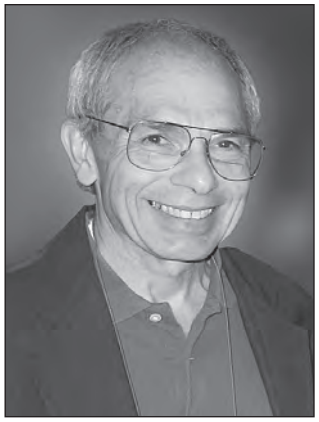
The Testing Column

Let the Games Begin: Jurisdiction-shopping for the Shopaholics (Good Luck with That)

by Mark A. Albanese, Ph.D.



It has been argued that variation in how Uniform Bar Examination (UBE) scores are produced across jurisdictions presents opportunities for examinees to shop around for a UBE jurisdiction that gives them the best shot at passing the bar exam.[1] A number of concerns related to grading consistency, fairness, and transparency have been raised, as well as a proposal that centralizing the grading process for the written portion of the exam would be a remedy to these ills. In addressing these concerns, I will start with what we know about the variability in grading among jurisdictions, assess the likelihood that examinees could benefit from shopping around for a more favorable jurisdiction in which to take the UBE, and close with what could be gained from centralizing the grading process.

Jurisdiction Variability in Grading Procedures and General Quality (i.e., Consistency)

In an ideal world, there would be no difference in the quality of grading among jurisdictions, particularly among UBE jurisdictions. But when we are talking about hundreds of humans looking at thousands or tens of thousands of essays, perfect consistency across graders, essays, time, and administrations is challenging and, perhaps, unrealistic.

Promoting Uniform High-Quality Grading

NCBE makes every effort to help jurisdictions follow uniform high-quality grading practices by holding grading workshops following each bar exam administration. Graders can attend these workshops in person at our Madison, Wisconsin, headquarters or attend by conference call. For the 2015 administrations, over 95 graders from at least 35 jurisdictions attended each workshop in person or by conference call. We also make available an on-demand video of the workshop proceedings, which was accessed extensively for each 2015 administration, and we provide secure online access to the written grading materials. The workshop and these supplemental materials provide ample opportunities for graders to review best practices before grading commences and throughout the grading process.

An article in the June 2016 *Bar Examiner* by Judy Gundersen, NCBE's Director of Test Operations, details all that NCBE does to help jurisdictions ensure and maintain high-quality grading.[2] We do not know exactly what fraction of all the graders participate in the grading workshops (jurisdictions do not furnish us with grader rosters); however, we know that most UBE jurisdictions do have at least some (possibly all) of their graders participate in the training (participation in the grading workshop is not a condition of use for the UBE). Those who participate review detailed grading guidelines generated by NCBE and gain hands-on practice grading essays. We expect that many of the individuals who participate in the grading workshop take back what they learn and share it with the other graders in their jurisdictions.

Does all this effort in grader training result in uniformly high-quality grading across UBE jurisdictions? That is a difficult question to answer with absolute certainty. We do know that there is variability in the grading scales used by different UBE jurisdictions. The majority of UBE jurisdictions use the 1–6 scale that NCBE recommends for rank-ordering papers, but others use a 1–5 scale, and some may extend the scale up to 100, although the maximum number of different grade values used by any UBE jurisdiction in February 2016 was 14. There is also variability between UBE jurisdictions in how many graders grade each paper, the size of the "training set" used to give graders a sense of the range of likely essay quality, and whether or not graders all grade a subset of the same papers to achieve and monitor grading consistency (i.e., calibration). Even though NCBE makes recommendations for best practices in grading,[3] jurisdictions maintain discretion in precisely which of the practices to adopt.

Monitoring Jurisdiction Variation

In addition to our efforts to promote uniform high-quality grading through training opportunities, we monitor possible variation in grading practices across jurisdictions. We compute two indices that reflect both consistency and quality in grading.

