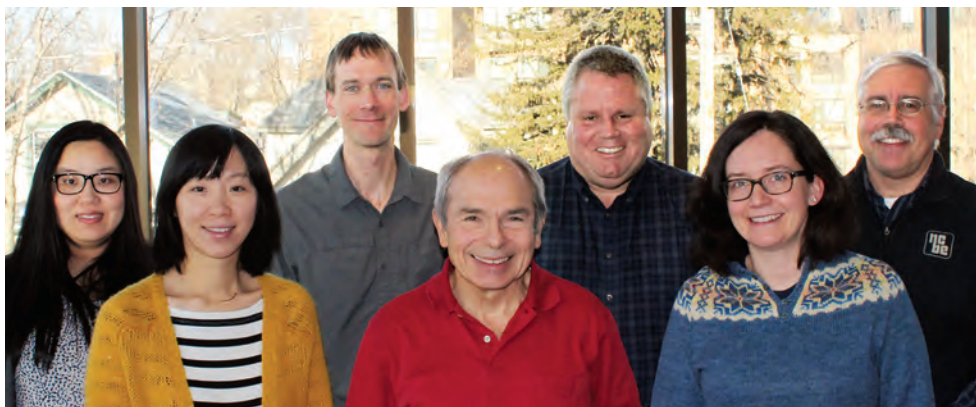


Q&A: NCBE Testing and Research Department Staff Members Answer Your Questions

BY NCBE Testing and Research Department

Members of NCBE's Testing and Research Department staff are periodically asked for concise responses to frequently asked questions. This article tackles seven such questions by providing brief answers on topics that have been covered more extensively in past issues of the Bar Examiner.



NCBE Testing and Research Department Staff:

Front row:

Juan Chen, Ph.D.;
Mark A. Albanese, Ph.D.;
Joanne Kane, Ph.D.

Back row:

Mengyao Zhang, Ph.D.;
Andrew A. Mroch, Ph.D.;
Mark Connally, Ph.D.;
Douglas R. Ripkey, M.S.

1 What Is Equating? Why Do It? How Does It Work?

Equating is a statistical procedure used for most large-scale standardized tests to adjust examinee scores to compensate for differences in difficulty among test forms so that scores on the forms have the same meaning and are directly comparable. (The term *test form* refers to a particular set of test items, or questions, administered at a given time. The February 2017 Multistate Bar Examination [MBE] test form,

for example, contains a unique set of items, and the July 2017 test form contains a different set of items.) With equating, a reported scaled score has a consistent interpretation across test forms.

Equating is necessary because using exactly the same set of items on each test form could compromise the meaning of scores and lead to unfairness for some examinees. For example, there would be no guarantee that items were not shared among examinees over time, which would degrade the

meaning of scores. Not only would the scores obtained by some examinees be corrupted by the examinees' advance knowledge of test items, but, if undetected, these inflated scores would advantage the clued-in examinees over other examinees. To avoid this problem, the collection of items on test forms for most large-scale standardized tests changes with every test administration.

Despite test developers' best efforts to build new forms that conform to the same content and statistical

specifications over time, it is nearly impossible to ensure that statistical characteristics like difficulty will be identical across the forms. Equating adjusts scores for such differences in difficulty among test forms so that no examinee is unfairly advantaged by being assigned an easier form or is unfairly disadvantaged by being assigned a more difficult form.

While there are many methods available to conduct equating, a commonly used approach for equating large-scale standardized tests—and the approach that is used for the MBE—is to embed a subset of previously administered items that serve as a “mini-test,” with the items chosen to represent as closely as possible the content and statistical characteristics of the overall examination. The items that compose this mini-test are referred to as *equators*. The statistical characteristics of the equators from previous administrations are used to adjust scores for differences in difficulty between the current test form and previous forms after accounting for differences between current and previous examinee performance on the equator items.¹ Conceptually, if current examinees perform better on the equators compared to previous examinees, we know that current examinee proficiency is higher than that of previous examinees and that current examinees’ MBE scaled scores should be higher (as a group) than those of previous examinees (and vice versa if performance on the equators is worse for current examinees). The equators are used as a link to previous MBE

administrations and, ultimately, to MBE scaled scores.

