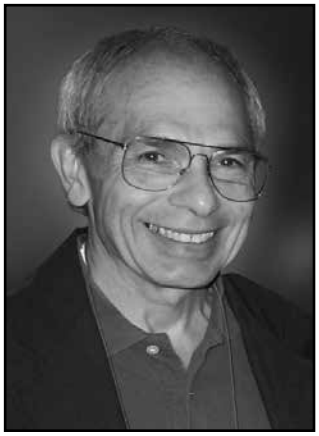
The Testing Column

Differences in Subject Area Subscores on the MBE and Other Illusions

by Mark A. Albanese, Ph.D.



In my last column, I addressed one part of a change to MBE score reports sent to jurisdictions effective with the February 2014 administration of the MBE: the elimination of raw total scores. In this column, I address the second part of the change: the elimination of raw subject area subscores and their associated percentiles. NCBE will continue to provide to jurisdictions the scaled total score and its associated percentile based upon examinees in the jurisdiction. (As a refresher, the raw total score is the simple sum of items answered correctly, and the raw subject area subscores are the sum of items answered correctly within each separate subject area. The scaled total score is the score generated from the statistical process of equating, which ensures that the scores have consistent meaning across different administrations.) This article is intended to explain more fully why NCBE decided to drop subscores from MBE score reports.

Why Subscores Are Unnecessary—and Unreliable

The scaled total score is really all that is needed. Subscores, which by necessity are raw because there are too few items to equate scores, are largely redundant with raw total scores. This redundancy has been confirmed by several independent analyses of examinee response data from past MBE administrations that consistently found subscores to be essentially less reliable estimates of the total score.[1] (They are less reliable because they are based upon only 31–33 items in each MBE subject area instead of 190 total items.) Differences in subscores are often an illusion and can even be deceptive, appearing to show lower proficiency in one subject area when in fact the true weakness may be in a different subject area. (I explain the reasons for this in more detail later in this column.) Thus, the use of subscores in helping students direct their study can result in misinformed decisions. Law schools will encounter problems as well if they attempt to use the subscores to improve their curricula. Higher scores in one subject area than in another do not necessarily mean that students were more proficient in that area, so making curriculum decisions based upon differences in subscores can be misguided. Further, three studies over the last decade that attempted to validate the subscores from analyses of MBE examinee responses have concluded that the subscores cannot be shown to be meaningfully different from the total score.[2]

Subscores and the *Standards for Educational and Psychological Testing*

A major challenge that any testing organization faces is meeting the standards of the testing profession. These standards were last published in 1999 in a document titled *Standards for Educational and Psychological*

Testing.[3] This is a joint publication of the major professional associations involved in standardized testing: the American Educational Research Association (AERA), the American Psychological Association (APA), and the National Council on Measurement in Education (NCME). It is not a document to be trifled with.

The most relevant statement pertaining to the subscore issue is Standard 1.12, which dictates that "the distinctiveness of the separate scores should be demonstrated."[4] When psychometricians say that something needs to be demonstrated, they mean, "Show me the data!" It is not enough to just *say* that the subscores represent different concepts; one must demonstrate that the different subscores are meaningfully different in terms of examinee responses. Therein lies the rub.

In the remainder of this article, I will explain why the MBE subscores do not meet the standards of the profession (and why they are not designed to do so). The continued use of subscores is not defensible in terms of concept, reliability, validity, or utility to examinees or to law schools. I will also describe how eliminating subscores will facilitate moving the MBE forward in evolving to meet the challenges of the changing world of the practice of law.

By Design, the MBE Will Not Produce Usable Subscores

The purpose of the MBE is to provide jurisdictions with evidence of candidate proficiency for granting a single unrestricted license. The subject areas are meant to guide the development of the test so that important content is represented in proportion to its importance to the newly licensed practicing lawyer. The subject areas are not a set of separate hurdles that must be leaped. Therefore, we do not build the MBE to provide data that would enable different decisions to be made for each of the six different subscores. To do so would require a much different approach to selecting items. We select items for inclusion on the MBE that have high correlations with the total score. If they have a higher correlation with a subscore than they do with the total score, more than likely the correlation with the total score will not be high enough for that item to be included on future examinations.[5] So, items that would contribute well to providing a particular subscore are unlikely to function well enough for the total score. Thus, by both concept and design, the MBE does not support the reporting of subscores.

