

# EQUATING THE MBE

*by Michael T. Kane, Ph.D., and Andrew Mroch*

**T**he different forms of the Multistate Bar Examination (MBE) administered on different test dates are designed to be very similar to one another in content and difficulty. However, because most of the questions on each new form of the MBE are developed for that particular form, the different forms can vary somewhat in difficulty. In order to ensure consistency and fairness across the different MBE forms given on different test dates, equating procedures are used to adjust for any variability in difficulty across the different forms.

Two commonly used methods of equating are linear equating and item response theory (IRT) equating. For both of these methods, some of the items on each new form of the test are drawn from prior forms of the test, and these common items are used to equate (or link) the new form of the MBE to prior forms of the MBE. Candidate performance on these common items provides a basis for comparing the performance of the candidates taking different test forms on different test dates. Once the competence of the group of candidates taking a particular test form has been estimated from their performance on the embedded set of common items, the statistical characteristics of the test form can be determined, and adjustments can be made for differences in difficulty across test forms (based on differences in the levels of performance on the common items of the groups taking the tests).

In the past, the MBE has been equated using a procedure referred to as “common-item linear equating.” As of February 2005, the MBE is being equated using IRT methods. While IRT equating and linear equating are both designed to ensure that different forms of the MBE administered on different test dates yield scores that are comparable, and while both approaches employ embedded common items, they use different mathematical models and different rules for selecting the common items.

In discussing equating, it is necessary to draw a distinction between two kinds of scores, raw scores and scaled scores. A candidate’s raw score on the MBE is simply the number of questions that the candidate answered correctly on the test form that he or she took, and can have values between 0 and 200. The equating procedure transforms each raw score into a corresponding scaled score that is designed to have the same meaning regardless of when the candidate took the test.

## COMMON-ITEM LINEAR EQUATING

One way to link two forms of a test is to include a set of identical common items in the two forms. Suppose, for example, that we want to equate a new test form to a prior test form. Although most of the items on the two tests will be different, we can include a set of items from the prior test in the new test. After the new test is given, we can determine the relationship between the raw scores on the new

test (total number correct) and the raw scores on the set of common items (number correct on the common set); in particular, we can develop a linear relationship between the raw scores on the new test and the raw scores on the common items.

Similarly, using the data from an administration of the prior test, we can establish a linear relationship between raw scores on the prior test and the scores on the same set of common items. Once we know the relationship between raw scores on the new test and the common-item scores and that between raw scores on the prior test and the common-item scores, we can determine the raw score on the prior test that would be expected of a candidate with any particular raw score on the new test. Using this linear relationship, scores on the new test can be transformed to the corresponding expected score on the prior test.

If the prior test had been equated to a still-earlier test (generally using a different set of common items), the expected score on the prior test form could then be transformed to an expected score on this still-earlier test form. In this way, a long string of

tests can all be equated to one another. In practice, all of the tests are equated back to the first test in the sequence, which is called the anchor test, and all scores on all forms of the test are transformed to the corresponding expected score on the anchor test; these transformed scores are called scaled scores. For the anchor test, the scaled scores are the same as the raw scores.

Figure 1, below, illustrates the process of common-item linear equating. The raw scores on the new test form are linked to the scores on the common items, which are, in turn, linked to the raw scores on a prior test form. Because the prior form is linked to the anchor form (perhaps through some intermediate steps), the new form can be linked to the anchor form, and the scores on the new form can be transformed to the expected scores on the anchor form, or the scaled scores.

Because they do not depend on the particular form that a candidate took, the scaled scores are more easily interpretable than the raw scores, and they provide a fairer basis for making licensure

