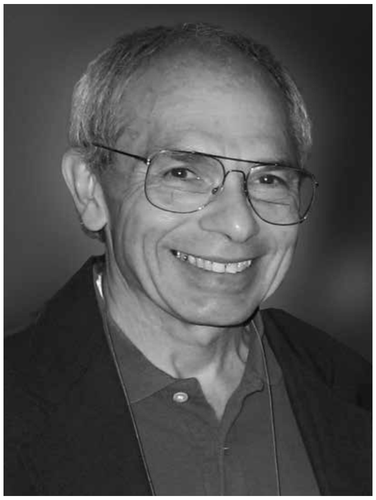
The Testing Column

Raw Scores on the MBE Tell You Little—and Probably Less Than You Think

by Mark A. Albanese, Ph.D.



Effective with the February 2014 administration of the MBE, MBE score reports sent to jurisdictions will no longer include raw total scores (the simple sum of items answered correctly) and raw subject area subscores (the sum of items answered correctly within each separate subject area) and their associated percentiles. We will continue to provide the scaled total score and the jurisdiction-specific percentile associated with that scaled score. This change is intended to conform to standard testing practice as well as to reduce confusion among stakeholders. If you are despairing because the raw scores were the only thing on the score report that you thought you understood, I hope this article will make you feel better.

The scaled total score (the score generated from the statistical process of equating, which ensures that the scores have consistent meaning across different administrations) is really all you need, and what can be interpreted from the scaled score is not that different from what you think the raw score gives you. Further, although the raw score would seem to have a straightforward interpretation, it really gives you far less information than you probably think it does, because interpreting raw scores is complicated by the intrinsic difficulty of the items (which can vary quite markedly from one administration to the next) and the proficiency of the examinees (which can also vary quite markedly; witness differences in the performance of examinees in February, who have been, on average, less proficient than those testing in July).

Clearing Misconceptions about Interpreting Raw Scores

Why Raw MBE Scores Do Not Range from 0 to 200

Because there are 200 items on the MBE, many people expect raw scores to range from 0 to 200. In practice, this is never the case. The upper bound is actually 190, since 10 items are being tried out for future use and are not used to compute the examinee's score. Theoretically, raw scores can go as low as 0 but in practice rarely go below 40–50. Someone who makes a good-faith effort to do well on the MBE will, for all practical purposes, never receive a score of 0. Even random guessing on questions will not likely result in a score of 0. (In fact, random guessing would result in a 0 score 1 time out of 1.8 x 10^{24} times. Since a trillion is 1.0 x 10^{12}, a 0 score will not happen unless the examinee leaves his or her answer sheet blank.)

Why a Higher Raw Score Does Not Necessarily Mean Greater Proficiency

We generally try to keep the average difficulty of the examination constant across time, but there are inevitable differences in the mean (the average value of the scores) and standard deviation (a measure of the extent to which scores typically deviate from the mean) of the raw scores at different administrations, as well as what raw score corresponds to a given scaled score.

For example, for the four examinations administered in 2012 and 2013, the raw score mean ranged from about 121 to 129, and the raw score that corresponded to a scaled score of 135, a commonly used passing score, ranged from 116 to 122. Suppose an examinee received a raw score of 120 in July and a raw score of 115 the following February. One might interpret the 120 score as representing greater proficiency. However, it could be that the 115 score represented greater proficiency than the 120 score because the February examination contained items that were intrinsically more difficult. Without adjusting for such differences in intrinsic difficulty of the items in any given test administration, the raw scores can give a fairly confusing, if not distorted, perception of examinee performance.

To assist in interpreting the raw and scaled scores, in the past we gave each jurisdiction the percentile rank that a given score represented based upon the examinees in that jurisdiction. The percentile rank is the percentage of examinees in that jurisdiction whose scores are below the given raw score and its corresponding scaled score for that particular test administration. Percentile rank will now be reported only for the scaled scores. The percentile rank for the raw score is the same as that for the scaled score, but as I showed earlier, the raw score can map to a scaled score that can be quite different depending upon the test administration. All in all, the raw score presents a very imprecise and restrictive picture of an examinee's performance.

The Problem with Using Raw Scores to Meaningfully Interpret Items Answered Correctly

So, what *can* be done with the raw score? The only situation where raw scores can be compared without ambiguity or without being misleading is when the raw scores are obtained from exactly the same set of items. In that situation, if two examinees have raw scores on the MBE that differ by two points, the one with the higher score answered two more items correctly than did the one with the lower score. The examinee receiving the higher score is generally going to be more proficient. If a score of interest is above the mean, it means that the score is above average, and conversely below average if it is below the mean. The percentile rank provides additional information about the score in terms of how other examinees performed. Any use of raw scores beyond comparisons involving scores obtained on the same set of items is ill-advised.