2 Can We Equate Essay or Performance Tests? (Or, What Is Essay Scaling?)

As with MBE items, the written components of the bar exam (essay questions and performance test items) change with every administration. The difficulty of the questions/items, the proficiency of the group of examinees taking the exam, and the graders (and the stringency with which they grade) may also change. All three of these variables can affect the grades assigned by graders to examinees’ responses to these written components of the exam and can have the potential to cause variation in the level of performance the grades represent across administrations. Unlike the MBE, the answers to the written questions/items of the bar examination cannot be equated, because previously used questions/items can’t be reused or embedded in a current exam—there are too few written questions/items on the exam and they are too memorable. If essay questions or performance test items were reused, any examinee who had seen them on a previous administration would be very likely to have an unfair advantage over examinees who had not seen them previously.

Because directly equating the written components is not possible, most jurisdictions use an indirect process referred to as *scaling the written component to the MBE*. This process has

graders assign grades to each question/item using the grading scale employed in their particular jurisdiction (e.g., 1 to 6). The individual grades on each written question/item are typically combined into a raw written score for each examinee. These raw written scores are then statistically adjusted so that collectively they have the same mean and standard deviation as do the scaled scores on the MBE in the jurisdiction. (*Standard deviation* is the measure of the spread of scores—that is, the average deviation of scores from the mean. The term *scaled score* refers to the score as it has been applied to the scale used for the test—in the case of the MBE, the 200-point MBE scale.)

Conceptually, this process is similar to listing MBE scaled scores in order from best to worst and then listing raw written scores in order from best to worst to generate a rank-ordering of MBE scores and written scores. The best written score assumes the value of the best MBE score; the second-best written score is set to the second-best MBE score, and so on. Functionally, the process yields a distribution of scaled written scores that is the same as the jurisdiction’s distribution of the equated MBE scaled scores. Another way to think about the process is that the raw written scores are used to measure how far each examinee’s written performance is from the group’s average written performance, and then the information from the distribution of the group’s MBE scores is used to determine what “scaled” values

should be associated with those distances from the average.

This conversion process leaves intact the important rank-ordering decisions made by graders, and it adjusts them so that they align with the MBE scaled score distribution. Because the MBE scaled scores have been equated, converting the written scores to the MBE scale takes advantage of the MBE equating process to indirectly equate the written scores. The justification for scaling the written scores to the MBE has been anchored on the facts that the content and concepts assessed on the MBE and written components are aligned and performance on the MBE and the written components is strongly correlated. The added benefit of having scores of both the MBE and the written component on the same score scale is that it simplifies combining the two when calculating the total bar examination score. In the end, the result of scaling (like equating) is that the scores represent the same level of performance regardless of the administration in which they were earned.

3 How Does Grader Leniency Affect a Jurisdiction's Pass Rate?

For jurisdictions that scale their written components to the MBE (as recommended by NCBE), grader leniency does not affect a jurisdiction's overall pass rate.² Because grades on individual written items are combined to create a raw written score that is then scaled to the MBE,

grader leniency (or stringency) is accounted for, thereby not affecting the jurisdiction's overall pass rate.³ Grades rank-order examinees, but pass rate is determined by the jurisdiction's cut score and the MBE scores for the group of examinees in a jurisdiction.

For example, if the mean raw written score in a jurisdiction is 50% of the points and the mean MBE score is 142, then the 50% will be converted to a 142 when the written component is scaled to the MBE. If graders are more lenient and the mean raw written score is 70% of the points, and assuming that the mean MBE score remains 142, the 70% will be converted to a 142 when the written component is scaled to the MBE.

