Originally appearing in THE BAR EXAMINER*, Vol. 72, No. 3, August 2003. Reprinted with permission of the National Conference of Bar Examiners.*

EQUATING THE MULTISTATE BAR EXAMINATION

by Deborah J. Harris

Each administration of the Multistate Bar Examination (MBE) undergoes a process called "equating"; this process is designed to ensure that examinee scores are not unfairly affected by the particular test form's relative difficulty as compared to other forms administered on different test dates. This article will explain the need for equating, the equating process, and how equating affects scores. The article will focus on two common equating methods: linear equating, which is currently being used for equating the MBE, and item response theory (IRT) equating, which is being investigated as a possible alternative for the MBE.

This article will provide an introduction to linear and IRT equating; those readers interested in a more complete discussion, including the assumptions that different equating methods make and the practical issues that need to be considered in operational equatings, should consult the equating literature.[1]

The discussion below will focus on equating for the Multistate Bar Examination. However, to avoid some of the complexities involved in operational equatings, the examples provided here are somewhat simplified, and do not accurately represent the actual MBE equating process; similarly, the results provided do not represent actual operational results.

The MBE is a 200-item multiple-choice test, which contains items from six content areas. An examinee's raw score is found by adding up the number of items the examinee answered correctly; the equating process translates these raw scores into scale scores. These scale scores, which run from 0 to 200, are reported to the jurisdictions for use in the decision-making process.

WHAT IS EQUATING?

Equating is a statistical adjustment used to compensate for any advantage or disadvantage experienced by examinees who might take an easier or harder form of a particular test, such as the MBE. In practice, equating is used when a testing or licensing entity wishes to compare examinee scores over test administrations, yet does not wish to administer the same test questions every time the test is given.

For example, suppose a jurisdiction's board of bar examiners uses the MBE as part of its criteria for admission to the bar. The MBE is given in February and July every year, and the board wants all the scores to be comparable. That is, a score of 138 obtained in February should indicate the same level of achievement on the exam as a score of 138 obtained in July. However, the board members do not want the July examinees to hear about the content of the particular test questions from the February examinees, nor to have examinees who sit for the test in February encounter the same test questions when they retake the test in July. Therefore, they want the July examinees to see a test form that is different than the form seen by the

February examinees, i.e., the July questions should be different from those on the February test. Because the test questions will be different on any two forms, there will virtually always be differences in the difficulty level across the forms. This happens despite the test developers' following careful test development procedures and building the test to conform to rigid test content and statistical specifications; it is virtually impossible to create two forms of a test that are identical in difficulty for all examinees.

Equating is a statistical adjustment that seeks to compensate for that small difference in difficulty; the aim is for no examinee to receive an advantage, or disadvantage, because of the particular test form taken. If the equating process is successful, an examinee should receive the same reported (equated) scale score regardless of which form is administered. Therefore, if the July 2002 MBE examinees on average score higher, after equating, than the February 2002 examinees, it should be because the examinees who test in July are more able or better prepared, and not because the July test form is easier than the February form.

IF TWO FORMS OF THE SAME TEST WERE ADMINISTERED TO A SINGLE GROUP OF EXAMINEES, IT WOULD BE RELATIVELY TRIVIAL TO DETERMINE WHICH FORM WAS EASIER BY SIMPLY COMPARING THE GROUP'S AVERAGE SCORES FOR THE TWO FORMS. . . .

COMMON ITEM EQUATING

In equating, the intent is to make adjustments for small differences in the overall difficulty of two forms of the same test. Both linear and IRT methods of equating may use an embedded common item design for collecting data. Under this design, the two forms of a test administered to two different groups of examinees contain, within the larger sets of items, an identical set of questions, the so-called "common items." The performance of the two groups on these common questions is compared to determine the difference in ability between the two groups, and the performance of each group on the common items and the unique items is used to estimate the difference in difficulty between the two forms.

