PR-84-O

3/19/84

EXPLANATION OF SCALING ON THE CALIFORNIA BAR EXAMINATION

OVERVIEW

This paper describes how scores on the Multistate Bar Examination (MBE) are adjusted ("scaled") to control for possible variations in average question difficulty across different administrations of the exam, such as between July 1982 and July 1983. The paper further describes how the MBE scale scores are used to adjust scores on the multiple choice portions of the Performance Test (PT) so that the PT's scores also are not affected by possible variations in average PT item difficulty from one administration of the exam to the next. An understanding of the scaling of PT multiple choice scores therefore requires an understanding of the principles and rationale underlying the scaling of MBE scores.

SCALING MBE SCORES

The MBE contains 200 multiple choice questions (or "items"). Most, but not all, of the items asked on one administration of the MBE (such as July 1983) have not been asked before. Thus, differences in average MBE raw scores (i.e., the number of items answered correctly) between two administrations of the exam could be due to two factors:

1) differences in the average difficulty of their unique items; e.g., the questions that were only asked on the July 1983 exam were on the average more difficult than the those asked on the July 1982 exam.

2) differences in the average ability level of the applicants taking each test, such as normally occurs between the July and February administrations of the exam

A standard statistical technique called "scaling" is used to make sure the MBE scores that are reported for applicants are not affected by either of these two factors. This process involves the following basic steps:

1) Include in the batch of 200 items that are asked on a given administration of the MBE a large group of items called "equators." Equators are items that appeared on a previous version of the MBE and are highly representative of the other items that appeared on that previous version.

2) Determine how well the applicants taking the current version of the MBE performed on the equators relative to how well the applicants who took the previous version also performed on these same items.

3) Adjust the total raw scores on the current version so that a given total adjusted (scaled) score indicates the same level of skill and knowledge regardless of the particular version of the test that was taken.

The principles underlying the complex statistical procedures used for doing the scaling are illustrated by the following hypothetical example. Suppose the applicants taking the July 1983 MBE had an average score of 70 percent on the equators; i.e., on the average, 7 out of 10 July 1983 applicants answered an equator correctly. The percentage of applicants answering an item correctly is an indication of its difficulty.

Suppose further that the applicants taking the July 1982 MBE also had an average score of 70 percent on these same items. In other words, when the July 1982 and 1983 groups had a common set of items, they had the same average score on these items. If this occurred, we would expect the two groups to do equally well on their respective sets of unique items (i.e., the items that appeared on only one administration of the exam); provided, of course, that the two sets of unique items were equally difficult.

Thus, if the July 1982 applicants had an average score of 75 percent on their unique items and the July 1983 applicants had an average score of only 70 percent on their unique items, we would have to infer that the July 1983 unique items were, on the average, more difficult than the July 1982 unique items. After all, when the two groups were asked the same items (i.e., the equators), they performed equally well. Thus, the difference between the groups in their average scores on their respective sets of unique items must have been due to differences in the average difficulty of those items.

If this hypothetical situation had actually occurred, it would not be fair to compare one applicant's score on the July 1982 version of the exam with another applicant's score on the July 1983 version because the items in the latter version were, on the average, more difficult than those asked on the July 1982 version. To correct for this unfairness when MBE scores are used to make pass/fail decisions, the July 1983 scores would have to be scaled up relative to the July 1982 scores.

The July 1982 test also could have been more or less difficult than the February 1982 exam, the July 1981 exam, and so on back to the base year of July 1972. Because of this, the scores on each exam are adjusted relative to the base year. Thus, if an applicant answered 122 items correctly on a test that was easier than the July 1983 exam but harder than the base test, that applicant might be assigned a California scale score of 390 whereas 122 items correct on the July 1983 exam was assigned a California scale score of 407. The formula used for converting July 1983 raw scores to California scale scores was:

$$\text{California MBE scale score} = 2.6460\ (\text{MBE raw score}) + 83.9841$$

Inserting various values of raw score into the equation above reveals that all applicants benefited from the scaling. For example, if an applicant's MBE score was based on just the percentage of items answered correctly, then a 122 would convert to a 366 out of a possible 600 points ($122/200 \times 600 = 366$). Thus, an applicant with a 122 raw score earned 41 more total score points than this applicant would have earned if his or her MBE score was based on the percentage of questions answered correctly.

The basic statistical formulas for converting raw scores to scale scores through a series of linked equators are described by William Angoff in Chapter 15 of Educational Measurement (2nd edition). R. L. Thorndike (editor). Washington DC: American Council on Education, 1971.

