Educational Measurement
ISSUES AND PRACTICE

NCME national council on measurement in education

Educational Measurement: Issues and Practice
Spring 2014, Vol. 33, No. 1, pp. 15–22

The Impact of Examinee Performance Information on Judges' Cut Scores in Modified Angoff Standard-Setting Exercises

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

This research evaluated the impact of a common modification to Angoff standard-setting exercises: the provision of examinee performance data. Data from 18 independent standard-setting panels across three different medical licensing examinations were examined to investigate whether and how the provision of performance information impacted judgments and the resulting cut scores. Results varied by panel but in general indicated that both the variability among the panelists and the resulting cut scores were affected by the data. After the review of performance data, panelist variability generally decreased. In addition, for all panels and examinations pre- and post-data cut scores were significantly different. Investigation of the practical significance of the findings indicated that nontrivial fail rate changes were associated with the cut score changes for a majority of standard-setting exercises. This study is the first to provide a large-scale, systematic evaluation of the impact of a common standard setting practice, and the results can provide practitioners with insight into how the practice influences panelist variability and resulting cut scores.

Keywords: modified Angoff, standard setting

The process of setting standards and establishing appropriate cut scores is a critical step in any context that uses test scores to make decisions about examinees. Test-based decisions are made in a variety of areas such as high school graduation or equivalency, course placement, and entry into a number of professions, including medicine, law, and accounting. Though it may be widely agreed that setting standards is an important part of testing, how standards and cut scores are established often does not receive widespread attention from individuals other than those who are involved in the test construction process.

Establishing cut scores is not simply an important independent step in the testing process; this activity is critical with respect to the overall validity of test score interpretations (Kane, 1992, 1994, 2001, 2006). Despite the clear importance of providing evidence for the validity of cut-score decisions and the fact that much has been written about standard setting (e.g., Cizek, 2012), definitive research is lacking in many areas. As a testament to the importance of this line of research, the 2006 National Council on Measurement in Education Career Award Address was dedicated to the subject of standard setting and emphasized the need for additional research in this area (Plake, 2008).

The process used to establish cut scores will differ based on the specifics of the testing context and the personal preferences of the practitioner; there are many different procedures (and variations on those procedures) that are used in practice. Perhaps the most common approach to setting standards for multiple-choice examinations is the Angoff method (Angoff, 1971). In the traditional Angoff approach, standard-setting judges first are asked to conceptualize a "minimally proficient" examinee (the examinee whose level of proficiency justifies passing, but just barely). They then are asked to review test items and, for each one, to provide an estimate of the proportion of minimally proficient examinees that would answer the item correctly. These proportions are aggregated across all of the test items and across all judges, and the resulting value provides an estimate of the score that a minimally proficient examinee would be expected to receive on the test (i.e., the cut score).

Over time, a number of modifications to the Angoff procedure have been suggested. One of the most common modifications involves providing judges with performance data for the items they review; this allows them to get a sense of how people taking the test actually perform on the items. Previous research suggests that in the absence of such data content experts may have a difficult time making judgments that sensibly correspond to actual item difficulties (Busch & Jaeger, 1990; Clauser, Mee, Baldwin, Margolis, & Dillon, 2009; Clauser, Swanson, & Harik, 2002).

Perspectives on the impact of providing performance data are mixed. Hambleton (2001) expressed the view that the impact "may be more psychological than psychometric" (p. 102) and asserted that the main impact often is on the variability among panelists rather than on the overall estimated cut score. Brandon (2004) reviewed six studies and found that providing performance data led to significant cut score changes in four of them. Hurtz and Auerbach (2003) provided a meta-analysis of studies on the Angoff

Melissa J. Margolis, Senior Measurement Scientist, National Board of Medical Examiners, 3750 Market Street, Philadelphia, PA 19104; mmargolis@nbme.org. Brian E. Clauser, Vice President, Measurement Consulting Services, National Board of Medical Examiners, 3750 Market Street, Philadelphia, PA 19104; bclauser@nbme.org.

procedure and concluded that providing performance data generally resulted in lowering the cut score. In other studies that have investigated how the provision of performance data during the Angoff standard-setting process impacts the resulting cut scores, the results are far from conclusive: some found increases in cut scores, some found decreases in cut scores, and some found both increases and decreases across different content areas on the same test (Busch & Jaeger, 1990; Clauser et al., 2002; Cross, Impara, Frary, & Jaeger, 1984; Plake, Impara, & Potenza, 1994; Truxillo, Donahue, & Sulzer, 1996). The more uniform findings across studies, as suggested by Hambleton (2001), were that providing data generally leads to convergence (i.e., group cut scores getting closer together; Busch & Jaeger, 1990; Clauser et al., 2002; Cross et al., 1984; Plake et al., 1994; Truxillo et al., 1996). Though much research has been dedicated to examining this issue, the lack of consistency in the findings is indicative of the need for a large-scale, systematic investigation of the problem.