If two forms of the same test were administered test to a single group of examinees, it would be relatively trivial to determine which form was easier by simply comparing the group's average scores for the two forms, with the form having the higher average score being judged to be the easier form. However, administering two test forms to the same group of examinees is difficult to do in practice because of security, fatigue, motivation, and other practical issues. Therefore, comparisons must generally be made between separate administrations of an examination, and the relative difficulty of the test forms will make a difference in those comparisons.

Let's say for example, that a high school French teacher decided to give two exams to her students. On one exam, the students who took the test answered 80 percent of the questions correctly, on average; on the other exam, the average score was 60 percent. Based on these numbers, one might think that the second exam was a more difficult exam. But if you know that the first exam was given to a third-year French class and the second exam was given to a first-year class, your perception of the relative difficulty of the exams might change.

In common item equating, one group of examinees is administered one form of a test, including a particular subset of items; a second group of examinees takes a second test form that includes the same subset of items. The common items are used to adjust

for differences in ability between the two groups of examinees. The set of common items included in both forms represents a "mini test" in terms of content and statistical properties. The scores of the two groups of examinees on the common items are used to adjust for differences between the two groups of examinees, as explained below.

If the two groups did equally well on the common items (and the common items are indeed a "mini test"), then any differences in the groups' average raw scores on the two forms should be due to differences in the difficulty levels of the two forms. For example, if the two groups did equally well on the common items and if Group A did better on Form A than Group B did on Form B, then Form A would be considered easier overall than Form B. If the group that took Form A did better on the common items than the group that took Form B, and if Group A did proportionally better on the test overall, it suggests that the two forms have the same difficulty. If Group A did better than would be expected on Form A relative to Group B's performance on Form B, then Form A would be determined to be easier overall than Form B and vice versa. The use of common items allows a comparison of the abilities of the two groups on identical items. Once the group factor is accounted for, the form difficulty difference can be determined, and the appropriate statistical adjustment to raw scores can be made to ensure that the reported scale scores for both forms are equivalent. Statistical equating procedures are designed to make these kinds of adjustments, so that the scale scores have the same meaning regardless of which test form an examinee took.

FOR MBE SCORES, A CANDIDATE'S REPORTED SCORE . . . TELLS A BOARD WHAT IT NEEDS TO KNOW ABOUT THE CANDIDATE'S PERFORMANCE; IT IS NOT NECESSARY TO ALSO KNOW THE PARTICULAR TEST FORM THE CANDIDATE TOOK . . . .

For MBE scores, a candidate's reported score of 138 tells a board what it needs to know about the candidate's performance; it is not necessary to also know the particular test form the candidate took, because equating ensures that a score of 138 represents the same level of achievement over time regardless of the particular form taken.

Linear Equating Methods

All equating methods make statistical adjustments to raw scores to compensate for small differences in form difficulty. These adjustments can range from a simple adjustment, where the same number of points is added to or subtracted from each raw score, to more complicated methods, where the amount of the adjustment, and even the direction of the adjustment, can differ for different raw scores.

MBE equating is currently done using common item linear methods. There are several linear equating methods, which differ in the assumptions they make in order to obtain the appropriate adjustments to raw scores. Rather than adjusting all scores on the new form by adding or subtracting the same number of points, linear equating allows for different adjustments to be made throughout the scale, though only in a linear fashion. Assume that a July form of the MBE is to be equated to a previous form of the test. Basically, raw scores on the July form are equated to raw scores on the previous form, typically resulting in non-integer raw scores (e.g., a raw score of 80 on the July form might be equivalent to a raw score of 80.976 on the previous form); these equated raw scores are then converted to reported scale scores, using the raw-to-scale score

conversions for the previous form, and rounding to integer scores from 0 to 200.

The raw scores on the new form are transformed to reported scale scores by a linear equation: Scale score = (slope x raw score) + intercept. For the example described above, the equating might be summarized in the following linear equation:

Reported July score = (.8067 x raw July score) + 38.9978.

Using this equation, a raw score of 110 would be reported as a scale score of 128, while a raw score of 150 would be reported as a scale score of 160.