The Reliability of the Written Component Total Score

The first index we use is the reliability of the written component total score. (As a reminder, reliability estimates the extent to which a group of examinees would be rank-ordered the same if a second similar test was administered.) This index ranges from 0 to 1.0, with 1.0 meaning that there is consistent performance across the different essay questions and MPTs. If the reliability is 1.0, we could swap out different essay questions and MPTs and the score would not change for any examinee. A 0 reliability means that there is no consistency in performance from one essay question or MPT to the next. If we were to swap one essay question or MPT for another, the examinee's score could change dramatically: theoretically, an examinee could move from having the lowest score to the highest score or vice versa if different questions were selected.

The reliability of the written component total score becomes larger as more scores contribute to creating the total score, because it reflects a larger sample of performance. For comparison purposes, the 190-item MBE for recent administrations has a reliability of 0.92; for the July 2016 administration, it had a reliability of 0.93.[4] For an examination like the written portion of the bar exam with only eight different scores in UBE jurisdictions (one for each of the six MEEs and two MPTs), the reliability will be much lower. In fact, if we project the MBE reliability of 0.92 to an eight-item multiple-choice test, the reliability of such a test would be only 0.33. However, because each MEE question is a 30-minute exercise and each MPT is a 90-minute exercise, we would expect the written component total score to be substantially more reliable than the score from a handful of multiple-choice items that we expect would take only about 15 minutes to answer. (The MBE has 100 questions per three-hour session, allowing 1.8 minutes to answer each question; projecting the same amount of time to answer each question in an eight-item MBE results in a session lasting about 15 minutes.)

In July 2015, the reliabilities of the written component total scores for the 14 UBE jurisdictions ranged from 0.62 to 0.82 and averaged 0.73. In February 2016, the reliabilities of the written component total scores for the 17 UBE jurisdictions ranged from 0.48 to 0.77 and averaged 0.72. So, there is variability in the reliability of the written component total scores generated in the different UBE jurisdictions. A bigger problem is that even the highest reliability achieved in any jurisdiction (0.82) does not reach 0.90, the minimum level normally considered adequate for high-stakes testing purposes. (Further, in the case of non-UBE jurisdictions, if local essay questions are used, they have been found to have even lower reliability than the MEE/MPT written scores.)[5] Fortunately, most jurisdictions (and all UBE jurisdictions) make their pass/fail decisions based on the more reliable combined total score, not on the basis of the written component score alone (i.e.,

most jurisdictions do not have separate hurdles for the MBE and the written component).

The Correlation of the Written Component Score with the MBE Scaled Score

The second index we examine to monitor possible variation in grading practices is the correlation of the written component score with the MBE scaled score. Both the written component and the MBE are designed to assess knowledge, skills, and abilities required of the newly licensed lawyer. The MBE has the advantage of covering a broad range of content in the somewhat limited manner available in the multiple-choice format, while the MEE and MPT cover a more limited range of content but have the advantage of doing so by requiring the examinee to demonstrate the ability to express thoughts in writing, a critical skill for the newly licensed lawyer. Although they have obvious differences, the two parts of the exam do fundamentally measure similar abilities, so the consistency of the two scores may be considered an indicator of consistency of grading of the written component.

In July 2015, the correlations of the written component score with the MBE scaled score ranged from 0.44 to 0.81 and averaged 0.66 across the 14 UBE jurisdictions. In February 2016, the correlations ranged from 0.51 to 0.67 and averaged 0.60 across the 17 UBE jurisdictions. (When these correlations are adjusted for their less-than-perfect reliability, they are generally above 0.60, indicating that the MBE and written components "assess some shared aspects of competency, and that each method also assesses some unique aspect of competency.")[6] As was the case for the reliability of the written component total score, there is variability between jurisdictions in the correlation of the written component score with the MBE scaled score. But is it a difference that makes a difference? And further, does it reflect meaningful differences in the average preparedness of examinees within a jurisdiction on particular subject areas?

