

THE TESTING COLUMN

ESSAY GRADING FUNDAMENTALS

by Judith A. Gundersen

As I write this column, bar exam graders across the country are in some stage of grading essays and performance tests. Every U.S. jurisdiction is responsible for grading the written component of its bar examination—whether the written component consists of the Multistate Essay Examination (MEE), the Multistate Performance Test (MPT), jurisdiction-drafted questions, or some combination of two or all three. Grading the written portion of the bar examination is a painstaking process that accounts for at least half of an examinee's grade—thus a significant component of the overall bar exam score. This column focuses on some essay (and performance test) grading fundamentals: rank-ordering, calibration, and taking into account an examinee's ability to communicate in writing. Adhering to these fundamentals helps ensure fair and reliable essay grading procedures and score results.



compared between the control group and current test takers. This equating process ensures comparable score meaning across MBE administrations.

But what about equating the MEE and the MPT? These tests cannot be equated in the same sense that the MBE is equated because their questions are too memorable to be reused or embedded in an exam—examinees spend 30 minutes on a given MEE question and 90 minutes on a given MPT question, as opposed to just a few minutes on an MBE question. Any examinee who had seen an MEE or MPT question before would remember it and have an advantage over an examinee who had never seen the question. (Once an MEE or MPT is administered, none of its questions is ever used again on another test form. Retired questions are made available for purchase or free of charge on our website as study aids or for use in law schools.)

First, a few words are in order about the role that equating plays in the overall context of grading. As stated many times in this column and elsewhere in the *Bar Examiner*, the purpose of the bar examination is to determine minimal competence to be licensed as an attorney. Both fairness to examinees and protection of the public dictate that the bar exam be reliable and valid across test forms and administrations. The Multistate Bar Examination (MBE) is the only part of the bar exam that is equated across all administrations. This is done by embedding a mini test form within the MBE with known statistical properties that is then

Because MEEs and MPTs cannot be equated in the same way as the MBE, but are a critical piece of the bar exam score, NCBE recommends the best practice of scaling the written scores to the MBE: raw scores earned on each MEE and MPT question are added up and then scaled to the MBE. This in effect puts the overall score earned on the written portion of the exam on the MBE scaled score distribution, thereby using the equating power of the MBE to give comparability to the written portion. Scaling preserves the important rank-ordering judgments that graders have made on answers.¹

RANK-ORDERING PAPERS

MEE and MPT questions are developed to test content and skills set forth in the MEE subject matter outline and the MPT list of skills tested. Within each MEE and MPT, multiple issues are raised that might be addressed by examinees—some issues are easier to identify and some are subtler. Multiple issues help graders make meaningful grading distinctions among papers. Some papers should get high scores, some average scores, and some lower scores, regardless of what score scale a jurisdiction uses (1–5, 1–6, 1–10, etc.), and regardless of whether, taken as a whole, all papers are strong or weak. What matters is rank-ordering among papers—relative grading.

Rank-ordering works best if distinctions are made between papers and scores are spread out over the whole score scale (whatever that may be). For example, if a jurisdiction uses a 1–6 scale (a “1” paper being a very poor answer relative to the other answers in the jurisdiction, and a “6” paper being an excellent answer relative to the other answers in the jurisdiction), it is important that graders assign 1’s, 2’s, 3’s, 4’s, 5’s, and 6’s, not just compress all of their grades between 3’s and 4’s. Were a grader to give every answer in her group of papers a “3,” for example, the question would, in effect, be thrown out—it would have no impact on examinees’ scores. It would be like keying all answers correct in a multiple-choice question. Similarly, but to a lesser degree, bunching all grades between just two of the points on a 6-point scale would diminish the relative value that this particular question would have on an examinee’s overall written score.

To prepare graders, NCBE provides detailed grading materials, which are subjected to review by outside content experts, editing by drafting committees, and proofing and cite-checking by NCBE lawyer-editors. User jurisdictions also have the option of reviewing the questions and grading materials before administration. NCBE hosts an MEE/MPT grading workshop after each administration, with three participation options for graders: in person, by conference call, or

via on-demand streaming. Finally, the grading materials are included in MEE and MPT study aids, so prospective examinees can become familiar with the questions and what graders are looking for in examinee answers.

Rank-ordering papers is harder when a grader perceives that the answers are all very good or all very poor. But meaningful distinctions between papers can and should be made no matter whether a paper evidences a weak or strong performance. That is, a grader should take into account an examinee’s use of the facts, the quality and depth of the examinee’s legal analysis, the examinee’s issue-spotting ability, and the quality of the examinee’s writing (more on this later). Considering each paper as a whole, informed by the grading materials, rank-ordering papers using the entire score scale will best ensure that examinees’ written scores reflect their performance on this portion of the exam.