From these two values, it can be seen that the equating process does not "add the same number of points" to each raw score; a raw score of 110 becomes a scale score of 128 (a difference of +18), while a raw score of 150 becomes a scale score of 160 (a difference of +10).

The current linear equating procedures for the MBE involve two sets of common items, one from a previous February-administered test and one from a previous July-administered test. Each set of common items is chosen to be representative of the overall content and difficulty of the MBE. By using two sets of common items from different MBE forms, rather than only a single set, the accuracy and stability of the equating process is enhanced.

IRT Equating Methods

An alternative to linear equating is Item Response Theory (IRT) equating. As is the case with linear methods, several IRT methods exist, with different models, different scaling methods, and different specific equating procedures, allowing for many combinations to be used in practice. It has been found that when good testing practices are used, the results of the different IRT methods (or, for that matter, IRT and linear methods) do not differ appreciably in many situations.

Each IRT method assumes that examinee performance for each item follows a particular statistical model, which can sometimes be quite complicated. However, the basic premise is the same as with linear equating: Examinee responses on the common items are used to adjust for differences in the two test forms. The IRT methods use the model to estimate characteristics of the items on each test form (such as item difficulty) separately, and then use the common item characteristics to adjust the raw scores.

The IRT adjustments are not constrained to be linear, which makes them more flexible. In the linear equating example provided earlier, every July raw score was converted to a reported score by first multiplying the raw score by .8067 and then adding 38.9978. With IRT equating, the adjustment cannot be summarized by a single equation. If we use the same data from the earlier example and compare hypothetical results using both linear and IRT methods, a raw score of 110 converts to a reported score of 128 using either method; a raw score of 166 converts to a reported score of 173 using the linear equating method but a reported score of 175 using the IRT method, while a raw score of 126 converts to a 141 using linear equating and a 140 using the IRT method.

In addition to providing more flexibility in making adjustments, IRT methods also allow more flexibility in linking designs. Instead of using a common item set to link to a single previously used test form, IRT methods can be used to link to a "pool" of items from multiple forms, which may allow for improved equating in some situations, such as where security or item relevance (e.g., passage of a new law outdates a specific item, making it unusable as a common item) are issues.

