The Effect of Numbers of Experts and Common Items on Cutting Score Equivalents Based on Expert Judgment

John Norcini, Judy Shea, and Louis Grosso
American Board of Internal Medicine

The effect of different numbers of experts and common items on the scaling of cutting scores derived by experts' judgments was investigated. Four test forms were created from each of two examinations; each form from the first examination shared a block of items with one form from the second examination. Small groups of experts set standards on each using a modification of Angoff's (1971) method. Cutting score equivalents were estimated for the matched forms using different group sizes and numbers of common items; they were compared with cutting score equivalents based on score equating. Results showed that a reduction in error is associated with using more experts or having more items in common between the two forms. For 25 or more common items and five or more judges, the error was about one item on a 100-item test. More than five experts or 25 common items made only a very small difference in error. *Index terms: cutting scores, equating, expert judgment, standard setting.*

Many certification and licensure testing programs require that different forms of the same examination be developed and administered over time. Because the forms are not usually equal in difficulty, scores are often equated so that examinees are not at an unfair advantage or disadvantage because of the form they are assigned. Among the scores equated is the cutting score or pass/fail point; application of the equated cutting score to the later examination(s) ensures equivalence of standards over time.

When the cutting score is derived by one of the judgmental standard-setting methods (e.g., Angoff, 1971), it is possible to ensure cutting-score equivalence across forms by applying one of the usual equating designs to experts' judgments rather than examinees' scores. A recent study demonstrated that cutting-score equivalents based on experts' standard-setting judgments for several sets of items were closer to the criterion of examinees equated to themselves than standards based on equating examinees' scores on the same sets of items (Norcini, 1990). This effect was most pronounced when the number of examinees used in equating was small.

In the Norcini (1990) study, the number of experts was fixed at 10 and the number of common items was fixed at 93. However, the number of experts has been found to have a significant impact on the reproducibility of a standard (Norcini, Lipner, Langdon, & Strecker, 1987). Similarly, the number of common or anchor items has been found to have an impact on the quality of equating (Skaggs & Lissitz, 1986).

The purpose of this study was to determine the effect of numbers of experts and common items on the scaling of cutting scores derived by experts' judgments. In particular, data from two medical certifying examinations administered two years apart were reanalyzed. Items from each examination were divided into four forms. Each of the four forms from the first year has a block of items in common with one of the four forms from the second year. Experts for each year were divided into four groups, and they set standards on one of the forms using a modification of Angoff's standard-setting method (Norcini, Shea, & Ping, 1988).

APPLIED PSYCHOLOGICAL MEASUREMENT
Vol. 15, No. 3, September 1991, pp. 241-246

0146-6216/91/030241-06$1.55

Cutting score equivalents were estimated for each of the four forms using different expert group sizes and different numbers of common items. These estimates were then compared with the cutting-score equivalents based on score equating that used examinees' performance on the same forms with the same sets of common items. The configuration of all data was consistent with Angoff's (1971) Design IV for equating in which groups of experts and examinees are nonrandom, and the tests contain a block of common items. An equating procedure attributed to Tucker (Gulliksen, 1950) was applied to obtain cutting-score equivalents that were based on experts' judgments and examinees' test performance.

Method

Tests

Standard-setting data were used from an application of a variation on Angoff's (1971) method to two certifying examinations in medicine. Before the meetings in which standards were set, all items from each examination were divided into four forms [i.e., Forms A1, B1, C1, and D1 for the first year (called the old forms), and Forms A2, B2, C2, and D2 for the second year (called the new forms)]. These forms were constructed with the restriction that each form be comprised of an approximately equal number of questions of each item type (one-best-answer and multiple true/false) that represented each primary content area within the table of specifications. Initially, old forms consisted of 189 items, whereas new forms consisted of 194 items. The number of items in common was 73 for Forms A1 and A2, 71 for forms B1 and B2, 67 for forms C1 and C2, and 73 for forms D1 and D2. A small number of items was randomly deleted so that each pair of forms had 65 items in common and all forms were composed of 180 items.

Raters and Rating Procedures

There were 40 expert raters for the first examination and 38 expert raters for the second examination. All were physicians certified in the specialty of the examination. Prior to the standard-setting meeting, the raters were divided quasi-randomly into four groups (i.e., expert groups GA1, GB1, GC1, and GD1 for the old forms, and groups GA2, GB2, GC2, and GD2 for the new forms). All group sizes were reduced to 9 raters; for each of the two years, three of the pairs of groups had one member in common. Before the standard-setting meeting, the experts were sent material describing the group process, although many were already familiar with Angoff's method and had actually used it previously.

