It's All Relative—
MEE and MPT Grading, That Is

by Judith A. Gundersen[1]

The Multistate Essay Examination (MEE) and the Multistate Performance Test (MPT) portions of the bar exam are graded by bar examiners in user jurisdictions. They are not centrally graded at NCBE, but NCBE prepares detailed grading materials for both exams and provides hands-on grader training designed to facilitate consistent, accurate, and fair grading across all MEE and MPT user jurisdictions and, in particular, for Uniform Bar Examination (UBE) jurisdictions. It is critical that grading of the MEE and MPT portions of the UBE be consistent across UBE jurisdictions, as the score portability afforded by the UBE is based on the assumption that the exams are graded in a consistent manner no matter where graded or by whom.

This article discusses the relative or rank-ordering grading philosophy NCBE uses in its grader training, the reasons we advocate this approach, and recommendations for optimal use of this grading method. I'll start with a review of how grading materials are prepared at NCBE and the nature of our grader training.

Preparation of MEE and MPT Grading Materials

Grading materials for the MEE and MPT include the MEE analyses and MPT point sheets, which are detailed discussions of all the issues raised in the items by the item drafters and suggested resolutions or analyses of the issues. The analyses and point sheets are drafted by the authors of the items and are then discussed, edited, and revised by the respective NCBE drafting committees at semiannual meetings. Preparing MEE and MPT items and their grading materials takes at least two years and is an iterative process with many lawyer-volunteers, NCBE staff, outside content experts, and pretesters involved.[2]

MEE and MPT drafters know the importance of crafting excellent grading materials. The process of preparing grading materials also serves as a good check for the drafting team on the item's internal consistency, degree of difficulty, and gradability; it is quite common for grading materials to unveil problems with the item that were not identified by the drafter or committee at an earlier stage.

Grader Training

MEE and MPT grading materials are very thorough and can effectively guide graders through the grading process. In addition, NCBE also conducts hands-on grader training sessions at its Madison, Wisconsin, headquarters the weekend following the bar exam. Graders may attend the grading workshop in person, by conference call, or via on-demand streaming as available following the workshop. Participation by user jurisdictions is high—hundreds of graders representing most MEE and MPT jurisdictions participate in one of these three ways.

The grading workshop lasts one day and consists of a dedicated session for each MEE and MPT item led by drafting committee members who are experienced grading workshop facilitators. Sessions begin with an overview of the item and grading materials, and any questions about the area of law (MEE) or the assigned task (MPT) are addressed. The participants then set about silently reading several real examinee answers (sent by bar administrators from all over the country) and grading them. Grades are assigned using a 1–6 relative score scale (as discussed later). As professors often do in law school, workshop facilitators rely on the Socratic method from time to time—graders are called on to explain the grades they gave. This is particularly true if a grade might be an outlier from grades assigned by other graders in the session. Based on the review and grading of the sample of examinee answers and the ensuing discussion between graders and facilitators, grading materials may be refined or grading weights adjusted. Final versions of the grading materials are then made available to graders in user jurisdictions a day or so after the workshop.

Grading workshop participation alerts graders to common answer trends and also gives them a head start on calibration—the development of coherent and identifiable grading judgments so that rank-ordering is consistent throughout the grading process and across multiple graders. (The focus of this article is *not* on calibration, but that doesn't mean it isn't a critical component of the grading process. See the section on calibration below.)

The Relative Grading Philosophy in Action

What Is Relative Grading and How Does It Work?

With NCBE's grading materials in hand, graders are ready to begin the grading process in their own jurisdictions with their own examinees' answers. But grading MEEs and MPTs isn't like marking a paper with a score from 1% to 100% or meting out an A, B, C, D, or F (or drawing smiley or frown faces on papers; one of my sons' third-grade teachers, whom I will call Ms. Brinkman for purposes of this article, was fond of drawing a big ☹ on papers that didn't meet her standards!). Instead, NCBE trains bar examiners to grade the MEE and MPT on a relative basis—making distinctions between papers and rank-ordering them according to whatever score scale the jurisdiction has in place. (Jurisdictions may use whatever score scale they wish—e.g., 1–5, 1–6, 1–10, etc.—although NCBE uses a 1–6 score scale at its grading workshop, for reasons detailed later in this article.)

Relative grading training helps graders identify consistent standards in ranking papers and then apply those standards to put papers in piles according to their relative strength. The 1–6 scale used at the workshop simply means that a score of 6 is reserved for the best papers among all answers assigned to a particular grader. It is better than a 5, which is better than a 4, and so on, all the way to 1—a paper that is among the weakest papers. Relative grading means that in any group of answers, even if no single paper addresses all the points raised in an item, the strongest papers still deserve a 6 (using a 1–6 score scale). They do not have to be perfect nor necessarily deserve a true A or 100% (or a ☺☺ according to Ms. Brinkman). Using the same principles, a paper need not be completely devoid of content to get a 1 if the other papers are strong.