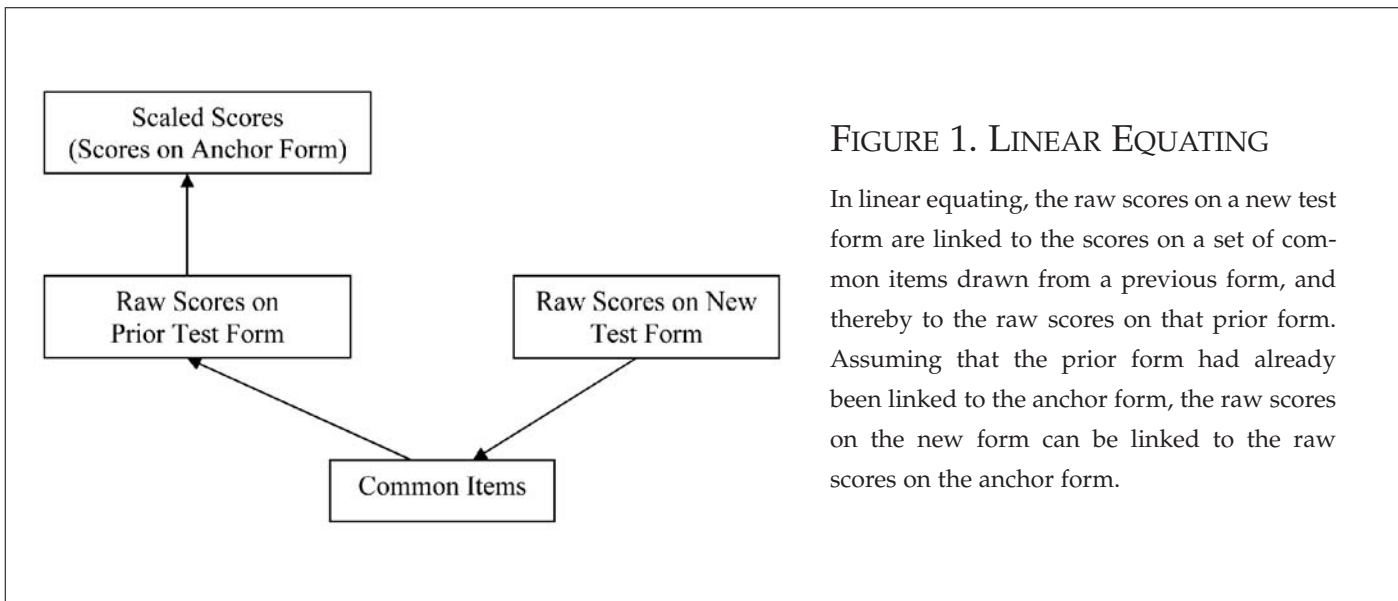


FIGURE 1. LINEAR EQUATING

In linear equating, the raw scores on a new test form are linked to the scores on a set of common items drawn from a previous form, and thereby to the raw scores on that prior form. Assuming that the prior form had already been linked to the anchor form, the raw scores on the new form can be linked to the raw scores on the anchor form.

decisions. Before 2005, each new form of the MBE was equated directly to two prior forms, each of which had been equated back to the anchor form using common-item linear equating.

Linear equating is relatively simple to implement and describe, and it yields a linear relationship between scaled scores and raw scores that can be represented by a straight line or a linear equation that is easy to use. However, linear equating has the disadvantage of relying on links to only one or two earlier forms of the test. A somewhat newer set of models based on item response theory uses multiple links for each new form and therefore tends to be more reliable and more flexible.

### IRT EQUATING

Item response theory (IRT) equating employs more complex statistical models than linear equating, but it has the same basic goal as linear equating, that is, to generate scaled scores for each form that are comparable across test forms. IRT equating employs the same basic logic as linear equating (i.e., the use of common items to link different forms), but it relies

on more sophisticated mathematical models and has some technical advantages over linear equating.

Item response models represent a candidate's performance on each item in terms of the candidate's ability on an IRT ability scale and in terms of certain item parameters that describe the statistical properties of the items. For example, one parameter is used to describe how difficult the item is, and another parameter describes how well the item differentiates different levels of ability. A third parameter, a guessing parameter, represents the chances that a low-scoring candidate will answer the item correctly by guessing. Once the item parameters have been estimated, the theory can predict how well candidates with different levels of ability will do on the item.

After a test form has been administered, the model can be used to estimate both each candidate's level of ability on the IRT ability scale and the values of the item parameters for each item on the new test form. As part of this estimation process, the IRT procedure defines a particular IRT ability scale on which the ability level of each candidate and the parameters for each item are specified.

