is reasonable to believe that if a judge tends to overestimate the difficulty of unfamiliar content, he or she will do so independent of the choice of standard setting method; this choice may still affect the extent to which the recommended passing score is affected. For example, in the bookmark method (Lewis et al., 1999) the degree to which the passing score is affected is likely to be dependent on the positions of the unfamiliar items within the booklet. If unfamiliar items appear sporadically within the booklet, those items may have no effect at all on the final recommended passing score. When this is the case, the bookmark method may be more resilient to the effects of unfamiliar items because few unfamiliar items will be in the region of the judge's bookmark placement. Conversely, if several unfamiliar items appear on consecutive pages near the judge's desired passing score the effect on the passing score may be practically significant. When this is the case, the bookmark method has limited opportunity to remove the influence of these unfamiliar items. Ultimately, further research is needed to determine how these unfamiliar items influence bookmark standard setting results.

Conclusions

The Angoff standard setting method relies on the opinions of content experts to determine the level of examinee performance associated with specific classification categories. In some cases, content experts may have gaps in their knowledge of the exam content. Such content deficiencies may be more likely to occur when the content domain is broad and/or highly technical (as is the case in many credentialing exams) or when panelists are selected for reasons other than their content expertise (as when parents, administrators, union representatives, or other nonexpert stakeholders are included in a standard setting panel). Regardless of the reason, when judges are presented with unfamiliar exam content practitioners must determine how these items and their associated ratings should be treated. The findings presented in this study—that unfamiliar items receive systematically lower ratings on average than familiar items—indicate that this decision has the potential to have significant practical consequences for examinees and other groups of stakeholders. Although many factors—both practical and statistical—should be considered before omitting panelists' ratings, results from this study suggest excluding ratings judges identify as untrustworthy may produce passing scores that are both theoretically more defensible and psychometrically more accurate.

Acknowledgments

The authors would like to thank the National Board of Medical Examiners for their tremendous support and guidance throughout the development of this research. We are particularly indebted to Brian E. Clauser, Melissa J. Margolis, Janet Mee, and Marcia L. Winward for conducting the data collection and analysis. This work would not have been possible without their contributions.

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) received no financial support for the research, authorship, and/or publication of this article.

References

- American Educational Research Association, American Psychological Association, National Council on Measurement in Education, & Joint Committee on Standards for Educational and Psychological Testing. (2014). *Standards for educational and psychological testing*. Washington, DC: Authors.
- Angoff, W. H. (1971). Scales, norms, and equivalent scores. In R. L. Thorndike (Ed.), *Educational measurement* (2nd ed., pp. 508-600). Washington, DC: American Council on Education.
- Busch, J. C., & Jaeger, R. M. (1990). Influence of type of judge, normative information, and discussion on standards recommended for the National Teacher Examinations. *Journal of Educational Measurement*, 27, 145-163.
- Chang, L., Dziuban, C. D., Haynes, M. C., & Olson, A. H. (1996). Does a standard reflect minimal competency of examinees or judge competency? *Applied Measurement in Education*, 9, 161-173.
- Clauser, B. E., Swanson, D. B., & Harik, P. (2002). A multivariate generalizability analysis of the impact of training and examinee performance information on judgments made in an Angoff-style standard-setting procedure. *Journal of Educational Measurement*, 39, 269-290.
- Clauser, J. C., Clauser, B. E., & Hambleton, R. K. (2011, April). *Variability across judges in their ability to estimate probabilities: The case for weighting*. Paper presented at the meeting of the National Council on Measurement in Education, New Orleans, LA.
- Davey, T., Fan, M., & Reckase, M. D. (1996, April). *Some new methods for mapping ratings to the NAEP θ -scale to support estimation of NAEP achievement level boundaries*. Paper presented at the meeting of the National Council on Measurement in Education, New York, NY.
- Ferdous, A. A., & Plake, B. S. (2008). Item response theory-based approaches for computing minimum passing scores from Angoff-based standard-setting study. *Educational and Psychological Measurement*, 68, 778-796.
- Hambleton, R. K., & Swaminathan, H. (1985). *Item response theory: Principles and applications*. Boston, MA: Kluwer Academic.
- Hudson, J. P., & Campion, J. E. (1994). Hindsight bias in an application of the Angoff method for setting cutoff scores. *Journal of Applied Psychology*, 79, 860-865.
- Kane, M. T. (1987). On the use of IRT models with judgmental standard setting procedure. *Journal of Educational Measurement*, 24, 333-345.
- Kane, M. T. (2001). So much remains the same: Conception and status of validation in setting standards. In G. J. Cizek (Ed.), *Standard setting: Concepts, methods, and perspectives* (pp. 53-88). Mahwah, NJ: Lawrence Erlbaum.
- Lewis, D. M., Mitzel, H. C., Green, D. R., & Patz, R. J. (1999). *The bookmark standard setting procedure*. Monterey, CA: McGraw-Hill.

