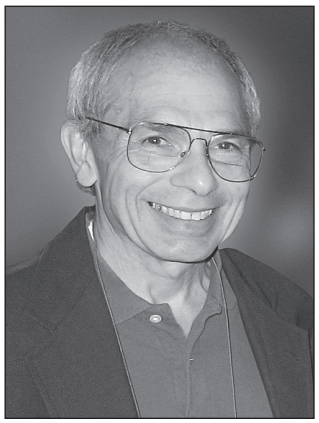
The Testing Column

Subscore National Percentile Ranks: The Undead

by Mark A. Albanese, Ph.D.



After three years in the grave, performance indicators on the MBE content areas have returned with the February 2017 bar examination results. However, they have returned in a form that will be less likely to lead to misinterpretation. Performance in each content area will now be reported as the percentile rank (PR) based upon national data—that is, the percentage of all examinees taking the same administration of the bar examination who scored below the score of a given examinee. Formerly, performance was reported as subscores, which consisted of the number of items answered correctly (raw score) in each content area, and the PR was based upon the specific jurisdiction where the bar exam was taken.

Jurisdictions will receive PR indicators with their score reporting materials, and they will decide what to do with the information, including whether to release it to examinees and/or law schools. This resurrection of performance indicators on the MBE content areas provides an opportunity to review why we eliminated the previous subscore information in the first place and then to suggest uses to which the new national PR values can be validly applied.

Why NCBE Eliminated Subscores

Subscores were eliminated three years ago because numerous studies had found that the MBE raw subscore measures provided nothing distinct from what was already contained in the total scaled score. Due to the particulars of how the MBE is constructed, raw subscores and their jurisdiction-based PRs are less reliable and potentially distorted versions of the total scaled score and its PR.[1]

Raw subscores are less reliable than the total scaled score primarily because they are based upon one-seventh of the number of items (25 items for each content area subscore versus 175 items for the total scaled score).[2] In addition to the inherent unreliability of subscores based on relatively few items, raw subscores are not equated, so they do not have consistent meaning across administrations, due to variation (albeit minor) in the intrinsic difficulty of the items. Without equating a test score, it is impossible to determine whether the particular performance at any given time results from variation in the intrinsic difficulty of the items or fluctuations in the proficiency of the examinees taking the bar exam at that time. Any law school trying to use the raw subscores to improve its curriculum or to evaluate its teachers could be doing more harm than good if changes are made in response to declines in subscores that are really due to minor variations in the intrinsic difficulty of the test items. Further, any examinee trying to concentrate his or her studies on perceived weaker content areas as suggested by the pattern of raw subscores could just as likely be hurting prospects of bar passage as helping. Because subscores are based on a relatively small set of items, an apparent difference in raw subscore profile across content areas could be

due to the particular sample of items included on a test form. As an extreme example, if the Torts items used on a particular administration of the MBE happened to be the only 25 Torts questions an examinee could answer correctly, the examinee's raw subscore on the Torts items would look wildly better than it should. Similarly, if the items selected for Real Property happened to be the only Real Property questions out of hundreds that an examinee couldn't answer correctly, the examinee's Real Property raw subscore would look much worse than it should. We aim to sample broadly from each of the seven content areas on the MBE in our selection of 25 items per topic, so these examples are indeed extreme.

While the content areas underpinning the MBE are critical for building the MBE, their value as separate indicators of performance are somewhat limited. Think of the subscores as you would the building blocks of a house. Without any single building block in the foundation, the house could collapse. However, a single building block tells you little about the house it was used in constructing. So, while building blocks are essential to the creation of a structure, they are not intended to stand alone. Similarly, the content areas of the MBE are essential supports in constructing the MBE, but they are not built to stand alone.