The two key variables used to scale the written component scores to the MBE are the mean and the standard deviation (SD) of the MBE in the jurisdiction. (As a reminder, the mean is the sum of scores divided by the number of scores; the standard deviation can be thought of as the average deviation of scores from the mean.) In July 2015, the mean MBE scores for the 14 UBE jurisdictions ranged from 134.94 to 147.18, and the SD ranged from 12.88 to 17.62. In February 2016, the mean MBE scores for the 17 UBE jurisdictions ranged from 126.55 to 146.20, and the SD ranged from 12.68 to 16.39. Thus, UBE jurisdiction mean MBE scores varied by 12.2 points in July 2015 and by 19.7 points in February 2016. There clearly is jurisdiction variability in the mean MBE scores and the SDs as well.

Could Examinees Change Jurisdictions to Improve Their Scores?

Could examinees capitalize on these differences among jurisdictions to enhance their scores on the written portion of the bar examination? For several reasons, it is highly unlikely that examinees could successfully "game the system" in this way.

Relative Grading Leaves Little Variation in Mean Raw Written Component Scores across Jurisdictions

A particular challenge to such an enterprise is the use of relative grading in most UBE jurisdictions. The relative grading approach is used to prompt graders to use every category in the grading scale (whether the scale used is 1–6, 1–5, etc.). Without this type of effort, it has been found that a sizable number of graders will pile their grades into the middle categories, which will bias or at least underweight the grades they award compared to those of other graders who use the full range of grading categories. Relative grading reduces the variation in grades awarded across jurisdictions and makes any differences that do exist inscrutable.[7]

For example, in July 2015, the mean of the unweighted raw written component scores for the 10 UBE jurisdictions using the six-point scale varied between 28 and 29 for 5 of the 10 jurisdictions, one had a mean of 27, and the remaining four varied

between 30 and 34. In February 2016, 7 of the 13 UBE jurisdictions using the six-point scale had means that ranged between 27 and 28, two had means of 26, and the remaining four ranged between 29 and 34. The majority of jurisdictions have so little variation in their mean raw written component scores that distinguishing them on that basis is a useless enterprise, and the remainder of jurisdictions have means that are mostly higher, which would complicate the search for a jurisdiction with low essay performance. All of this effort to win the jurisdiction-shopping game would be further complicated by the fact that the SD of the raw written component score, a factor that also plays into scaling the written score, is quite variable, ranging from 4 to 9 in February 2016, and complicating matters even further, jurisdictions' mean raw written component scores correlate essentially 0 with their mean written component scores scaled to the MBE (correlation in July = 0.01, in February = 0.00).

What Would the Ideal Jurisdiction Be?

Even if an examinee did believe that the bar exam could be "gamed" by jurisdiction-shopping, it would be very challenging to successfully determine to which jurisdiction one should go to enhance one's written component score. It has been argued that going to a low-performing jurisdiction would enhance one's written component score because the performance would look better among the others compared to how it would appear in a high-performing jurisdiction.[8] The fly in the ointment with this approach is that the low-performing jurisdiction may make the performance on the written component look better, but it would have to look sufficiently better that it would overcome the lower mean MBE score in the jurisdiction, because the scaling of the written component score to the MBE sets the mean of the written component score to the mean of the MBE. From the data above, there is more variability in the MBE mean component scores than there is in the written scores, so this could be a bigger hurdle than one might think.