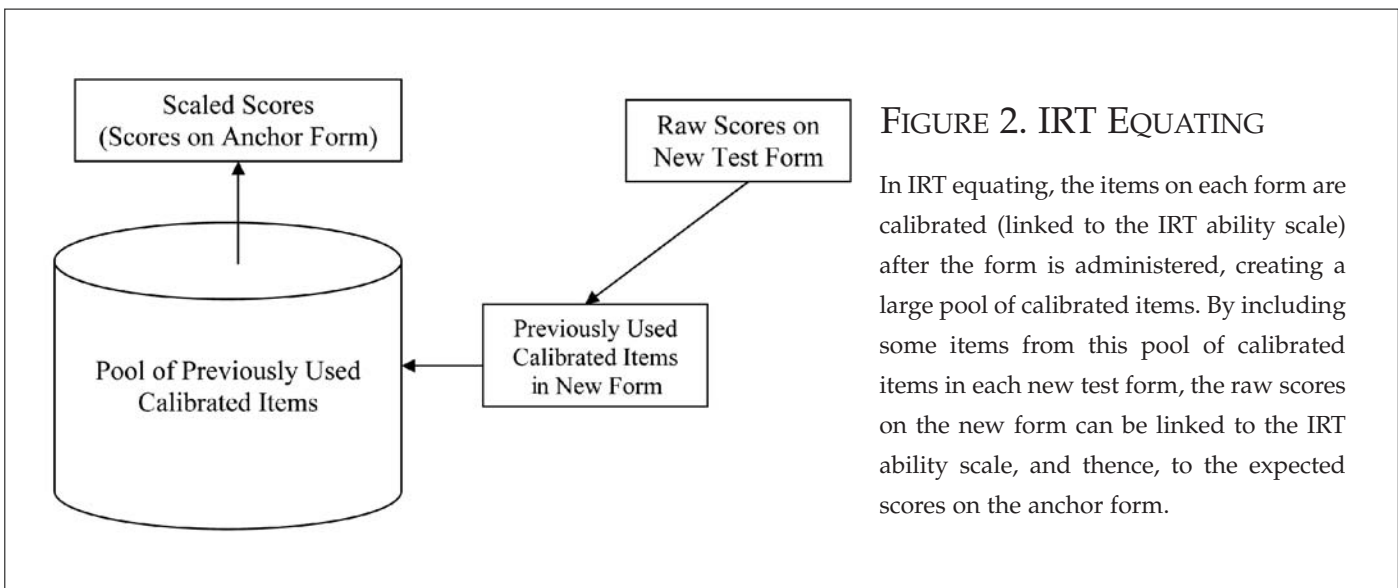


FIGURE 2. IRT EQUATING

In IRT equating, the items on each form are calibrated (linked to the IRT ability scale) after the form is administered, creating a large pool of calibrated items. By including some items from this pool of calibrated items in each new test form, the raw scores on the new form can be linked to the IRT ability scale, and thence, to the expected scores on the anchor form.

## A QUICK GUIDE TO IRT EQUATING FOR THE MBE

Although different forms of the Multistate Bar Examination (MBE) are designed to be very similar to one another in content and difficulty, they do vary somewhat, because most of the questions on each new form are new. This variability means that some test forms are necessarily more difficult than others, even if only slightly. In order to ensure consistency and fairness across the different MBE forms, statistical equating procedures are used to adjust for any variability in difficulty across the different forms.

In discussing *equating*, it is useful to draw a distinction between two kinds of scores, *raw scores* and *scaled scores*. A candidate's raw score on the MBE is simply the number of questions (0 to 200) that the candidate answered correctly on the test form that he or she took. The equating procedure transforms each raw score into a corresponding scaled score that is designed to indicate the same level of performance regardless of which form of the test the candidate took. The scaled scores also range from 0 to 200, but can have decimal values.

The equating model used for the MBE (known as *item response theory equating* or *IRT equating*) describes items (i.e., questions) in terms of certain item parameters. One item parameter is used to describe how difficult the item is, and another parameter describes how well the item differentiates different levels of ability. A third parameter, a guessing parameter, represents the chances that a low-scoring candidate will answer the item correctly by guessing.

Using the item parameters, it is possible to anticipate how candidates with various levels of ability will do on particular items. For example, if a candidate has a high ability level and the item has a low difficulty level (i.e., is an easy item), the candidate would be expected to have a good chance of getting the item correct. The same candidate would have a lower probability of getting a more difficult item correct.

After a test form has been administered, the results can be analyzed to determine parameter values for each of its items. As noted above, using the item parameters, it is possible to project how well candidates at different levels of ability would do on any item, and therefore, it is possible to project how well candidates at different ability levels would do on any test form for which the item parameters are known. The raw scores on different test forms that are expected for candidates at a particular ability level are considered equivalent.