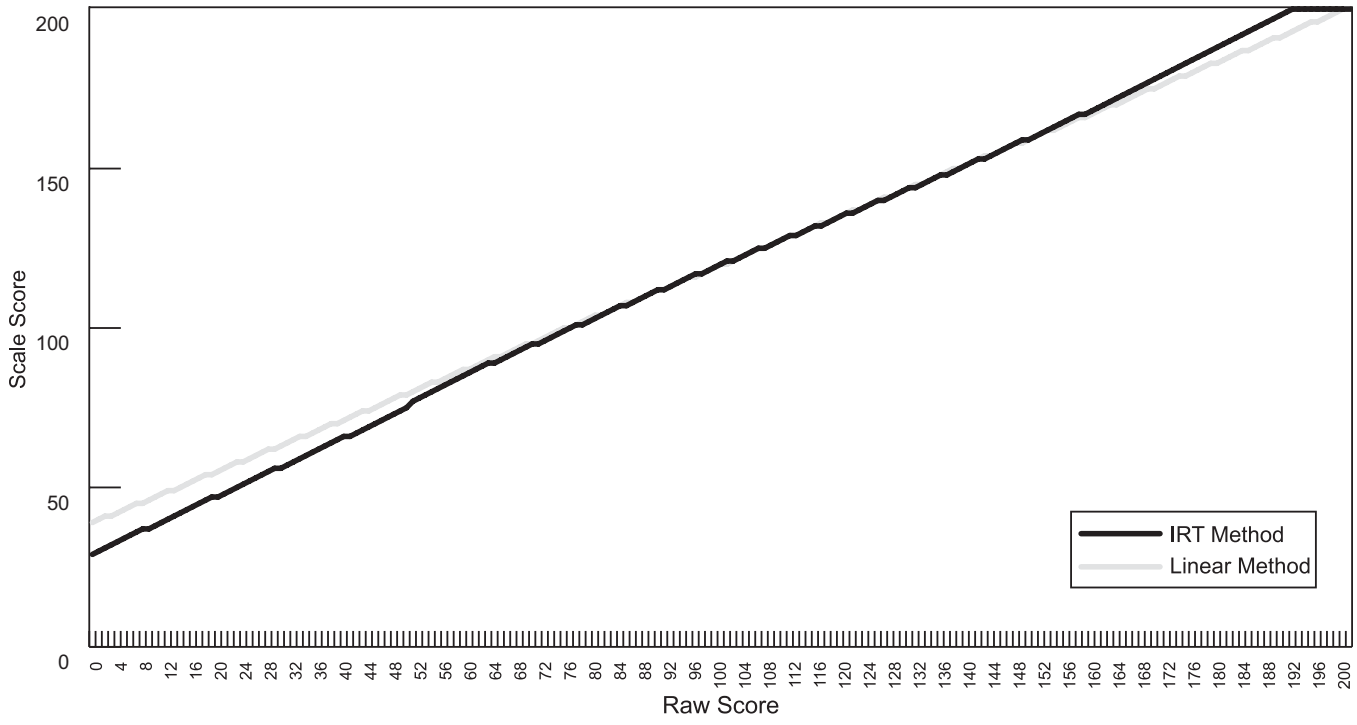
Illustrative Comparison

In order to get a sense of how IRT equating would compare to linear equating on the MBE, and as an experiment for possible future IRT equating, psychometricians at ACT, Inc., equated a July form of the 200-item MBE to the score scale using a single 30-common-item link to one previously administered form, which was already on the 0 to 200 MBE score scale.[2] The number of examinees taking the exam was large enough to support equating, with many thousands of examinees being administered each form. The July form was equated to the previous form using both linear and IRT methods. The linear and IRT equatings both resulted in conversions for transforming the raw scores on the July form to scale scores on the 0 to 200 scale. The conversion can be summarized by a line for linear equating, but requires a full conversion table for the IRT method. The raw-to-scale score conversions are summarized visually in Figure 1. The lighter line in Figure 1 shows the conversions for the raw scores using linear equating; the darker line shows conversions for the raw scores using IRT equating.

As can be seen from the figure, the raw-to-scale score conversions for the middle scores (roughly raw scores of 65 to 162) are quite similar for linear and IRT equating; the scores in the upper and lower ends are more dissimilar, which is usual when linear and curvilinear IRT equating methods are compared, and is generally is not of concern on the MBE because licensing decisions are virtually never made in the tails of the score distribution.[3]

In the range of raw scores from 65 to 162, where over 97 percent of the examinees on the July form scored, the conversions resulting from the two different equating methods differ by one point or less (for 21 of the raw scores, the IRT method resulted in

Figure 1. Comparison of IRT and Linear Raw-to-Scale Conversions



a conversion that was 1 point higher than the linear method conversion, for 13 scores, the IRT method was one point lower, and for 64 scores, the conversions were identical). In the tails, the differences are larger, but from a practical standpoint, few examinees score there (the lowest score in this data set was in the 50s; the highest was in the 180s). Even decisions that combine MBE scores with other measures in a compensatory model would be likely to make the same ultimate decision (e.g., examinees scoring in the extreme ends of the raw score scale are likely to be so far above or below the passing score that the conversion used is unlikely to matter, even when the MBE score is combined with other measures).

THE IRT METHODS ARE MORE COMPLICATED TO IMPLEMENT AND TO UNDERSTAND OR EXPLAIN THAN LINEAR METHODS, BUT THEY ARE ALSO MORE FLEXIBLE. . . . THIS FLEXIBILITY HAS PRACTICAL ADVANTAGES IN TERMS OF BOTH THE QUALITY OF ITEMS USED (ONE CAN SELECT THE BEST ITEMS FROM ACROSS SEVERAL FORMS) AND SECURITY.

WHICH EQUATING METHOD IS "RIGHT"?