For each year of examination data, the four groups met separately. Each session began with a statement of the purpose of the meeting, a description of the rating process, and a discussion of the characteristics of the borderline group of examinees (i.e., a group that is neither clearly qualified nor unqualified for certification). This was followed by individual judgments of the proportion of the hypothetical borderline group of physicians that would respond correctly to the first item. The judgments were written on a blackboard and the experts with the extreme estimates justified their decisions. This was followed by a general discussion in which all participants were free to change their estimates. Performance data for a borderline-like group were available to serve as a reference point for those who chose to use it.

Each group followed this procedure on the first examination for the same 49 items; the experts took the remaining 140 items and the instructions with them and completed the ratings at home. For the second examination, the experts rated the same 45 items in the group setting and completed the remaining 149 items at home. Earlier research had demonstrated that estimates gathered after a meeting produced cutting scores similar to those gathered before or during a meeting (Norcini, Lipner, Langdon, & Strecker, 1987).

Criteria

To judge the accuracy of the scaling using experts' estimates, a criterion for each of the new

forms was developed through traditional equating of examinees' scores. For the four sets of forms, the examinee scores on the new form were equated to the examinee scores on the old form. All old forms were taken by the same 5,701 examinees and all new forms were taken by the same 6,216 examinees; examinees were limited to those taking the test for the first time. Scores were equated by applying Tucker's linear procedure (Gulliksen, 1950) for groups not widely different in ability. The transformation that resulted from this process was applied only to the cutting scores generated by the first year's expert groups, and it produced the four criterion equatings. The cutting scores were expressed on the proportion-correct scale.

Variations in Numbers of Experts and Common Items

The items that were common to the four form-pairs (i.e., Forms A1 and A2, Forms B1 and B2, Forms C1 and C2, and Forms D1 and D2) constituted the basis for varying the number of common items. For these analyses, random sets of 5, 25, 45, and all 65 items were drawn from the common items of each of the four form pairs.

The four groups that derived cutting scores for the old forms and their four counterparts for the new forms constituted the basis for varying the number of experts. From each of these eight expert groups, random samples of 3, 5, 7, and all 9 experts were drawn with the restriction that the groups of 3, 5, and 7 would not have an expert in common. For each of the eight groups, the experts' estimates were summed, averaged across the experts in the group, and divided by the number of items in order to produce cutting scores on the proportion scale.

Design

When the four different numbers of common items were crossed with the four different form pairs and the four different group sizes, 64 unique combinations were available for analysis (4 numbers of common items × 4 form pairs × 4 group sizes). This design is displayed in Table 1. For each unique combination, it was assumed that Group I set the base standard on the old form. The goal was to ensure an equivalent standard on the new form. The purpose of Group II was simply to provide a means for producing a cutting-score equivalent. Consequently, the cut-

Table 1
Study Design: Form Combinations Used by Group I and Group II for 3 to 9 Experts

Form Pair		Number of Common Items	Number of Experts							
			3		5		7		9	
Old	New		I	II	I	II	I	II	I	II
A1	A2	5	GA1	GA2	GA1	GA2	GA1	GA2	GA1	GA2
A1	A2	25	GA1	GA2	GA1	GA2	GA1	GA2	GA1	GA2
A1	A2	45	GA1	GA2	GA1	GA2	GA1	GA2	GA1	GA2
A1	A2	65	GA1	GA2	GA1	GA2	GA1	GA2	GA1	GA2
B1	B2	5	GB1	GB2	GB1	GB2	GB1	GB2	GB1	GB2
B1	B2	25	GB1	GB2	GB1	GB2	GB1	GB2	GB1	GB2
B1	B2	45	GB1	GB2	GB1	GB2	GB1	GB2	GB1	GB2
B1	B2	65	GB1	GB2	GB1	GB2	GB1	GB2	GB1	GB2
C1	C2	5	GC1	GC2	GC1	GC2	GC1	GC2	GC1	GC2
C1	C2	25	GC1	GC2	GC1	GC2	GC1	GC2	GC1	GC2
C1	C2	45	GC1	GC2	GC1	GC2	GC1	GC2	GC1	GC2
C1	C2	65	GC1	GC2	GC1	GC2	GC1	GC2	GC1	GC2
D1	D2	5	GD1	GD2	GD1	GD2	GD1	GD2	GD1	GD2
D1	D2	25	GD1	GD2	GD1	GD2	GD1	GD2	GD1	GD2
D1	D2	45	GD1	GD2	GD1	GD2	GD1	GD2	GD1	GD2
D1	D2	65	GD1	GD2	GD1	GD2	GD1	GD2	GD1	GD2

ting scores from the old forms were adjusted by applying Tucker's linear procedure for groups not widely different in ability. These adjusted cutting scores were compared to the criteria.