Subscores Do Not Meet Reliability Standards for High-Stakes Examinations

> A man with one watch knows what time it is.
> A man with two watches is never quite sure.
>
> —*Segal's Law*

The MBE presents us with the equivalent of 190 watches (items) to determine whether an examinee is minimally competent to practice law. With all examinees having success on some items but not on others, uncertainty abounds.

When psychometricians talk about uncertainty, it is in the context of test *reliability*. Reliability reflects the degree to which an examinee's score would be likely to change if the examinee were to be tested again with a comparable but different set of items. Reliability values range from 0.0 to 1.0, with 1.0 being a perfect (but, practically, unattainable) level of certainty. A score with a low level of reliability (e.g., less than 0.70) will have a high level of uncertainty, and the same examinee's score could be quite different if a different set of comparable items comprised the test.

The minimum standard for reliability that has generally been accepted for high-stakes examinations

is 0.90.[6] The MBE total scaled score meets this minimum required threshold value for reliability, whereas the subscores do not come close. Because items are not selected for their subscore contributions, but rather for their contribution to the total score, the subscores have relatively low reliabilities. It is not uncommon for one or more of the subscore reliabilities to be less than 0.50, far below the 0.90 value generally accepted as the standard for making high-stakes decisions. Since 2007, the subscore reliabilities (adjusted for a uniform number of items) have ranged from 0.49 to 0.70 and have averaged 0.60. In total, out of 84 subscore reliability estimates obtained for the MBE from February 2007 through July 2013, 52% have been below 0.60 with 98.8% below 0.70.

It could be argued that a lower reliability threshold value could be used for the subscores if their primary use was to guide a failing candidate in preparing to retake the bar exam. But even if the reliability threshold value for subscores were relaxed to 0.75, to increase the reliability of a subscore from 0.60 to 0.75 the number of items would need to be doubled. Thus, in order to obtain sufficient reliability on subscores and to produce scaled versions of each, the number of items in each subject area would need to be increased, on average, to 62–66 items (as opposed to the current 31–33 items—which, beginning in 2015, will decrease to 27–28 with the addition of Civil Procedure as the seventh MBE subject area). I might point out that doubling the items in each subject area is an underestimate for almost half of the subscores that currently have reliabilities below 0.60. Doubling the overall number of scored items on the MBE would therefore increase them from 190 to 380, doubling the testing time from 6 to 12 hours, an impractical prospect.

Validity Cannot Be Established for Subscores

The *validity* of a test score refers to whether a test is measuring what it is intended to measure, and whether a score reflects what it is intended to reflect. Neither tests nor scores are validated in and of themselves, but the wealth of data supporting their use creates what has been termed the "validity argument" in support of their use for various purposes.

Over the past 10 years, numerous studies have attempted to establish the validity argument for whether the subscores on the MBE measure something distinct from the total scaled score. Studies of differences between subject area subscores for each examinee have found that fewer than 1% of the examinees had differences between their subscores that were large enough that they exceeded what would be expected to occur by chance alone.[7] This means that for the vast majority of examinees, what may appear to be meaningful differences between their subscores is an illusion. If they were to act on the differences by studying their lower subscore topics to a greater degree than their higher subscore topics, they would more than likely be basing their decisions on faulty information.

Other studies have found that correlations between subscores are so high as to be virtually identical (greater than 0.90 when adjusted for measurement imprecision).[8] Correlations this high fail criteria for subscore distinctiveness.[9] High correlations also mean that if we were to use a certain subscore to rank-order examinees from the lowest score to the highest score, we would get essentially the same ordering as that obtained from rank-ordering examinees using any other subscore or the total score. However, the rank-ordering obtained from any of the subscores would be much less stable than that of the total score.