What has changed to cause subscore performance indicators to be raised from the dead? In the current environment, with law school applications and enrollments at 30-year lows and MBE mean scores at similar lows, examinees and law schools have high interest in anything that might help improve scores. With this in mind, we went back to the drawing board to see if there was anything we could provide that would meet users' desire for easily interpretable content-level performance information. The result is that we will provide to every MBE jurisdiction PRs based upon national data for the total scaled score and for the raw subscore for each of the seven content areas beginning with the February 2017 exam administration.

In the remainder of this article, I describe the features of these new performance indicators and suggest ways in which they can be validly applied. Data from the February and July 2016 administrations of the MBE are used for illustrative purposes. It is important to note that the data are *projections* of what the results will be in February 2017.

Percentile Ranks Provide Clear and Unambiguous Interpretation

The PR for the *total scaled score* tells examinees their relative performance overall compared to the group of examinees in all jurisdictions taking the test during the same test administration. Total scaled scores are equated, meaning that they are statistically adjusted to compensate for differences in the intrinsic difficulty of the items. They also maintain consistent meaning across time, in part because they are equated, but also because they are based upon test forms that are carefully created to match the same content and statistical specifications to the extent possible.

While the total scaled scores have consistent meaning and are stable, the PR for a particular total scaled score may vary depending on the relative proficiency of examinees testing at a given time. For example, an examinee with a total scaled score of 135 in July 2016 would receive a national PR of 37%. That same total scaled score in February 2016 would receive a national PR of 49%. The score of 135 in both February 2016 and July 2016 represents the same level of competence, but the PRs (the indications of performance relative to all other examinees testing at the same time) are 12 points different because the composition of examinees in the two administration periods are different. (Historically, in July, fewer than 20% of the examinees are repeat takers, whereas in February more than 60% are repeat

takers; most repeat takers have previously failed the bar exam, and they generally do not perform as well as first-time takers.)

As a reminder, the subscore PR values are not based upon equated scores, but simply on the raw subscore (the number of items answered correctly) on the 25 items in each content area. The result is that a given raw subscore can yield wide variation in PR across content areas. For example, let's consider a raw subscore of 15. This subscore is 60% correct out of the 25 items in each of the seven content areas. In July 2016, this percent correct would have projected to a scaled score of approximately 134. The PRs for a raw subscore of 15 in July 2016 ranged from 26% to 39% across the different content areas. In February 2016, the corresponding range was from 29% to 46%. Generally, a larger percentage of examinees failed to correctly answer 15 or more items in a content area in February than in July. The best interpretation of the content-area PRs is in relation to the PR for the total scaled score. If a content-area PR is substantially lower than the scaled-score PR, it might represent an area where additional study could be beneficial. What constitutes being "substantially lower" is an open question at this point. As we gain more experience with examinee profiles of PR values, more clarity may arise.

National Data Increases the Stability of Percentile Ranks

Providing PRs based upon a national group of examinees will be more stable than PRs based upon an individual jurisdiction, especially for small jurisdictions. The stability of PRs is directly a function of how many examinees contribute to providing the data. For example, if there are 30 examinees in a jurisdiction, and an examinee's score is better than the scores of 6 of the other examinees, the examinee's local PR is 6/30 = 20.0%. However, taken to extremes, if the examinee's PR is calculated based on the national examinee population of approximately 50,000 in July, rather than the 30 from the jurisdiction, and the examinee's score is better than the scores of 6 of the other examinees nationally, the PR would be 6/50,000 = 0.01%. The 20% PR based on the jurisdiction is a less stable estimate of the examinee's PR than is the 0.01% PR from the national distribution, because the sample of 30 examinees in the jurisdiction does not adequately represent the total examinee population.

Jurisdictions also vary so much in their size and in the composition of their examinee populations that the PR values produced at that level are much less dependable than those based upon the national examinee population. This is a factor of the number of examinees tested in a jurisdiction, the number of examinees repeating the examination, and the law schools from which those examinees graduated. For example, a total scaled score of 135 had February 2016 PR values across jurisdictions that ranged from 19% to 75% versus a national PR of 49%. The PR values for content-area subscores show similar if not greater instability across jurisdictions. The PR values obtained from the national examinee group are therefore going to be better for judging an examinee's performance than those obtained from any jurisdiction. This is particularly true for the UBE jurisdictions, where a score is portable.