Results

Descriptive Data

Table 2 provides data concerning the performance of examinees on each form. Examinees' means for Forms A1 to D1 ranged from .713 to .736 with standard deviations (SDs) ranging from .080 to .086. Because the examinees taking these forms were the same, the forms differed in difficulty by as much as .25 SD. Means for Forms A2 to D2 ranged from .729 to .738, and showed more similarity in form difficulty; SDs ranged from .088 to .091. The generalizability coefficients $E_{\hat{p}}^2$ (Brennan, 1983) for all forms ranged from .86 to .89, which demonstrates that the scores from the forms had similar and reasonably good generalizability.

Table 2 also provides data concerning the cutting scores (means) that were based on the experts' estimates of borderline group performance. Most cutting scores were between .25 and .50 SD below the examinees' means, which implies pass rates of roughly 60% to 75% on each of the forms. The cutting scores across the forms varied considerably, given the magnitude of the SDs. However, they were in about the same rank order as the means of examinees, which suggests that the experts were sensitive to the difficulty of the forms when they made their judgments.

Table 2
Form Means, SDs, and Generalizabilities ($E_{\hat{p}}^2$) for Examinees and Experts

	Examinees			Experts	
Form	Mean	SD	$E_{\hat{p}}^2$	Mean	SD
A1	.718	.085	.87	.652	.035
B1	.713	.086	.87	.679	.019
C1	.736	.080	.86	.716	.021
D1	.725	.081	.86	.707	.039
A2	.738	.088	.88	.711	.019
B2	.737	.091	.89	.713	.021
C2	.738	.088	.88	.712	.012
D2	.729	.089	.88	.692	.017

The Effect of Different Numbers of Experts and Common Items

To estimate the differences between the cutting-score equivalents that were based on score equating and those based on scaling of experts' judgments, each standard was subtracted from the appropriate criterion and squared, and the square root was then calculated. These results are shown in Table 3 for each of the 64 unique combinations. In general, the differences are relatively small: they range from 0 to .026, with 61% being less than .01.

For each number of common items (5, 25, 45, and 65) within a form pair, the root mean square error (RMSE) was calculated. The differences between the cutting-score equivalents and the criteria for the different numbers of experts within a particular number of common items were squared, summed, and averaged, and the square root was then calculated. The same procedure was followed to calculate the RMSEs for different numbers of common items within a particular number of experts and for a specific form pair across numbers of experts. Overall, the errors were small—they ranged from .003 to .019. With some exceptions, they tended to be lower with larger numbers of experts or common items.

Table 4 provides RMSEs over all form-pairs for each combination of the four numbers of experts and common items. For 25 or more common items, or 5 or more judges, the error was .011 at most, or approximately one item on a 100-item test. Having more than five experts made only a very small difference in the RMSE. Likewise, having more than 25 common items had only a small effect on RMSE.

Discussion

The results of this study indicated that some reduction in error is associated with using more experts or with having more items in common between the two forms. For 25 or more common items and five or more judges, the error was equivalent to approximately one item on a

Table 3
RMSEs for Combinations of Numbers of Experts and Items

Form Pair		Number of Common	Number of Experts				
Old	New	Items	3	5	7	9	Combined
A1	A2	5	.012	.001	.008	.010	.009
A1	A2	25	.021	.013	.014	.012	.016
A1	A2	45	.005	.004	.002	.003	.003
A1	A2	65	.008	.007	.004	.004	.006
Combined			.013	.008	.008	.008	
B1	B2	5	.023	.021	.019	.013	.019
B1	B2	25	.010	.004	.006	.007	.007
B1	B2	45	.002	.003	.007	.002	.004
B1	B2	65	.003	.004	.004	.002	.003
Combined			.013	.011	.011	.007	
C1	C2	5	.014	.017	.003	.005	.011
C1	C2	25	.008	.003	.005	.000	.005
C1	C2	45	.023	.011	.009	.015	.015
C1	C2	65	.026	.009	.007	.015	.016
Combined			.019	.011	.006	.011	
D1	D2	5	.025	.020	.004	.010	.017
D1	D2	25	.013	.014	.008	.007	.011
D1	D2	45	.010	.004	.004	.015	.009
D1	D2	65	.006	.003	.003	.016	.009
Combined			.015	.012	.005	.013	

100-item test. Moreover, having more than five experts or more than 25 common items made only a very small difference in the RMSE.