This study provides a contribution to the literature by presenting a systematic investigation of the practice of providing examinee performance data to content-based standard-setting judges in multiple high-stakes standard-setting exercises. Though this methodological approach has become relatively common in the context of standard setting for credentialing examinations, the impact of providing such data is not fully understood and may have considerable implications for the validity of the resulting decisions. This paper attempts to answer three questions: (1) Does providing performance data result in changes in the estimated cut score? (2) Are these changes large enough to be of practical importance? And (3) are the changes consistently in the same direction? The advantage of the current data set is that it allows for asking these questions across three different examinations, each of which had two separate standard-setting exercises (conducted in different years) that included multiple panels; across examinations, years, and panels other specifics of the standard-setting procedure were held constant. Previous research has been based on smaller data sets and previous studies trying to combine results across studies may suffer from a selection (i.e., publication) bias. The fact that the current study reports on all standard-setting exercises from a high-stakes testing program over a number of years to some extent guards against selection bias.

Methods

The United States Medical Licensing Examination (USMLE) provides a single examination pathway for allopathic (i.e., MD) physicians in the United States. The USMLE process is divided into three independent steps, each of which necessitates taking and passing one or more examinations. The three multiple-choice examinations that comprise USMLE and were the focus of this research are as follows. Step 1 is a single-day examination that assesses knowledge of the sciences that are basic to the practice of medicine. Step 2 is comprised of two independent examinations: Step 2 Clinical Knowledge (CK) and Step 2 Clinical Skills (CS). Step 2 CK is a single-day examination that assesses the clinical knowledge that is essential for the safe and effective practice of medicine under supervision. Step 2 CS is a day-long examination in which examinees interact with a series of actors trained to play the part of patients. The standard-setting approach for this test is unlike that used for the other three examinations and therefore was not considered as part of this research. Step 3 is a two-day examination comprised of multiple-choice questions and computer-based case management scenarios that assesses application of medical knowledge and understanding of clinical science that are considered essential for the unsupervised practice of medicine; only the multiple-choice component of Step 3 was included in this research.

Study Design

As part of operational practice, a content-based standard-setting exercise is conducted approximately every three years for each of the USMLE examinations. Data for the present research were taken from two independent content-based standard-setting exercises for each of the three step examinations. For each exercise, three independent panels of approximately 8–10 content experts were convened over a span of one–two months; the methodology was the same for all panels.

Process training. Training on the content-based standard-setting process began with a discussion of the concept of the Minimally Proficient Examinee (MPE): the examinee whose performance is just acceptable when measured against the standard of interest. This is a critical step in the process, because understanding of and comfort with this concept is the foundation for all of the work that the panelists do. Panelists first were asked to think about the idea of the MPE and then to try to describe the characteristics of this person. Following initial discussion of the topic, some additional process-related details were provided and the training moved into the next phase: practice judgments.

Practice judgments. For this initial exercise, judges were given a booklet containing 15 practice items and the associated answer key for those items. These practice items were selected to represent a range of item presentations that the judges would see throughout the main item set (i.e., the items were single-best answer multiple choice questions, but the specifics of the items varied; for example, items differed in numbers of options, some items had associated images, some items had associated tables or charts.). The items also were selected to represent a range of difficulty levels so that the judges would have initial experience with providing estimates for and being able to discuss items of easy, moderate, and high difficulty. The judges reviewed each item one at a time, provided written judgments about their performance expectations, reported their judgments to the group and had them recorded on a whiteboard, and then discussed the reasons for providing the judgments they made. The question about performance expectations was asked in the following way: "What is the probability that a MPE would answer this item correctly?" Responses were recorded as whole numbers, and judges were permitted to use any numbers they deemed appropriate (from 0 to 100).