The scaling procedures for the MBE assume that the applicants who take a given version of the exam have not had prior access to that exam's set of equators. For example, if the July 1983 exam had used equators from the July 1982 test, then it is assumed that the July 1983 applicants have not had an opportunity to study the equating items (and learn the answers to them) prior to their taking the July 1983 exam. This assumption requires that a given version of the MBE be released to the public only after it has been determined that none of its items will be used for equating future versions of the exam. A breach in the equators' security would have the effect of increasing the scale scores on the most recent version of the exam.

SCALING THE PT MULTIPLE CHOICE SECTIONS

The purpose of scaling the multiple choice scores on the PT is the same as the purpose for scaling MBE scores, namely: to control for possible variations in average item difficulty from one exam to the next. However, the method used for scaling MBE scores cannot be used with the PT because PT items from a previous version of the exam cannot be repeated on the current version. Thus, some other method for scaling PT multiple choice scores had to be used.

The method selected for scaling the July 1983 PT multiple choice scores is analogous to converting meters to yards and involved the following steps:

1) Compute the average score (mean) and standard deviation (SD) for each of the following three measures: a) raw score on the multiple choice portion of the morning PT, b) raw score on the multiple choice portion of the afternoon PT, and c) California MBE scale scores.

2) Convert all the scores on the multiple choice portion of the morning PT to a distribution that had a mean and standard deviation that were equal to one sixth the size of these parameters in the distribution of California MBE scale scores (because each PT multiple choice section was assigned 100 points and the MBE was assigned 600 points).

3) Repeat step 2 for the afternoon PT.

The simplified form of the formulas that were used to convert PT raw scores to scale scores appear below. Note that a given applicant's PT score was not adjusted in terms of that applicant's MBE score, but rather, in terms of the total distribution of California MBE scale scores among all the applicants that took both the MBE and the PT.

PT-AM Scale Score = (4.5088 x AM raw score) + 20.9988

PT-PM Scale Score = (3.4692 x PM raw score) + 32.4314

The decision was made to adjust scores on the PT's multiple choice sections to a score distribution that was based on the MBE because:

1) There was no way of knowing in advance of the exam whether the PT multiple choice items were unduly easy or difficult and, with only a total of 30 items, a few of them could have a large affect on the passing rate. If the items turned out to be unusually difficult and if no scaling method was employed to control for this problem, then the passing rate on the exam would have been reduced significantly.

2) The distribution of MBE scale scores provided an objective and independent basis for determining the difficulty of the PT items because its scale scores were already adjusted for possible variations in average question difficulty from one exam to the next.

3) The procedures for scoring the multiple choice sections could be announced in advance of the administration of the exam and thereby avoid any suspicion that the Committee adversely influenced the percent passing.

4) Previous research had indicated that applicants who had high scores on the machine scorable portion of the PT's prototype also tended to have high MBE scores.

5) Scaling to the MBE rather than to the Essay was expected to increase the percent passing the whole exam because studies of past exams indicated that about 21 percent more applicants passed the MBE than passed the Essay and the mean written scores on prototype versions of the PT were no higher than mean Essay scores.

An analysis of the July 1983 data confirmed the assumptions on which the PT scaling was based. These analyses demonstrated that there was a highly statistically significant correlation between MBE scores and raw scores on the multiple choice portion of both PTs. Applicants who did well on the MBE also tended to do well on a PT multiple choice section. These analyses also revealed that scaling to the MBE (rather than to the Essay or the PT's written score) maximized the percent passing because the MBE was much easier than either the Essay or the PT written sections.

CONCLUSION

The purpose of scaling raw multiple choice scores is to adjust these scores for possible variations in average question difficulty from one exam to the next because most of the questions asked on one exam are not the same as those asked on a prior exam.

The net effect of scaling the July 1983 MBE and PT multiple choice scores was to increase the percent passing. For example, an applicant who had raw scores on the morning and afternoon PT multiple sections of 14 and 13, respectively, and an MBE raw score of 122 would earn a total of 569 points on these three sections. If this applicant's score had been based on the percentage of questions answered correctly, he or she would have received a score of 546 on these sections; i.e., 23 points less than the score assigned by the Committee of Bar Examiners. It also was discovered that scaling the PT multiple choice scores to the MBE produced a higher pass rate than would have been produced if the PT multiple choice scores had been scaled to either the Essay or PT written sections.