While grader leniency (or stringency) does not affect a jurisdiction's overall pass rate, it could affect which examinees pass or fail. For example, suppose that two graders are assigned to grade the same essay question, and that one grader is lenient and the other is harsh (which would result from the graders not being properly calibrated). Further, suppose that the essays provided to the two graders are of equal quality. The essays evaluated by the harsh grader will receive lower scores and would be more likely to result in failure than those evaluated by the lenient grader, even though the overall pass rate remained the same. Failure to adequately calibrate graders is of consequence to individual examinees even though the overall passing rate for the jurisdiction is unaffected.

Jurisdictions should help ensure that examinee grades accurately reflect performance on each question by verifying that graders are properly calibrated and that they are using the entire available grading scale for each essay.⁴

4 If a Jurisdiction Changes the Weights of the Bar Exam Components, Will the Pass Rate Change?

If a jurisdiction scales its written component to the MBE, changing the weights of the MBE and the overall written component when calculating the total bar examination score (e.g., weighting the MBE 40% vs. 50% in relation to the written component) will not have a large impact on the pass rate for a given administration. Because scaling places the performance on the written component on the same (equated) score scale as the MBE, the means and standard deviations will be identical and how scores are distributed will usually be very similar. When two distributions with a similar shape (i.e., the same means and standard deviations) are combined (i.e., added together or averaged), the shape of the resulting distribution of scores will also be similar.

For example, if a jurisdiction's average MBE score is 140, the average written component score will also be 140 after the written component is scaled to the MBE. When those average scores are added together, the combined score will not vary based on the weight assigned to

each component. The following examples illustrate that the group's average for the combined score is 140 regardless of whether the MBE is weighted 40% or 50% of the combined score.

MBE weighted 40%
(40-60 weight):

$$(0.40 \times 140) + (0.60 \times 140) = \mathbf{140}$$

(weight of 40% times the average MBE score plus weight of 60% times the average written component score)

MBE weighted 50%
(50-50 weight):

$$(0.50 \times 140) + (0.50 \times 140) = \mathbf{140}$$

(weight of 50% times the average MBE score plus weight of 50% times the average written component score)

Note that to calculate the combined score for an individual examinee, the examinee's scores would be substituted for the group average scores in the equations shown above. As a result, an individual examinee's combined score will be affected by the weighting, but the overall average score for the jurisdiction and the overall pass rate for the jurisdiction won't change much. Examinees whose score is higher on the component with the greater weight will be more likely to pass the exam. For example, suppose an examinee received a score of 140 on the MBE and a score of 130 on the scaled written component. With the 40-60 weighting, the examinee's combined score would be 134, whereas it would be 135 with the 50-50 weighting. The result is that if the cut score was 135 in a jurisdiction, the examinee would fail by

the 40-60 weighting and pass by the 50-50 weighting, yet the percentage passing in the jurisdiction would be essentially the same under both weighting approaches.

5 Is the MBE (Only) a Test of Memorization?

Two content validity studies, conducted in 1980 and 1993, invited law practitioners and professors to evaluate (among other things) the extent to which MBE items placed greater emphasis on memorization or on analytic skills. Across those studies, subject matter experts indicated that the items placed roughly equal emphasis on memorization and legal reasoning skills across content areas—with the exception of Constitutional Law, where the panelists reported greater emphasis on reasoning skills.⁵

More recently, in research conducted by a recipient of NCBE's Covington Award,⁶ think-aloud protocols were used to measure the cognitive processes examinees used in responding to MBE items. (Think-aloud protocols involve having participants verbalize their thoughts as they perform a given task, thereby providing insight into the cognitive processes involved in performing the task.) The results revealed that efforts to recall isolated facts were a relatively small proportion of the cognitive processes examinees engaged in when responding to MBE items.⁷

In addition, correlational research has indicated that MBE scores are

related to other measures of legal reasoning ability, including the written portion of the bar examination,⁸ law school grades,⁹ and simulations of tasks such as questioning witnesses, negotiating a settlement, and preparing a brief.¹⁰ This correlational research indicates that regardless of the degree to which the MBE requires memorization, performance on the MBE is indeed associated with important legal skills.