After going through each item, providing initial judgments, and discussing the judgments with the group, judges were provided with data that indicated how a cohort of examinees had performed on each of the test items. Two distinct types of performance data were presented: (1) item-level graphs showing how examinees at each performance decile (based on total test score) performed on the item and (2) a

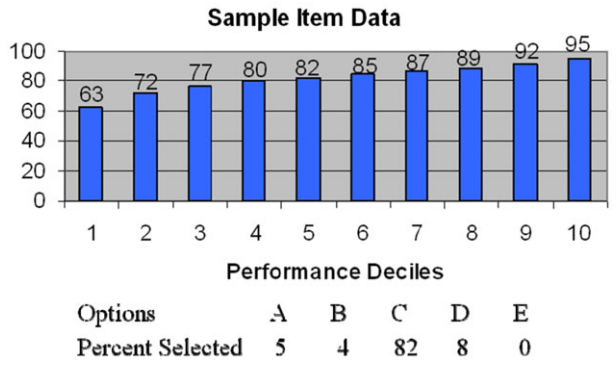

FIGURE 1. Sample Performance Data Presented to Standard-Setting Judges.

breakdown of the percentage of examinees that had chosen each of the multiple-choice options (see Figure 1). Judges were asked to review the performance data, review the item and their initial judgment again, and decide whether or not they thought that they should change their original judgment based on the data. The reason for providing the data is that previous research indicates that judges have difficulty making the required judgments without some sense of how examinees actually perform on the items (Busch & Jaeger, 1990; Clauser et al., 2002; Clauser, Harik et al., 2009; Clauser, Mee et al., 2009). Judges were instructed to use the data merely as a guide; if their judgment and the data were quite different, going back and taking a more careful look at the item could reveal that the judge missed something during the initial review. After reviewing the data for an item, judges were asked to provide a final judgment (which may or may not have been the same as the initial judgment) and again the judgments were recorded and discussed.

After judges completed discussing their second round of judgments for the 15-item practice set, they began the same iterative process of reviewing and providing judgments for a set of 75 items. Unlike the 15-item practice sets, the 75-item sets were created with careful attention to content and difficulty parameters so that they represented shortened versions of actual test forms. Judges first reviewed and provided initial judgments for all items without data; they then reviewed data and provided final judgments for each item. For this set of items there was no discussion amongst judges; all of the work was done independently. The 75-item set for all three panels within an exercise was the same.

In practice, the 75-item set is considered part of training and there are two additional steps in the standard-setting process: (1) results from the 75-item set are used to provide judges with feedback on how their judgments compared to those of the other judges and how individual and group judgments would impact a reference group of examinees, and (2) following this feedback session, judges complete a larger set of independent item judgments (typically ranging from 150 to 225 items). It is the results from this final set of judgments that are presented to the policymaking groups to inform the cut-score decision. There were two main reasons for using the 75-item sets for this research: (1) within each exercise, the three panels each saw the same set of 75 items (this allows for direct comparison of the estimated cut scores), and (2) results from the first independent review of the 75-item set represent the only set of judgments that occurs before judges have an opportunity to calibrate themselves based on representative performance data or impact feedback. The initial and final Angoff judgments resulting from these 75-item sets are the data that were used for the current analysis.

Analysis

Descriptive statistics (mean and standard deviation) for the initial and final (pre- and post-data) cut scores were calculated by year, by panel, and across panels for each examination. To examine whether there was a general effect that resulted in raising or lowering cut scores across examinations, years, and panels, a regression analysis was completed. Individual judge-level cut scores acted as the dependent measure and nominal variables coding examination, year, panel, and whether the cut score was initial or final were included in the model. If this final parameter estimate were significantly different from zero it would suggest that the final cut scores were significantly different from the initial cut scores and therefore that the provision of performance data makes a difference. The specific model is:

$$\begin{aligned}\text{Cut score} = {} & b0 + b1(\text{initial/final}) + \text{b2(step2)} \\ & + b3(\text{step3}) + \text{b4(panel2)} \\ & + b5(\text{panel3}) + \text{b6(year2)} + \text{error}.\end{aligned}$$

The step, panel, and year variables are coded one if they correspond to that step, panel, or year and zero otherwise. The initial/final variable is coded zero for the initial and one for the final cut score.