These results are consistent with prior work (Norcini, 1990), which indicates considerable agreement between cutting-score equivalents that are based on score equating and those based on scaling expert judgments. It expands on this prior work by indicating that five or more judges or 25 or more common items are sufficient to achieve reasonably precise results. On the test forms used in this study, the difference of about 1% of the items was less than 15% of 1 SD.

Although the results are encouraging, they must be considered as preliminary. Experts were assigned to groups in a quasi-random fashion, and although each group met separately, this method might be atypical and thus limit the generalizability of the results. Also, the experts in this study were involved at some point in the test development process. This may have led them to be relatively consistent in their estimates. Therefore, a larger number of experts might be required in other disciplines or under different circumstances. Moreover, inclusion of a larger number of experts is often needed in order to reduce variability and to represent the various viewpoints necessary to give the standard credibility.

Despite such limitations, this study demonstrated that the scaling of cutting scores with the use of expert judgment produces relatively precise results with only modest numbers of experts and common items. This scaling method has been shown to produce standards closer to

Table 4
RMSEs for Different Numbers of Experts and Common Items

Number of Common	Number of Experts				
Items	3	5	7	9	Combined
5	.019	.017	.011	.010	.015
25	.014	.010	.009	.008	.010
45	.013	.006	.006	.011	.009
65	.014	.006	.005	.011	.010
Combined	.015	.011	.008	.010	

criterion than score equating when the number of examinees is small (Norcini, 1990); it should also yield standards closer to a criterion than score equating when the cutting score is relatively extreme. The latter is a reasonable expectation because the cutting score for many mastery examinations will be at or near the mean of the experts' estimates and will be relatively far from the mean of examinees.

In addition to examinations with extreme cutting scores, there are several other conditions in which score equating might not be possible and might not produce very precise results. Examples include conditions in which testing time is limited, the entire form of a test is released to the public, and performance tests and simulations in which very few items are used. In each of these instances in which a sufficient number of internal anchor items cannot be administered, the scaling of cutting scores permits the use of an external anchor without requiring examinee time and its attendant administrative costs.

Only one design was used here for collecting the scaling data from experts, and a single linear method was applied to the data. There are other designs and equating methods that might produce cutting-score equivalents that are more precise. Moreover, the same adjustment can also be applied to the estimates of borderline group performance for individual items. This permits construction of an item pool in which each question has an equivalent borderline group value associated with it. All test forms derived from such a pool would then have equivalent standards.

Score equating was used here as the criterion against which to judge this method of generating cutting-score equivalence. It is important to note that the two methods produced similar results. Aside from precision, however, there is no theoretical reason to prefer one method over the other.

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.

Brennan, R. L. (1983). *Elements of generalizability theory.* Iowa City: The American College Testing Program.

Gulliksen, H. (1950). *Theory of mental tests.* Hillsdale NJ: Erlbaum Associates.

Norcini, J. J. (1990). Equivalent pass/fail decisions. *Journal of Educational Measurement, 27,* 59–66.

Norcini, J. J., Lipner, R. S., Langdon, L. O., & Strecker, C. A. (1987). A comparison of three variations on a standard-setting method. *Journal of Educational Measurement, 24,* 56–64.

Norcini, J. J., Shea, J. A., & Ping, J. C. (1988). A note on the application of multiple matrix sampling to standard-setting. *Journal of Educational Measurement, 25,* 159–164.

Skaggs, G., & Lissitz, R. W. (1986). IRT test equating: Relevant issues and a review of recent research. *Review of Educational Research, 10,* 495–529.

Acknowledgments

The authors gratefully acknowledge the comments of the anonymous reviewers. This research was supported by the American Board of Internal Medicine but does not necessarily reflect its opinions or policies.

Author's Address

Send requests for reprints or further information to John Norcini, American Board of Internal Medicine, 3624 Market Street, Philadelphia PA 19104, U.S.A.