ACHIEVING AND MAINTAINING GRADING CONSISTENCY: CALIBRATION

Whether a grader grades all the answers to a certain question himself or with other graders, getting and staying calibrated is critical. Calibration is the process by which a grader or group of graders develops coherent and identifiable grading judgments so that the rank-ordering is consistent throughout the grading process and across multiple graders. It shouldn’t matter to an examinee if her answer is paper number 1 for grader A or paper number 233 for grader B.

To calibrate, graders begin by reading a set of 10 or more common papers and assigning tentative grades. Multiple graders compare their grades on the sample group and see where they need to resolve grading judgments. Once any differences between grading judgments are worked out, then another sample group of 10 papers should be read to see if the graders are in alignment. Again, grading differences on this second set of sample papers must be resolved. Finally, a third set of 10 common papers might be necessary to ensure

that graders are grading consistently. If the total number of examinees or papers to be graded in an administration reaches the hundreds or thousands, it might be a good idea to embed a few common papers among multiple graders, those papers then being checked to ensure that consistency is maintained over the course of the grading process.

Single graders should also start with a defined set of papers to gauge what the pool of answers will look like and assign tentative grades until they've seen more papers. Because grading is relative and papers are to be rank-ordered, context is everything. Early grades will probably need rechecking as more answers are read. Some graders find it helpful to keep benchmark papers—representative papers for each point on the score scale—to help re-orient themselves after a grading break. It may also be helpful for a grader or graders to try to put papers in buckets or piles representing each point on the score scale to ensure that they are, in fact, using the whole score scale and not bunching all answers between two points on their score scale.

TAKING INTO ACCOUNT EXAMINEES' ABILITY TO COMMUNICATE IN WRITING


One way for graders to make distinctions between papers is to take into consideration examinees' ability to communicate in writing—this is a construct of the MEE and MPT and is set forth in the purpose statement of the MEE and the skills tested in the MPT. A lawyer's ability to communicate in writing is a critical lawyering skill. NCBE's 2012 job analysis confirmed this—100% of all respondents to the survey we distributed to new lawyers stated that the ability to communicate in writing was "extremely significant" to their jobs as lawyers.² If writing didn't matter, then the bar exam could consist solely of multiple-choice questions—which would save a lot of time and effort. But it does matter.

Demonstrating the ability to communicate in writing does *not* mean using legalese or jargon. Rather, it

means writing a well-organized paper that demonstrates an understanding of the law and how to apply it to the facts in the problem. It means, as stated in the MEE instructions, "show[ing] . . . the reasoning by which you arrive at your conclusions."

The MPT has more specific criteria for assessing the quality of an examinee's writing than the MEE, as MPT examinees are instructed on the proper tone for the assignment (e.g., persuasive, objective), the proper audience (e.g., court, client, opposing counsel), and sometimes the desired formatting (e.g., the use of headings, statement of facts, case citations). Thus, in general, it can be easier for graders to make distinctions on the quality of writing when grading MPTs. However, graders can make a meaningful assessment of writing ability on both the MPT and the MEE.

CONCLUSION

Graders have an important job, and they know it. I've met hundreds of graders over the years, and they all strive to make consistent and fair decisions, and take their jobs very seriously. Employing the practices and principles of rank-ordering, achieving and maintaining calibration, and assessing written communication ensures a fair and reliable process for grading the all-important written portion of the bar examination. 

NOTES

1. For a detailed explanation about scaling, see the December 2014 Testing Column: Mark A. Albanese, Ph.D., *The Testing Column: Scaling: It's Not Just for Fish or Mountains*, 83(4) THE BAR EXAMINER 50–56 (December 2014).
2. The NCBE job analysis is part of a content validity study conducted by NCBE in conjunction with its testing program. The job analysis was carried out through a survey distributed to a diverse group of lawyers from across the country who had been in practice from one to three years. Its goal was to determine what new lawyers do, and what knowledge, skills, and abilities newly licensed lawyers believe that they need to carry out their work. The job analysis, entitled *A Study of the Newly Licensed Lawyer*, is available on the NCBE website at <http://www.ncbex.org/publications/ncbe-job-analysis/>.

JUDITH A. GUNDERSEN is the Director of Test Operations for the National Conference of Bar Examiners.