In addition to this general test, a 2 × 3 (initial or final cut score by panel) repeated-measures analysis of variance (ANOVA) was conducted for each year within each examination to investigate whether any identified changes between pre- and post-data cut scores within and across panels were statistically significant.

The pre- and post-data cut scores also were applied to cumulative frequency distributions for a reference group of examinees (first-time takers who are students or graduates of U.S. medical schools based on the most recent data available at the time the standard setting was conducted), and comparisons between the fail rates that resulted from the two sets of judgments were made. These comparisons provide a practical indication of the impact of the different cut scores on a defined population of examinees.

Results

Tables 1, 2, and 3 present descriptive statistics (mean and standard deviation) for the initial and final cut scores by year and by panel for the Step 1, Step 2, and Step 3 examinations, respectively. For 6 of the 18 total panels across the three examinations, the cut score decreased after receiving performance data; for the remaining 12 panels, final cut scores were higher than initial cut scores. Cut scores increased following the review of performance data for all of the Step 1 panels across both standard-setting years; though the other two examinations did not have a pattern that was consistent across both years, the Step 3 Year 2 data did display a decrease in cut scores following the review of performance data for all three panels.

The results of the overall regression analysis are presented in Table 4. The important result is that the variable representing the contrast between the initial and final cut scores is not

Table 1. Step 1 Panel-level Descriptive Statistics for Initial (Pre-Data) and Final (Post-Data) Cut Scores

			Initial Cut Score		Final Cut Score	
Standard- Setting Year	Panel	Sample Size	Mean	Standard Deviation	Mean	Standard Deviation
1	1	10	61.19	17.60	63.61	17.01
	2	9	61.53	19.22	62.21	18.05
	3	10	59.09	18.57	60.69	18.20
2	1	9	54.61	16.04	58.71	12.72
	2	10	56.74	20.44	60.58	15.88
	3	10	51.81	20.40	61.87	13.86

Note. Step 1 assesses knowledge of the sciences that are basic to the practice of medicine.

Table 2. Step 2 Panel-Level Descriptive Statistics for Initial (Pre-Data) and Final (Post-Data) Cut Scores

			Initial Cut Score		Final Cut Score	
Standard- Setting Year	Panel	Sample Size	Mean	Standard Deviation	Mean	Standard Deviation
1	1	10	60.19	19.59	61.17	16.45
	2	8	55.19	20.63	59.23	18.64
	3	9	65.92	19.71	65.14	14.74
2	1	9	64.89	18.28	68.22	15.00
	2	8	63.91	17.84	64.50	15.91
	3	11	67.35	20.40	66.52	16.94

Note . Step 2 CK assesses the clinical knowledge that is essential for the safe and effective practice of medicine under supervision.

Table 3. Step 3 Panel-Level Descriptive Statistics for Initial (Pre-Data) and Final (Post-Data) Cut Scores

			Initial Cut Score		Final Cut Score	
Standard- Setting Year	Panel	Sample Size	Mean	Standard Deviation	Mean	Standard Deviation
1	1	9	61.61	16.46	64.91	13.38
	2	9	71.89	14.42	71.07	11.96
	3	10	62.78	16.54	64.66	12.31
2	1	12	75.23	14.23	70.55	12.95
	2	12	75.35	15.82	73.84	14.17
	3	11	68.19	15.84	67.15	13.26

Note . Step 3 assesses application of medical knowledge and understanding of clinical science that are essential for the unsupervised practice of medicine.

Table 4. Results of Regression Analysis for Full Data Set

Effect	Unstandardized Coefficients		Standardized Coefficients		
	B	Std. Error	Beta	*t*	Sig.
(Constant)	56.101	1.535		36.545	.000
Step2	4.306	.989	.236	4.355	.000
Step3	9.454	.955	.536	9.895	.000
Panel2	.815	.984	.045	.828	.408
Panel3	–.812	.959	–.046	–.847	.398
Year2	1.953	.794	.116	2.460	.014
Initial/Final	1.550	.792	.092	1.957	.051

significant ($p = .051$). The result does approach significance, but the results reported in Tables 1–3 suggest that this effect is driven substantially by the change in Step 1 cut scores. To more closely examine this effect, the analysis was repeated for the Step 2 and Step 3 data only. The full results are not reported due to space constraints, but the significance level for the variable representing the contrast between the initial and final cut scores was altered substantially ($p = .648$). Similarly, repeating the analysis for the Step 1 data yielded a clearly significant result ($p = .006$).