Finally, NCBE's former president, Erica Moeser, was unequivocal in stating that requiring examinees to have working knowledge of and fluency in legal concepts is not an undue burden but a central part of ensuring fitness for practice:

One bit of rhetoric I have encountered lately seems to take issue with the MBE as a test of memory. For starters, that is not an apt characterization. We know that the MBE tests analysis and reasoning. There is a kernel of truth in the testing of memory, though, and that is an acknowledgment that the MBE does require knowledge that must be remembered when taking the test.

For me, there is no defensible argument that emergence from law school should occur without assurance that the graduate has acquired the knowledge required for entering the practice of law. While skills and values are certainly essentials, anyone who makes the case that baseline knowledge is not an essential for practice and not

appropriate for testing has got it all wrong.

The knock on memorization falters if it is a euphemism for excusing knowledge as an essential, or if it serves as an excuse for institutional or individual failure to acquire fundamental knowledge about the core subject matter that belongs at the threshold of a lifetime general license to practice law.¹¹

6 How Do We Know That the Change from 190 to 175 Scored Items on the MBE Didn't Hurt Scores?

Effective with the February 2017 bar examination, NCBE increased the number of pretest items on the MBE from 10 to 25, thereby reducing the number of live items from 190 to 175. (The unscored pretest items are questions whose performance will be evaluated for use on a future exam, and live items are those that contribute to an examinee's score.) A common concern with regard to the reduction from 190 to 175 scored items on the MBE is whether this change would make it harder to get a high score and pass the bar exam, as the scaled scores are now based upon 15 fewer items.

Scaled scores on the MBE are not constricted by having fewer items. The 190 or 175 raw score points will map to the same 200-point MBE scale. More importantly, by reducing the number of scored items, we were able to pretest 15 additional new or revised items per

exam booklet while maintaining the same total test length of 200 items and the same testing time. The pretest data provides us with important information concerning item performance, which facilitates construction of future MBE forms built more closely to statistical specifications.

Before implementing a change in the number of scored items, we modeled the impact of the reduction in the number of scored items using the MBE exams administered in 2015 and 2016. We found minimal effects on the MBE 200-point scaled scores. The scaled score mean and standard deviation were only affected at the second decimal point, indicating that overall performance was not affected by a reduction to 175 scored items.


The consistency with which examinees would pass or fail with a passing score of 135 on the MBE (the most commonly used passing score across the jurisdictions) was very high between the 190- and 175-item exams.

We also modeled the expected impact on the reliability of scores in 2015 and 2016 (reliability being the extent to which a group of examinees would be rank-ordered in the same way over multiple testing sessions) and found that we should expect no change or maybe even a slight increase. The reduction in the number of scored items would be offset by an ability to select better-functioning items because of the added pretest data and an ability to be more selective

because of the need for fewer items. This finding has been confirmed by both the February 2017 and July 2017 MBE administrations, when the reliability of scores either rose or tied the highest value achieved previously, in spite of being based upon fewer items.

7 Is the MBE Getting Harder? Easier?

When new MBE items are written and go through the extensive review and revision process required for any MBE item, the judgments of the content experts drafting the items (law professors, practicing lawyers, and judges)—as well as the judgments of outside content experts who conduct an additional review on each item—provide content-related evidence that each item is appropriately targeted to be a prerequisite for the newly licensed lawyer. However, before allowing an item to count toward any examinee's score, NCBE prefers to try out the item by including it on an exam as an unscored pretest item (that is, an item that does not count toward an examinee's score). Pretesting provides verification that each item has acceptable statistical characteristics—for example, that it is not so difficult that only a small number of examinees answer it correctly. While the pretest items do not count toward an examinee's score, gathering statistical information from pretest items helps to build future exams with the best and most stable statistical characteristics.