Furthermore, separate studies undertaken by NCBE staff and other researchers have attempted to demonstrate that the subscores are statistically distinct from the total scaled score.[10] In total, these studies have used almost the entire universe of statistical approaches designed to detect the existence of subscores as statistically distinct scores from the total scaled score. None of these studies has found that the subject area subscores are statistically distinct from the total scaled score. This conclusion would still apply even if we were to scale each of the subscores, as we do for each examinee's total score.

Subscores Are Not Useful to Examinees and Law Schools

As I touched upon earlier, the subject area subscores can be confusing and misleading. Until we eliminated them from MBE score reports, we reported subject area subscores in their raw form, meaning the simple sum of the correct answers to items in that subject area. Because raw scores vary based upon the intrinsic difficulty of the items from which they are computed, because the MBE contains a relatively small number of items per subject area, and because some subject area subscores are based upon more items than others, a higher subscore in one subject area does not necessarily indicate a higher level of proficiency in that subject area.

Thus, an examinee who received a score report showing a higher subscore in Torts than in Real Property may think that he or she was more proficient in Torts; however, because there are 33 Torts items and only 31 Real Property items, subscores in Torts are likely to be higher than those in Real Property by force of numbers. Further, even for subscores with the same numbers of items, depending on the relative difficulty of the items in those subject areas, a higher score may possibly represent lower proficiency.

Similarly, if an examinee retakes the MBE, a higher subscore in a particular subject area on the retake does not necessarily demonstrate a higher level of proficiency in that area. A higher score could reflect no difference in proficiency on the second attempt, or it could even represent lower proficiency, again depending upon the relative difficulty of subject area items on the two administrations.

Law schools, which are understandably interested in the MBE performance of their students, could also misinterpret subscore information. For example, if a law school's students, on average, obtain higher subscores in Torts than in Real Property, the school's administration may interpret that score difference as indicating that the school's students were more proficient in Torts and perhaps conclude that the faculty members teaching Torts were more effective than those teaching Real Property. However, as noted earlier, there are more Torts items on each exam than Real Property items, so one would expect Torts subscores to be somewhat higher. Further, the Torts items on an exam could be overall intrinsically easier than the Real Property items, so a higher mean score on Torts could reflect no difference in the students' proficiency in the two areas, or it could even be that the students were more proficient in Real Property.

This kind of potential misinterpretation can cause other problems if law schools are interested in making changes to courses over time and using the mean subscores to evaluate those changes. A higher mean Evidence subscore may or may not indicate a positive outcome from changes made in an Evidence curriculum. As explained above, a higher mean score in Evidence in 2013 could actually reflect a less proficient student performance than the lower mean score in 2012. Without scaling the subscores, these types of

direct comparisons cannot be made, and even with scaled subscores, the associated levels of measurement imprecision would substantially complicate any comparisons.

Subscore Reporting Hinders the MBE's Ability to Evolve

The practice of law continues to evolve, and the MBE must evolve in concert to remain a valid assessment of the newly graduating lawyer's ability to practice law. In 2012, NCBE completed a job analysis of newly licensed lawyers aimed at providing evidence to support the extent to which NCBE's four tests are evaluating examinees in ways that are relevant to entry into the legal profession. In response to the information gathered from this effort, new subject areas are likely to be added to the MBE over time. To add this new content will require either a substantial increase in the number of test items and therefore the length of the MBE, or a change in the relative allocation of items to all of the subject areas covered by the MBE (as is being done in 2015 to accommodate the addition of Civil Procedure). If reporting of subscores were to be continued, and in order for the subscores to meet acceptable testing standards, there would need to be enough items to provide a reliable score for each subject area reported. That alone would require greatly lengthening the MBE. As described earlier, the MBE in its current form would have to at least double its length just to meet reliability standards; it would have to perhaps triple or quadruple its length to accommodate new subject area content.

Further, as the practice of law continues to evolve, we are likely to find that the importance of the various existing subscore topics may change and new content areas may rise in importance. These evolutionary changes will create new limitations in allocating items to different subject areas to maintain reliable subscores. By only reporting the scaled total score, minor adjustments can be made to the internal allocation of items to subject areas without concern for the impact on subscore reliability. The benefit of no longer reporting subject area subscores has been demonstrated with the addition of Federal Civil Procedure in 2015. The allocation of items to the existing six subscores will be reduced from 31–33 items per subject area to 27–28 items per subject area to make room for the new content without increasing the length of the MBE or compromising the reliability of the total score. Had we needed to keep subscores at their current, albeit low, levels of reliability in order to report subscores, we would have had to increase the length of the MBE by at least 27 items to accommodate the new content.