A particular test form (the *anchor form*) is chosen to define a unique scale for reporting scores, and the raw scores on all test forms are converted to their equivalent score on the anchor form. These transformed scores are the scaled scores. Because the scaled scores have a common interpretation, they do not depend on the particular form that a candidate took. The scaled scores provide a fairer basis for making licensure decisions.


The IRT ability scale is not easy to interpret in itself. However it is possible, using the IRT model and the item parameters, to predict the raw score that candidates at different levels on the IRT ability scale would get on any test form for which the item parameters are known. For example, if a candidate has a high ability level and the item has a low difficulty level, the candidate would be expected to have a high probability of getting the item correct. The same candidate would have a lower probability of getting a more difficult item (one with a higher difficulty parameter) correct. A candidate with a low ability level would have a very low probability of getting a difficult item correct. By summarizing these estimates over all items in a test, we can predict raw scores for candidates at different levels of ability. It is therefore possible to pick a form to be used as an anchor form, and to transform the ability estimates for all subsequent test forms to expected raw scores on that anchor form. Once an item's parameters have been estimated, the item is said to be calibrated. When a new form of the test is developed, some calibrated items are drawn from previously administered test forms and included in the new form of the test. These "common items" are used to ensure that the item parameters and the candidate ability estimates derived from the new test form are comparable to those derived from the old form.

[I]T IS POSSIBLE, USING THE IRT MODEL AND THE ITEM PARAMETERS, TO PREDICT THE RAW SCORE THAT CANDIDATES AT DIFFERENT LEVELS ON THE IRT ABILITY SCALE WOULD GET ON ANY TEST FORM FOR WHICH THE ITEM PARAMETERS ARE KNOWN. FOR EXAMPLE, IF A CANDIDATE HAS A HIGH ABILITY LEVEL AND THE ITEM HAS A LOW DIFFICULTY LEVEL, THE CANDIDATE WOULD BE EXPECTED TO HAVE A HIGH PROBABILITY OF GETTING THE ITEM CORRECT. THE SAME CANDIDATE WOULD HAVE A LOWER PROBABILITY OF GETTING A MORE DIFFICULT ITEM CORRECT.

As new test forms are administered, it is possible to create an item pool of all calibrated test items from previously administered forms of the test. After each new form is administered, the new items can be calibrated by estimating their item parameters, and these calibrated items can then be added to the pool. In IRT equating, items from many previously administered forms of the test are embedded in the new form of the test. Using candidate performance on the previously calibrated items drawn from the pool, the raw scores on the new form can be linked to the IRT ability levels, and then to the scaled scores (see Figure 2). While the linear approach requires a large number of common items from one or two prior test forms, the IRT approach can draw common items from a large number of different prior test forms.

In the transition from common-item linear equating to IRT equating for the MBE in 2005, the anchor form was not changed. The IRT ability scale was linked to the scaled scores on prior MBE forms, and thereby, to the same anchor form that has been used for linear equating. Therefore, the scaled scores on new forms of the MBE continue to have the same meaning as they did before the introduction of IRT equating.

IRT equating tends to be more flexible than linear equating. Instead of consistently drawing a large

number of items from one or two previous forms (which is required for linear equating methods), IRT allows test developers to draw a few items from a large number of forms or many items from a few forms, or any combination of these approaches. The added flexibility of IRT is especially useful if the security of some part of a test form is compromised (e.g., the morning half of an MBE form is lost). In such cases, the form cannot be used for linear equating (because for statistical purposes the whole form must be intact); the IRT models, however, would allow the use of items from that part of the test that was not compromised. Common-item linear equating uses a subset of items from one or two prior forms of the test, whereas IRT equating uses items from multiple prior forms of the test. Common-item linear equating has one or two big links (each involving many items) to one or two prior forms of the test, and IRT equating relies on a large number of links (each involving just a few items) from many prior forms of the test. As a result, IRT equating is generally more resilient to any special circumstances that might interfere with a particular linking of one form to another. 



MICHAEL T. KANE, Ph.D., is the Director of Research for the National Conference of Bar Examiners.



ANDREW MROCH is an intern at the National Conference of Bar Examiners and is a doctoral student at the University of Wisconsin-Madison (Educational Psychology). He holds a B.S. from Iowa State University (Psychology) and an M.S. from the University of Wisconsin-Madison (Educational Psychology).