There is no good way to model with existing data what would happen if a set of essays from a high-performing jurisdiction were graded in a low-performing jurisdiction (or any other jurisdiction, for that matter). At present, all jurisdictions grade their own essays within what is essentially a closed system. Even if graders do not use a relative grading approach, which directly compares the different essays with one another in assigning grades, the only context graders have for grading any particular essay is the group of essays produced by other examinees in that jurisdiction. Since, as shown earlier, there is substantial variation in jurisdictions in their MBE mean scores, reliability of the written component scores, and correlations of the written component scores with the MBE scaled scores, merging raw scores of essays graded in one jurisdiction with those of another cannot be assumed to produce scores equivalent to those that would be awarded if the actual grading of all the essays was done by graders in the jurisdiction. If the written component scores were scaled to the MBE before being merged, it would adjust for differences in jurisdictions in their MBE mean and SD, but it is not clear if these changes would adequately adjust for differences in jurisdictions in the reliability of the written component scores and in the correlation between the MBE scaled scores and the written component scores. To adequately determine how written component scores from one jurisdiction would change (or not) if the essays were graded in another jurisdiction, the essays would actually have to be graded in another jurisdiction. While such a study could be done, it would be up to one or more jurisdictions to do it, since they control the grading process. Even if a couple of jurisdictions banded together to see what would happen, it would be impossible to know whether the results indicated a pattern or were an idiosyncratic result specific to the particular jurisdictions conducting the study. To do a study large enough to give a firm answer would require a concerted effort on the part of a critical mass of jurisdictions over multiple administrations.

Because scores are anchored to the MBE mean of each jurisdiction, one could make a counterargument that examinees might do better going to a

high-performing jurisdiction and taking the bump up that the MBE mean would give them. However, it would have to be a better bump up than the hit they would take from having their written performance judged against the high flyers in the jurisdiction. Again, without actually sending written materials to different jurisdictions, it is not possible to know exactly what would happen.

How Could the Ideal Jurisdiction Be Identified?

A third challenge is that even if one was able to determine that a high- or low-performing jurisdiction would be the best to go to, it is not at all clear how such a jurisdiction could be identified. For starters, what constitutes high or low performance? Is it the bar passage rate? Is it the mean scaled written component score? If bar passage rate is the criterion, the standard score for passing varies across UBE jurisdictions from a low of 130 to a high of 140. The passing percentage tends to decline as the standard rises, so determining what is a high- or low-performing jurisdiction based upon passing percentages will have to account for the standard being used. If the criterion is the mean scaled written component score, it will be close to the mean of the scaled MBE score. However, the mean MBE score of a jurisdiction is not as stable as one might think. For example, the rank-order of the mean MBE scores for the 14 UBE jurisdictions in July 2015 compared with their rank-order in February 2016 differed by up to eight places. The highest-ranked of the 14 UBE jurisdictions in July 2015 was sixth from the bottom among the 17 UBE jurisdictions in February 2016. The median rank difference was three places. Only the bottom two jurisdictions maintained their relative ranking at both administrations.

So, there might be some hope in locating a consistently low-performing jurisdiction, but otherwise there is a lot of movement from one administration to the next. What may have been a high- or low-performing jurisdiction in the past may not be so in a future administration. It would be highly improbable that any examinee, any law school, any bar admission administrator, or NCBE itself could predict the rank-ordering of jurisdictions by average MBE performance with much chance for consistent success. Any hope for successful gaming of the system would also have to rely on insider information held by a relative few, and the information to which NCBE is privy cannot be disclosed because it is considered the property of the jurisdictions.

Even if there were a certain skill set that might give an examinee an advantage in a jurisdiction with particular characteristics (e.g., an examinee with strong writing skills looking for a jurisdiction with poor performance on the written component but strong performance on the MBE), large-scale attempts by examinees with the particular skill set to take advantage of the situation would be reactive, such that the jurisdiction characteristics they had hoped to capitalize upon would no longer exist due to the flood of examinees with high essay scores. This will likely always be true, but it is especially true in the current environment when examinees might be more likely to flock en masse to a jurisdiction administering the UBE, or to a jurisdiction with relatively strong employment prospects, given the weak legal employment market nationally.