Percentile Ranks Across Content Areas Can Identify Areas of Relative Strength and Weakness

The most useful score NCBE provides is the MBE total scaled score, which is equated so that scores reflect consistent levels of performance across administrations.[3] The PR associated with the total scaled score provides an overall estimate of an examinee's performance among all those taking the examination during that test administration. The extent to which the subscore PR values depart from the overall total

scaled score PR may reflect the degree to which the examinee performed differently in a given content area. Within a given administration, the subscore PRs can reveal whether an examinee maintained his or her relative position within the examinee population across content areas or whether the examinee displayed one or more areas of relative dominance or weakness. The anchor for the interpretation of content-area subscore PR values is the total scaled score and its PR value. The total scaled score reflects performance relative to passing or failing the bar examination. If the total scaled score is substantially below the cut score, it will generally take improvement across all content areas to make up the difference. Targeting study to one or two areas of particularly poor PR values is not likely to make up the difference, even if successful in improving performance in those content areas, unless there is improvement in the other areas as well.

For law schools hoping to use PR information for course/curriculum evaluation, the situation is particularly challenging. Students are highly motivated to pass the bar and will do what they must to succeed. If they think they have had a bad course in a content area, they are likely to spend more time studying that area. Bar prep companies also have a stake in seeing their students succeed, so if diagnostic tests show students testing poorly in an area, they are likely to emphasize it more in their preparation activities. Somewhat perversely, the MBE results can be a negative barometer in some cases. If students focus their bar preparation studies on the content taught in their bad courses and neglect studying for content taught in their good courses, their subscore PR values can look good for their bad courses and bad for their good courses. Compounding the problem, because content-area raw subscores are not equated, they have less stability than the total scaled score.

The best recommendations I can give are to make no irreversible decisions based upon results from a single administration and to use the subscore PRs only as a starting point for investigating whether a problem exists in a course or curriculum. Just because the bar examination results in a content area do not look bad, or even if they look good, does not mean that all is well. The national PR data is what it is, but its meaning for examinee study and law school curriculum purposes can be inscrutable. At its best, PR information will tell an examinee how he or she fared on a common test on a given day in comparison with thousands of other test-takers. May the force be with you.

References

Mark A. Albanese, Ph.D., "MBE Subscore Validity and Viability" (National Conference of Bar Examiners, November 2011).

Michelle Langer, Ph.D., "The Consideration of Subscores: Dimensionality Analysis of the Multistate Bar Examination" (presentation, Annual Meeting of the American Educational Research Association, Vancouver, British Columbia, April 13–17, 2012).

O. L. Liu, "Rasch and Multidimensional Rasch Analysis of the MBE Items" (National Conference of Bar Examiners, June 2007).

C.A. Stone and 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) *Educational and Psychological Measurement* (2006) 193–214.

P. Yin, "A Multivariate Generalizability Analysis of the Multistate Bar Examination," 65 *Educational and Psychological Measurement* (2005) 668–686.

Notes

1. For a detailed explanation of why raw subject area subscores were eliminated from MBE score reports, see my column in the June 2014 issue: Mark A. Albanese, Ph.D., "The Testing Column: Differences in Subject Area Subscores on the MBE and Other Illusions," 83(2) *The Bar Examiner* (June 2014) 26–31.
2. Twenty-five of the 200 MBE items are pretest items, which are items being considered for possible future use and are not used to compute the examinee's score. The number of pretest items increased from 10 to 25 effective with the February 2017 exam administration.
3. Many jurisdictions set their cut scores with the 200-point MBE scale as part of their consideration, often with passing score equivalents on the 200-point scale. This relationship gives real, consequential meaning to the MBE total scaled score.

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