When analyzed separately, the change in mean cut scores between the initial and final judgments was statistically significant ($p < .05$) for all examinations and all years. In addition, with the exception of Step 1 Year 1 ($F = 1.57, p = .21$) there was a significant interaction ($p < .05$) between mean initial and final cut scores by panel, indicating that the way that cut scores changed between initial and final judgments (i.e., whether they increased or decreased) varied across the individual panels. To save space, the complete ANOVA tables are not shown. These tables are available from the first author.

Descriptive statistics also were computed across panels for each year and examination by calculating the mean and standard deviation of the panel means. For all but one of the six standard-setting years (Step 3 Year 2), the overall mean cut score increased following review of performance information.

Table 5. Mean Initial (Pre-Data) and Final (Post-Data) Cut Scores and Associated Percents of Failing Examinees for All Examinations and Standard-Setting Years

		Initial Cut Score			Final Cut Score		
	Standard-Setting Year	Mean	Standard Deviation	Percent of Failing Examinees	Mean	Standard Standard Deviation	Percent of Failing Examinees
Step 1	1	60.60	1.33	4.2	62.17	1.46	5.5
	2	54.38	2.47	2.1	60.38	1.59	6.4
Step 2	1	60.43	5.37	5.1	61.85	3.02	6.9
	2	65.38	1.78	8.8	66.41	1.86	8.8
Step 3	1	65.43	5.63	5.1	66.88	3.63	7.8
	2	72.92	4.10	39.1	70.51	3.35	22.8

Note . Steps 1, 2, and 3 assess knowledge of the sciences that are basic to the practice of medicine, clinical knowledge that is essential for the safe and effective practice of medicine under supervision, and application of medical knowledge and understanding of clinical science that are essential for the unsupervised practice of medicine, respectively.

The issue of convergence was evaluated by reviewing the standard deviation results for pre- and post-data judgments. The results indicate a post-data decrease in standard deviation within all 18 panels; this decrease ranged from .37 (Step 1 Year 1 Panel 3) to 6.54 (Step 1 Year 2 Panel 3). The chance of all 18 standard deviations changing in the same direction by chance may be viewed as $1/2^{17}$, so the result clearly is statistically significant. The variability of the panel means within each of the six standard settings is slightly different; the standard deviation for the final cut score actually increased slightly for two of the standard-setting years: for Step 1 Year 1 the standard deviation increased from 1.33 to 1.46 and for Step 2 Year 2 the standard deviation increased from 1.78 to 1.86. For the other years, the average decrease in standard deviation was 1.5 and ranged from .75 (Step 3 Year 2) to 2.35 (Step 2 Year 1).

The final set of analyses investigated the practical implications of the results by applying the pre- and post-data cut scores to year- and examination-specific cumulative frequency distributions for a defined group of examinees (first-time takers who are students or graduates of U.S. medical schools). Comparisons between resulting fail rates from the two sets of judgments provide a practical indication of the impact of the cut scores on a defined population of examinees. (Due to space constraints, only selected fail rate information is presented here. Detailed information can be provided by the first author upon request.)

Review of the Step 1 results for both standard-setting years indicates that for all but one of the panels the percentage of failing examinees increased after judges reviewed performance data; the smallest change was a .8% increase for Year 1 Panel 3 (from 3.4% to 4.2%), and the largest change was a 6.6% increase for Year 2 Panel 3 (from 1.3% to 7.9%). The one panel that was the exception was Year 1 Panel 2; though there was a slight increase in the mean cut score for final judgments, both the initial and final cut scores would have failed 5.5% of examinees.

The Step 2 results present a slightly different picture: for this examination, review of the data led to an increase in the percentage of failing examinees for three panels, a decrease in the percentage of failing examinees for two of the panels, and no change in the percentage of failing examinees for one panel. The smallest increase was a change of 2.7% for Year 2 Panel 2, and the largest change was an 8.1% increase in failing examinees in Year 2 Panel 1. For the two panels that displayed lower percentages of failing examinees following review of the data, one decreased by 3.5% (Year 1 Panel 3) and one by 2.6% (Year 2 Panel 3). While the cut score increased slightly after review of data in Year 1 Panel 1, the change (from 60.19 to 61.17) did not result in any change to the 4.2% fail rate.