Is one type of equating "better" than the other? Not consistently. In choosing an equating method, several issues need to be considered, such as the number of examinees available, the likely distributions of scores, the assumptions different methods make, the turnaround time required, and so on. There is no easy rule to apply to determine which equating method is preferable in any given situation. The IRT methods are more complicated to implement and to understand or explain than linear methods, but they are also more flexible. In particular, IRT methods facilitate equating to a pool of items, rather than to a particular form, meaning all the common items in a set need not be drawn from a single previous form of a test. For example, if one wished to draw some of the Torts common items from previous forms A and B, Contracts common items from previous forms C, D, and E, and so on, the IRT equating method would allow this selectivity while a linear equating method would not. This flexibility has practical advantages in terms of both the quality of items used (one can select the best items from across several forms) and security. Item usage is improved because previous forms that cannot provide enough "good" items in all six content areas for a set of common items can still contribute the "good" items they have to a pool from which the set of common items is selected. Security is improved because the common items do not all come from a single previous form, thus lessening the risk that an examinee who retakes the test will see a number of the same items again. If the common items are drawn from a pool of eight previous forms, the examinee would need to have taken the test on all eight previous occasions to have already seen all the common items; when the common items all come from a single previously administered form, the examinee would need to have taken the test on only the one occasion where the previous form was administered to have seen all the common items.

Both linear equating and IRT equating are effective in eliminating differences in test difficulty among test forms. So which equating method is best for the MBE? Each method has its practical advantages and disadvantages. Because there were some differences in the conversions for each method, some examinees will score higher if linear equating is chosen, and some examinees will score lower. However, where most of the examinees (97 percent) generally score, the differences between the two conversions

were one point or less. For this example then, and likely for every administration of the MBE, both linear and IRT equating give similar results, and either approach would be defensible. IRT equating, however, has the advantage of superior flexibility and is a popular equating method for many testing and licensing agencies.

ENDNOTES

1. Good resources are those such as Brennan and Kolen (1995), Harris (1993), and Angoff (1971). Two Instructional Topics in Educational Measurement modules provide a nice introduction to linear equating (Kolen 1988) and IRT equating (Eignor and Cook 1991). An additional module (Harris 1989) provides an introduction to IRT.
2. The example presented here is a simplified version of the research conducted, and involves only one set of common items. The research was completely separate from the operational equating, where the standard linear methods and two sets of common items were used for actual score reporting for this administration of the MBE.
3. For most jurisdictions, a "passing" scale score on the MBE will fall in the 130 to 145 range; therefore an examinee with a raw score of 65 or 165 is already a clear fail or a clear pass, and variations in the scale score results for those raw scores will not make any difference in the licensing decisions.

REFERENCES

Angoff, W. 1971. Scales, norms, and equivalent scores. In EDUCATIONAL MEASUREMENTS, 2nd ed. Edited by R. Thorndike, 508-600. Washington, DC: American Council on Education.

Cook, L., and Eignor, D. 1991. An NCME instructional module on IRT equating methods. ***Educational Measurement: Issues and Practice*** 10:37-45.

Harris, D. 1988. An NCME instructional module on comparison of 1-, 2-, and 3-parameter IRT models. ***Educational Measurement: Issues and Practice*** 8:35-41.

Harris, D. 1993. Practical issues in equating. Paper presented at the Annual Meeting of the American Educational Research Association, Atlanta.

Kolen, M. 1988. An NCME instructional module on traditional equating methodology. ***Educational Measurement: Issues and Practice*** 8:29-36.

Kolen, M., and Brennan, R. 1995. TEST EQUATING METHODS AND PRACTICES. New York: Springer-Verlag.

DEBORAH J. HARRIS is Director of the Measurement Research Department at ACT. She has been working in the area of equating for more than 18 years in both research and operational settings, and has authored or co-authored numerous presentations, papers, and articles on equating and other measurement topics. She is also an adjunct at the University of Iowa, where she has taught measurement and statistics. She received her doctorate from the University of Wisconsin-Madison (Educational Psychology), and her M.A. and B.S. degrees (Elementary Education) from Central Michigan University.