Conclusion

In summary, subscores are being eliminated for the following reasons:

- they do not add any information that is not already present in the scaled total scores,
- they are insufficiently reliable to meet testing standards or even for use in guiding student study or curriculum management by law schools,
- over a decade of research has concluded that they are not statistically distinct from the scaled total scores,
- they are often misinterpreted by stakeholders, and
- their continued use would make the MBE too inflexible to respond to the changing legal environment.

Eliminating reporting of subscores is not a new idea. A search of our archives found such a recommendation from 1975 by the Educational Testing Service, the first testing vendor serving NCBE. So, why has it taken us so long? First, the arguments at that time were all conceptual, and there were no standards for testing and assessment. The recommendation was really only a suggestion. It could be (and was) ignored. Since that time, the testing industry has adopted standards that should be considered more seriously. Second, there were no data to back the recommendation. Since the turn of the present century, NCBE has accumulated research expertise that has enabled us to conduct a series of studies aimed at demonstrating the distinctiveness of the subscores. As described in this article, these studies have failed to do so. So after over 40 years, the time has finally come to say good-bye to subscores.

NOTES

1. *See* C.A. Stone & C-C Yeh, *Assessing the Dimensionality and Factor Structure of Multiple-Choice Exams: An Empirical Comparison of Methods Using the Multistate Bar Examination*, 66(2) EDUC. AND PSYCHOL. MEASUREMENT 193–214 (2006); M. LANGER, "The Consideration of Subscores: Dimensionality Analyses of the Multistate Bar Examination" (presentation, annual meeting of the American Educational Research Association. Vancouver, BC, Canada. April 13–17, 2012); M.A. Albanese, "MBE Subscore Validity and Viability" (National Conference of Bar Examiners, November 2011).
2. *Id.*
3. AMERICAN EDUCATIONAL RESEARCH ASSOCIATION, AMERICAN PSYCHOLOGICAL ASSOCIATION, AND NATIONAL COUNCIL ON MEASUREMENT IN EDUCATION, STANDARDS FOR EDUCATIONAL AND PSYCHOLOGICAL TESTING (American Educational Research Association, 1999).
4. *Id.*, at 20.
5. You might find it confusing to think that an item specifically designed to measure knowledge of Torts, for example, could have a higher correlation with other subject items on the test than with the other items designed to measure Torts knowledge. There are several reasons why this is so. First, each of the items on the MBE is designed to assess the examinee's ability to analyze fact patterns. The skills to do so are not limited to the particular content the item is based upon. Further, it is often the case that an MBE item has elements that are related to other subscores. So, while the main thrust may be Torts, the item may have elements that relate to one or more of the other subscore content areas. Finally, an examinee who performs well in one area of law is also likely to perform well on all of the other areas, so his or her overall performance will be relatively consistently good. On the flip side, the same tendencies exist for the less proficient examinee. An examinee does not fail the bar examination because he or she lacks proficiency in Torts alone; it takes a more widespread lack of proficiency to fail the bar examination.
6. J.C. NUNNALLY, PSYCHOMETRIC THEORY 226 (McGraw-Hill Book Company, 1968). In "settings where important decisions are made with respect to specific test scores, a reliability of 0.90 is the minimum that should be tolerated."
7. D. Ripkey & S. Case, "An Analysis of Subscore Differences" (National Conference of Bar Examiners, December 2013).
8. ACT, MULTISTATE BAR EXAMINATION TECHNICAL REPORT: MBE FORM 710 (ACT, Inc., July 2010).
9. S. Sinharay, *How Often Do Subscores Have Added Value? Results from Operational and Simulated Data*, 47 J. EDUC. MEASUREMENT 150–174 (2010).
10. *Supra* note 1.

MARK A. ALBANESE, PH.D., is the Director of Testing and Research for the National Conference of Bar Examiners.