For Step 3, the results were quite different from those for the other two examinations. The fail rate for Year 1 increased for two panels (Panels 1 and 3) and stayed the same for one panel (Panel 2) following review of the data. The two panels for which the fail rates increased had failing percentages that were among the lowest across all examinations (1.2% and 2.0% before data and 3.3% for both following data, respectively), and the increases were similarly small, increasing 1.3% for Panel 3 and 2.1% for Panel 1). The most notable finding was that all three panels in Year 2 had decreased fail rates following data review. In both Panel 1 and Panel 2, the fail rate resulting from the initial judgments was 58.5%; this fail rate dropped to 22.8% and 48.5% for Panels 1 and 2, respectively, following the review of data. Panel 3 displayed a more modest initial fail rate of 16.6% which decreased to 11.3% following the review of data.

Table 5 presents the fail rate results collapsed across panels within each standard setting for each examination. For four of the six standard settings, judgments made following the review of performance data resulted in a higher fail rate than did those made in the absence of data; these changes ranged from an increase of 1.3% for Step 1 Year 1 to an increase of 4.3% for Step 1 Year 2. Only one standard setting had a final percentage of examinees failing that was the same as that which resulted from the initial judgments: Step 2 Year 2 judgments would have failed 8.8% of examinees regardless of whether or not the judges reviewed examinee performance data. Finally, the Step 3 results for Year 2 are the only ones in which review of the performance data led to a decrease in the percentage of examinees failing from 39.1% to 22.8%. While the changes in fail rate may seem modest in many cases, fail rates generally were low overall; even seemingly modest changes between initial and final cut scores therefore are likely to represent practically significant increases in the percentage of examinees failing.

Discussion

The present research provided a large-scale investigation of the impact of examinee performance information on the outcomes of modified Angoff standard-setting exercises. Some researchers have questioned the impact of providing performance data, suggesting that resulting cut scores are less likely to be impacted than is panelist variability (Hambleton, 2001).

Other researchers have suggested that cut scores generally were lower after judges reviewed data (Hurtz & Auerbach, 2003). Results of the present research provided no support for a general decrease in cut scores after judges reviewed performance data. Although there was no significant general trend across the multiple data sets, there was clear evidence that increases in cut scores are not uncommon after judges review performance data. Specifically, the results indicate that both panelist variability and resulting cut scores were affected by the data. In general, panelist variability decreased after judges reviewed performance data. In addition, post-data cut scores were significantly different from those in the pre-data condition; these differences were observed for each of the three examinations for both years. Fail rate changes were associated with the cut score changes for a majority of standard-setting years and provide additional evidence of the practical significance of the results. These significant changes for each of the six standard-setting exercises were not uniform in terms of the direction of the change; some cut scores increased after review of data and some decreased. This result is consistent with the fact that the test for the main effect across all panels, years, and examinations was not statistically significant.

It should be noted that, although the changes that result from providing judges with performance data were statistically significant and nontrivial in terms of impact, the differences in cut scores before and after provision of performance data were not particularly large when compared to other sources of variability in estimating cut scores. These differences tended to be of a similar magnitude as—or smaller than—the differences across panels within a single examination and standard-setting year. Although it is not the primary focus of this study, the results make it clear that variability across panels (within a single examination and year) is substantial. This result has significant implications for interpreting the results of standard-setting exercises that are based on a single panel. A detailed analysis of this effect is presented in Clauser, Margolis, and Clauser (2012).

One of the methodological strengths of the present research is that there were multiple replications of the same process both within and across different examinations. This aspect of the research design lends some confidence to the resulting conclusions, but the fact remains that only one specific standard-setting method was studied and only one procedure that included two specific types of performance data was used. Making broad generalizations about the present findings—particularly in the context of differences in standard-setting procedures or methodologies—therefore should be done with caution. It also is worth noting that this study compares results after item review without performance data to the subsequent results after data were presented. As one reviewer pointed out, it is possible that after completing initial review of the item set judges may have provided systematically different results if they were asked to rereview the items without performance data. (This is different than the assertion that the changes reported in this paper arose from sampling error—the individual significance tests performed for each of the standard-setting years rules this out.) A different study would be necessary to separate the effects of reviewing the items from those of reviewing the items after receiving performance data. A substantial effect resulting from simple rereview—without performance data—seems unlikely, but it cannot be ruled out.