Now, you may be saying to yourself, "That's exactly what I love about raw scores: a difference in points means a difference in test items answered correctly. I know what raw scores mean." Well, maybe, but maybe not. Going back to the example, if the two items answered correctly were answered correctly by 98% of the examinees, would you consider that difference still to be meaningful? What if the two items were answered correctly by 40% of the examinees: would you now consider that difference to be meaningful?

In the former situation, the two items were extremely easy and were answered correctly by almost all examinees. Although there might be meaning derived from the lower-scoring examinee answering such easy questions incorrectly, it will be different from the latter situation, where the two items were very difficult and were answered correctly by less than half the examinees.

In terms of the examinees' proficiency differences, if the two items were very easy and almost all examinees answered them correctly, one would not have great confidence that the two-point difference made a difference. It could be that the lower-scoring examinee was less attentive or distracted when answering the questions or that the examinee was even copying from another examinee; or it could be that it was a meaningful proficiency difference and that the lower score represented deficiencies in core concepts. Thus, the interpretation of a two-point difference on easy items is not clear. If the two items were very difficult, then one might have more confidence that the two points showed a meaningful difference in proficiency. So, the interpretation of a difference in raw total scores is not quite as straightforward as one might think.

Working with Preferred Scaled Scores

Scaled Scores Offer Consistent Meaning over Time

Scaled total scores, however, are another story. Unlike raw total scores, which will change with the intrinsic difficulty of the test items from which they are computed, scaled scores will have the same meaning no matter what the intrinsic item difficulty may be for any given test administration. This means that scaled scores will have the same meaning whether the examination was taken in February 2014 or July 2008 or February 2003 or . . . . (You get my drift.)

Scaled scores will have a mean of approximately 143 in July and 137 in February, although the means may shift if the proficiency of the examinees changes en masse. For example, were the ABA to adopt a standard mandating that a higher percentage of a law school's graduates pass the bar, this might lead to better preparation of students. This in turn might lead to examinees performing better on the test, which could cause mean MBE scores to rise. (Similarly, drastic shifts in law school enrollment that lower enrollment criteria could take the mean in a different direction.) The standard deviation could also be affected if the result is also to increase (or decrease) the range of scores.

However, barring such a change, the standard deviation has been fairly stable at approximately 15. For a bell-shaped score distribution like that produced by MBE scaled scores, scores from one standard deviation below the mean to one standard deviation above the mean (128–158 in July, 122–152 in February) will include approximately 68% of the examinees. If it is extended to two standard deviations above and below the mean, approximately 95% of the examinees will be included. The total range of scaled scores will be from approximately 42 to 197.

Scaled Scores and Percentiles

The percentile associated with a given scaled score, however, will still vary with the particular test administration. As mentioned earlier, because February examinees have been, on average, less proficient than those testing in July, the percentiles for a given scaled score will be higher in February than they will be in July. Thus, care must still be exercised in interpreting the percentiles. However, a two-point difference in scaled scores will be a two-point difference whether comparing a score from February to one from July or comparing a score from 2014 to one from 2008 or even 2003. The scaling process makes a higher score higher for all time (or at least until the law profession changes to the point that we may be compelled to redefine what constitutes proficiency in law and redesign and rescale the MBE).

Scaled Scores and Items Answered Correctly

The one thing that scaled scores lose is the one-to-one link to the number of items answered correctly. An increase of one item in the number answered correctly will result in an increase in the scaled score, but it may not be enough that the scaled score rounded to an integer value will increase by a point. If this keeps you awake at night, we can give you scaled scores with one decimal point. Scaled scores with decimal values will always rise with an increase in raw score, just not always on a one-to-one basis.

Conclusion

So, if you are still pining away for raw total scores, you might try thinking of the scaled total score as being like a clone of the raw total score, only with all the warts removed. The score range for the scaled score is not that different from that of the raw score, and the interpretations many people generally try to make with raw scores are actually better suited to scaled scores.

In the next Testing Column, I will explain our decision to eliminate the subject area subscores. They are not quite an exercise in futility, but in the psychometric world they are close. Till next time….

Mark A. Albanese, Ph.D., is the Director of Testing and Research for the National Conference of Bar Examiners.