The Feasibility of Centralized Grading

It has been argued that centralized grading of essays by a team of national graders rather than jurisdiction-specific graders, and scaling those scores to the national distribution, would increase consistency in scoring.[9] NCBE does not disagree that centralized grading would improve consistency in grading practices and procedures: centralized grading would mean use of a single grading scale, standardized recruitment and training of graders, and uniform quality-control procedures. There could also be some efficiency gained through use of automated essay grading systems that have been successfully used in some other professions. Computer grading can be used as a second grader for quality control or as a primary grader with human backup for papers receiving failing grades. From what we have learned in exploring this possibility, however, there

need to be at least 400–500 graded examinee papers to "train" the computer software and another 100 or so graded examinee papers to validate the algorithm employed by the trained computer software. For most jurisdictions, these numbers of examinees would exceed the number they test at any given time. If they were interested in having computer grading, they would probably have to band together with other jurisdictions to jointly grade their essays. This would not only provide the needed numbers of examinees, but would also provide the opportunity to amortize some of the fixed costs of examination over a larger number of examinees. However, computer grading aside, the most likely improvement in quality that centralized grading would provide would be in the reliability of the written component scores, particularly for those jurisdictions that currently produce the least reliable scores.

However, centralized grading is not a panacea: as long as humans provide grades for essays, it will be impossible to remove every potential idiosyncrasy. Further, jurisdictions would likely be reluctant to cede the control they currently have over the grading process, particularly the discretion to make adjustments to the grading guidelines as their boards of bar examiners or highest courts may demand. At present, and as far as we can tell at NCBE, there is very little desire by jurisdictions to have centralized grading instituted by us or by anyone else. However, as mentioned above, we are actively working with vendors to conduct research into the feasibility and quality of automated grading. If we determine that state-of-the-art machine grading has the potential to improve the fairness and quality of the grading process, we would work with jurisdictions to move in that direction. At present, however, jurisdictions contemplating and participating in the UBE have expressed a strong desire to maintain local control of grading, so the most feasible current solution for maintaining consistent meaning in written component scores across UBE jurisdictions (and indeed across bar exam administrations within UBE and non-UBE jurisdictions) is to scale the written component to the MBE.

Summary Remarks

In closing, if examinees think that they can find a jurisdiction that will give them a leg up when it comes to the grading of their essays and want to put the time into finding it, well, that is the American way. Our economy thrives on people who shop till they drop. It has made America great. So if you shopaholics know what characteristics a jurisdiction must have to let you play the game, and if you have the unknowable knowledge of how jurisdictions will perform on the administration when you plan to sit for the bar exam, and if you would rather jurisdiction-shop than use the time for study, shop away; there are 37 jurisdictions, 14 of which are UBE jurisdictions, that have no limit on the number of times you can take the bar exam. Of course, they may not be the ones you are shopping for, but they will be there for you when you have finished shopping.

Notes

1. Suzanne Darrow-Kleinhaus, *UBE-Shopping: An Unintended Consequence of Portability?* (March 30, 2016) Touro Law Center Legal Studies Research Paper Series No. 16-14, *available at* SSRN, http://ssrn.com/abstract=2756520.
2. Judith A. Gundersen, *It's All Relative—MEE and MPT Grading, That Is,* 85(2) The Bar Examiner 37–45 (June 2016).
3. Judith A. Gundersen, *The Testing Column: Essay Grading Fundamentals,* 84(1) The Bar Examiner 54–56 (March 2015); Susan M. Case, Ph.D., *The Testing Column: Procedure for Grading Essays and Performance Tests,* 79(4) The Bar Examiner 36–38 (November 2010).
4. The 190 items are those that are scored questions; 10 of the items are unscored pretest questions. Effective with the February 2017 administration, the 200 questions on the MBE will consist of 175 scored questions and 25 unscored pretest questions.
5. Susan M. Case, Ph.D., *The Testing Column: Relationships Among Bar Examination Component Scores: Do They Measure Anything Different?,* 77(3) The Bar Examiner 31–33 (August 2008).
6. *Id.*
7. For an explanation of relative grading, *see* Gundersen, *supra* note 2.
8. Darrow-Kleinhaus, *supra* note 1, at 7.
9. Darrow-Kleinhaus, *supra* note 1, at 9.

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