A follow-up to the question of whether cut scores changed is one of *how* they changed and whether it was in a systematic or predictable way. Results of previous studies dedicated to investigating this topic vary; one meta-analysis concluded that cut scores following data review generally were lower than those from the pre-data condition (Hurtz & Auerbach, 2003). Again, the present results suggest that, although significant effects are apparent, it seems most reasonable to assume that the direction of the effect is likely to vary based on the specifics of the assessment and other unpredictable characteristics of the specific implementation. Again, for 12 panels the final cut scores were higher than the initial cut scores and for the remaining six panels the cut scores decreased after data review. Cut scores consistently increased across all of the Step 1 panels, while for the other two examinations a systematic pattern of increasing or decreasing final cut scores was not found. One factor that may contribute to the systematic change for the Step 1 panels has to do with the realities of the Step 1 testing scenario. Step 1 is a test of basic science knowledge and most commonly is taken by U.S. medical students in or around their second year of study. The panelists who participate in standard-setting activities are practicing physicians, and they have not studied the specific test content nor taken Step 1 or its equivalent for a minimum of five years (and, as is more realistically the case, for a decade or more). As a result, they often find the test material somewhat difficult, and this leads them to overestimate the difficulty for minimally proficient examinees. When they review the data, often it is the case that a much higher proportion of all examinees than expected get the items correct. This is for a number of reasons. First, in contrast to the panelists who have not studied basic science in years, Step 1 examinees have spent the two years prior to taking the examination studying the specific material that is covered on the examination. In addition, they also spend several months specifically preparing for the Step 1 examination itself. As a result, their knowledge of the material may never be more extensive than it is at that point. When panelists are reminded of the timing and test preparation issues, it generally results in an upward adjustment to their estimates. Test preparation also occurs for the other step examinations, but it is most intensive for Step 1. This might suggest that the judges' familiarity with the test material and examinee population impacts how they use the performance data, and this likely is true. It should, however, be noted that, whereas the particular conditions of the Step 1 examination may have impacted the direction of the change after reviewing performance data, the results provide less evidence that those conditions impacted the magnitude of the change.

Unfortunately, no similarly simple explanation is available to explain the patterns of change for the Step 2 and Step 3 results. It is worth noting that for both years of the Step 2 results and for the first year of the Step 3 results the panel with the highest initial cut score lowered its cut score after reviewing performance data and the panels with lower initial cut scores raised them. This apparent convergence may result from the fact that most panelists already know the current fail rates for the U.S. medical student population; when they do not, the question is almost always asked during the course of the initial training. It may be that the changes represent the panelists converging on what they view as plausible (or acceptable) fail rates given their prior knowledge.

Although it is fine for judges to have information about past fail rates, it could be argued that it should be of little relevance for the purposes of the content-based standard-setting exercises. Their role as judges is to review test content and to make decisions about that content based on expectations for minimally proficient examinees; there is no expectation that new standards will be similar to old standards, and in fact the judges' input would not be necessary if the desire was to maintain existing fail rates. Despite the fact that the panelists are explicitly instructed that they are to make content-based judgments, some panelists may have used the data to make decisions that were in line with their expectations or with existing statistics. This clearly would not be a concern in new examination contexts (without historical fail rates) or in contexts where there is no expectation that fail rates will be similar from one administration to the next. In the present context, however, the extent to which judges allowed the knowledge of current fail rates to influence their judgments directly impacts the extent to which the judgments were truly content-based (as intended) rather than norm-referenced.

Convergence of judgments within panels provides some evidence that post-data changes were not random and instead that judges interpreted and used the data similarly when revising their initial judgments. This suggests that they used the data to bring results into some general correspondence with prior expectations about fail rates. This view is consistent with the fact that the most substantial decrease in estimated cut scores occurred for the second Step 3 standard setting, which also had the initial fail rate that was most out of line with historic fail rates for that examination.

The specifics of how the judges use the data is beyond the scope of this research, but previous studies based on related data sets suggest that they bring their judgments into close correspondence with empirical item difficulties (Clauser, Mee et al., 2009; Clauser, Mee, & Margolis, 2013; Mee, Clauser, & Margolis, 2013). These studies either used subsets of the data included in the studies presented in this paper or similar data sets from the same examination system. These two types of convergence (the convergence with empirical item difficulties reported in previous studies and the convergence of estimated cut scores across panels reported in this study) are to some extent independent. The fact that judges modify their judgments so that they rank order similarly to the empirical item difficulties does not imply that they will produce more similar cut scores.