Each set of items comprising an MBE (i.e., each test form) is built to the same content and statistical specifications to be as consistent as possible over time. However, it is very challenging to build test forms that are of identical difficulty when the collection of questions changes across test forms. Some test forms may be easier and some may be more difficult. This is where the statistical process of equating is used to address any differences in difficulty. Because scaled scores are equated, an examinee receiving an easier test will not be unfairly advantaged and an examinee receiving a more difficult test will not be unfairly disadvantaged in the MBE scaled score each examinee receives. The MBE scaled scores maintain consistent meaning across test forms so that scores reflect consistent levels of performance. The MBE is not getting harder or easier. 

Notes

1. For more detailed descriptions and examples of equating, see Mark A. Albanese, Ph.D., "The Testing Column: Equating the MBE," 84(3) *The Bar Examiner* (September 2015) 29–36; Deborah J. Harris, "Equating the Multistate Bar Examination," 72(3) *The Bar Examiner* (August 2003) 12–18; Michael T. Kane, Ph.D. & Andrew Mroch, "Equating the MBE," 74(3) *The Bar Examiner* (August 2005) 22–27; Michael J. Kolen & Robert L. Brennan, *Test Equating, Scaling, and Linking: Methods and Practices* (Springer 3rd ed. 2014); Lee Schroeder, Ph.D., "Scoring Examinations: Equating and Scaling," 69(1) *The Bar Examiner* (February 2000) 6–9.
2. For the few jurisdictions that do not scale their written components to the MBE, grader leniency would inflate their



We welcome your ideas for future Testing Columns!

Do you have other questions that you'd like answered?

Are there topics on which you'd like a refresher?

Are you curious about how things work behind the scenes?

Send your ideas to contact_editor@ncbex.org.

- pass rates, because a lenient grader (or graders) would lead to higher written scores. Likewise, grader stringency would deflate their pass rates, because a stringent grader (or graders) would lead to lower written scores. Without scaling the written component to the MBE, the degree of stringency or leniency is not accounted for, which will affect the jurisdiction's pass rate.
3. Susan M. Case, Ph.D., "The Testing Column: Procedure for Grading Essays and Performance Tests," 79(4) *The Bar Examiner* (November 2010) 36–38.
4. See Judith A. Gundersen, "It's All Relative—MEE and MPT Grading, That Is," 85(2) *The Bar Examiner* (June 2016) 37–45.
5. Stephen P. Klein, *An Evaluation of the Multistate Bar Examination* (National Conference of Bar Examiners 1982); Stephen P. Klein, *Summary of Research on the Multistate Bar Examination* (National Conference of Bar Examiners 1993).
6. NCBE's Joe E. Covington Award for Research on Testing for Licensure is an annual award intended to provide support for graduate students in any discipline doing research germane to testing and measurement, particularly in a high-stakes licensure setting.
7. See Sarah M. Bonner, Ph.D., "A Think-Aloud Approach to Understanding Performance on the Multistate Bar Examination," 75(1) *The Bar Examiner* (February 2006) 6–15, at 10, category 8, example B).
8. Mark A. Albanese, Ph.D., "The Testing Column: Let the Games Begin: Jurisdiction-Shopping for Shopaholics (Good Luck with That)," 85(3) *The Bar Examiner* (September 2016) 51–56; Susan M. Case, Ph.D., "The Testing Column: Relationships Among Bar Examination Component Scores: Do They Measure Anything Different?," 77(3) *The Bar Examiner* (August 2008) 31–33.
9. Klein 1993, *supra* note 5.
10. Stephen P. Klein, "On Testing: How to Respond to the Critics," 55(1) *The Bar Examiner* (February 1986) 16–24.
11. Erica Moeser, "President's Page," 86(2) *The Bar Examiner* (June 2017) 4–6, at 5.