- 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.
- Loomis, S. C. (2012). Selecting and training standard setting participant: State of the art policies and procedures. In G. J. Cizek (Ed.), *Setting performance standards: Foundations, methods, and innovations* (2nd ed., pp. 107-134). New York, NY: Routledge.
- Margolis, M. J., Mee, J., Clauser, B. E., Winward, M., & Clauser, J. C. (2016). Effect of content knowledge on Angoff-style standard setting judgments. *Educational Measurement: Issues and Practice*, 35, 29-37.
- Meara, K., Hambleton, R. K., & Sireci, S. G. (2001). Setting and validating standards on professional licensure and certification exams: A survey of current practices. *CLEAR Exam Review*, 12(2), 17-23.
- National Research Council. (1999). Setting reasonable and useful performance standards. In J. W. Pellegrino, L. R. Jones & K. J. Mitchell (Eds.), *Grading the nation's report card: Evaluating NAEP and transforming the assessment of educational progress* (pp. 162-184). Washington, DC: National Academies Press.
- Norcini, J. J., Shea, J. A., & Kanya, D. T. (1988). The effects of various factors on standard setting. *Journal of Educational Measurement*, 25, 57-65.
- Plake, B. S., Impara, J. C., & Potenza, M. T. (1994). Content specificity of expert judgments in a standard-setting study. *Journal of Educational Measurement*, 31, 339-347.
- Plake, B. S., & Kane, M. T. (1991). Comparison of methods for combining the minimum passing levels for individual items into a passing score for a test. *Journal of Educational Measurement*, 28, 249-256.
- Plake, B. S., Melican, G. J., & Mills, C. N. (1991). Factors influencing intrajudge consistency during standard-setting. *Educational Measurement: Issues and Practice*, 10(2), 15-16.
- Raymond, M. R., & Reid, J. B. (2001). Who made thee a judge? Selecting and training participants for standard setting. In G. J. Cizek (Ed.), *Standard setting: Concepts, methods and perspectives* (pp. 119-157). Mahwah, NJ: Lawrence Erlbaum.
- Reckase, M. D. (1998). Converting boundaries between national assessment governing board performance categories to points on the National Assessment of Educational Progress Scale: The 1996 Science NAEP Process. *Applied Measurement in Education*, 11, 9-21.
- Saunders, J. C., Ryan, J. P., & Huynh, H. (1981). A comparison of two approaches to setting passing scores based on the Nedelsky procedure. *Applied Psychological Measurement*, 5, 209-217.
- van der Linden, W. J. (1982). A latent method for determining intrajudge inconsistency in the Angoff and Nedelsky techniques of standard setting. *Journal of Educational Measurement*, 19, 295-308.
- Verheggen, M. M., Muijtjens, A. M. M., Van, Os, J., & Schuwirth, L. W. T. (2008). Is an Angoff standard an indication of minimal competence of examinees or of judges? *Advances in Health Sciences Education*, 13, 203-211.
- Verhoeven, B. H., Verwijnen, G. M., Muijtjens, A. M. M., Scherpbier, A. J. J. A., & van der Vleuten, C. P. M. (2002). Panel expertise for an Angoff standard setting procedure in progress testing: Item writers compared to recently graduated students. *Medical Education*, 36, 860-867.
- Wayne, D. B., Cohen, E., Makoul, G., & McGaghie, W. C. (2008). The impact of judge selection on standard setting for a patient survey of physician communication skills. *Academic Medicine*, 83(10 Suppl.), S17-S20.