Existing research on the topic of procedural modifications in Angoff-style standard-setting exercises is plentiful but has been somewhat limited in scope, generally reporting on results from a small number of standard-setting exercises without replication based on multiple panels; this limited scope impacts the ability to interpret the findings. The structure of this study, which included 18 panels across three different examinations, provides the most robust evaluation of the practice of providing performance data that has been reported in the literature. Results of this large-scale investigation make it clear that providing standard-setting judges with performance data makes a difference. The extent to which this practice can be supported ultimately will depend on more detailed research on how judges use performance data and how this use impacts the validity of the interpretations made based on the standard-setting results. Whether or not such use ultimately is considered appropriate, this study makes it clear that performance data certainly can make a difference.

Acknowledgments

The present research grew out of the first author's doctoral dissertation work. I would like to thank Dr. Joseph DuCette, Professor of Educational Psychology at Temple University, for his guidance and support through the dissertation process.

References

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.

Brandon, P. R. (2004). Conclusions about frequently studied modified Angoff standard-setting topics. *Applied Measurement in Education*, *17*, 59–88.

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.

Cizek, G. J. (2012). *Setting performance standards: Concepts, methods, and perspectives*. Mahwah, NJ: Lawrence Erlbaum.

Clauser, B. E., Harik, P., Margolis, M. J., McManus, I. C., Mollon, J., Chis, L., & Williams, S. (2009). Empirical evidence for the evaluation of performance standards estimated using the Angoff procedure. *Applied Measurement in Education*, *22*(1), 1–21.

Clauser, B. E., Mee, J., Baldwin, S. G., Margolis, M. J., and Dillon, G. F. (2009). Judges' use of examinee performance data in an Angoff standard-setting exercise for a medical licensing examination: An experimental study. *Journal of Educational Measurement*, *46*, 390–407.

Clauser, B. E., Mee, J., & Margolis, M. J. (2013). The effect of data format on integration of performance data into Angoff judgments. *International Journal of Testing*, *13*, 65–85.

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., & Clauser, B. (2012, April). *An examination of the replicability of Angoff standard setting results within a generalizability theory framework*. Paper presented at the meeting of the National Council on Measurement in Education, Vancouver, BC, Canada.

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

Hambleton, R. K. (2001). Setting performance standards on educational assessments and criteria for evaluating the process. In G. Cizek (Ed.) *Standard setting: Concepts, methods and perspectives* (pp. 159–173). Mahwah, NJ: Lawrence Erlbaum.

Hurtz, G. M., & Auerbach, M. A. (2003). A meta-analysis of the effects of modifications to the Angoff method on cutoff scores and judgment consensus. *Educational and Psychological Measurement*, *63*, 584–601.

Kane, M. T. (1992). An argument-based approach to validity. *Psychological Bulletin*, *112*, 527–535.

Kane, M. (1994). Validating the performance standards associated with passing scores. *Review of Educational Research*, *64*, 425–461.

Kane, M. (2001). So much remains the same: Conception and status of validation in setting standards. In G. Cizek (Ed.) *Standard setting: Concepts, methods and perspectives* (pp. 53–88). Mahwah, NJ: Lawrence Erlbaum.

Kane, M.T. (2006). Validation. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 17–64). Westport, CT: American Council on Education/Praeger.

Mee, J., Clauser, B. E., & Margolis, M. J. (2013). The impact of process instructions on judges' use of examinee performance data in Angoff standard setting exercises. *Educational Measurement: Issues and Practice*, *32*(3), 27–35.

Plake, B. S. (2008). Standard setters: Stand up and take a stand! *Educational Measurement: Issues and Practice*, *27*(1), 3–9.

Plake, B. S., Impara, J., & Potenza, M. (1994). Content specificity of expert judges in a standard setting study. *Journal of Educational Measurement*, *31*, 339–347.

Truxillo, D. M., Donahue, L. M., & Sulzer, J. L. (1996). Setting cutoff scores for personnel selection tests: Issues, illustrations, and recommendations. *Human Performance*, *9*, 275–295.