This relative grading philosophy (also referred to as norm-based grading) may be a little different from the way many of us had our papers graded in school, where we were held to an "absolute" or "criterion-referenced" standard: we had to answer a certain number of parts of a question correctly to get a high score or an A regardless of how our fellow students answered. Or if we missed some points, we would get a low grade even if many of our fellow students also missed the same points.

NCBE's focus on relative grading does not mean, however, that absolute or criterion-referenced grading does not belong on the bar exam; it does, particularly on the Multistate Bar Examination (MBE)—the only part of the bar exam that is equated across time and across exam forms. Equating is the process of determining comparable scores on different exam forms. For the MBE, the absolute standard or "cut score" has the same meaning across administrations and jurisdictions. A scaled score (scaled means that it is a standardized score derived after equating) of 135 on the MBE is a 135 no matter when or where earned and will always mean that the examinee passes if the cut score is 135. By contrast, essays and performance tests cannot be equated in the way a multiple-choice exam like the MBE can be, so a total raw score of, say, 24 (or 100 or 1,000) on the written part of the bar exam may have a different meaning depending on the particular exam form, the examinee pool, the grader, and the jurisdiction.[3]

Because of the high-stakes nature of the bar exam, we must account for the differences in written exams across administrations, jurisdictions, and graders, and we do this by using the equated MBE score distribution as a highly reliable anchor. We weight the MEE and MPT raw scores for each examinee according to the jurisdiction's weighting scheme (e.g., on the UBE, the MEE is weighted 30% and the MPT 20%). We then map the total weighted MEE and MPT raw scores for each examinee to the MBE scaled score distribution according to performance level. This process is referred to as *scaling* and has the effect of adjusting the MEE and MPT scores so that they have the same mean and standard deviation as the MBE scores do in the testing jurisdiction (standard deviation being the measure of the spread of scores—that is, the average deviation of scores from the mean).

Scaling written scores to the MBE is a psychometrically valid practice because examinee performance on the MBE is strongly correlated to examinee performance on the combined MEE and MPT. Because the MBE is an equated exam, MBE scores have constant meaning across time and across jurisdictions, even though the items on particular exams may vary slightly in intrinsic difficulty. By scaling the combined MEE and MPT scores to the MBE scaled score distribution, we capitalize on (or leverage) the equating done to the MBE to give the MEE and MPT scores the same constancy in interpretation, despite the fact that MEE and MPT items may vary in difficulty from administration to administration.

It is important to point out that if the relative grading approach is used consistently across jurisdictions and administrations, the MEE and MPT raw scores will have the same mean and standard deviation in all jurisdictions and administrations no matter if the intrinsic difficulty of the MEE or MPT items changes or if the examinee population becomes more or less proficient. In jurisdictions that use the same grading scale, each jurisdiction will also have approximately the same raw score mean and standard deviation as well as having the same mean and standard deviation for all administrations. It is only by scaling to the MBE that differences in either the items or the examinees can be reflected in the scores.[4]

Why Use Relative Grading?

There are compelling psychometric and policy reasons why, given the current process for grading the MEE and MPT, NCBE trains graders to use a relative grading approach (with subsequent scaling to the MBE) to consistently grade the MEE and MPT.

Score Scales and Grading Procedures Vary Among Jurisdictions

Because of a decentralized approach to grading the MEE and MPT, no matter how successful we are at training graders across jurisdictions to promote uniformity, we must allow for the fact that there could be some scoring variation among jurisdictions.

Relative grading does not require that all jurisdictions use the same score scale. Rather, papers placed in a particular pile (assigned grade) reflect a level of proficiency that is more similar to others in the same pile than to papers placed in a different pile, and higher grades reflect higher degrees of proficiency. As stated earlier, an examinee's raw MEE and MPT scores are weighted appropriately, added together, and then mapped to the MBE scaled score distribution for the given jurisdiction. An examinee who performs well on all or most parts of the written portion of the exam will generally have scores that "land" on the upper end of the distribution of the MBE scaled scores for that jurisdiction. Someone who earns a lot of 6's on her MEE and MPT answers (in a jurisdiction using a 1–6 score scale) will generally have her total written score mapped to the top of the MBE scaled score distribution for her jurisdiction; an examinee who consistently earns 1's and 2's on his MEE and MPT answers will usually find that his total written score maps close to the bottom of the MBE scaled score distribution. This will be true no matter what score scale is used.

Relative grading is also adaptive enough to work with different approaches to the grading process. It does not matter if each paper is read by only one grader or by two graders who have to agree; or if a single grader grades all answers to a particular item or answers are divided among several graders. Nor does it matter if grading is done over the course of a day or weekend of intense grading or over the course of two months. As long as graders achieve and maintain calibration (as discussed later), relative grading should serve to keep answer assessment consistent across time and across graders.

MEE and MPT Items May Vary in Difficulty from One Administration to the Next

As much as the drafting committees and our test development process try to standardize MEE and MPT difficulty across exam administrations, it is impossible to create items that represent exactly the same degree of difficulty. And MEE and MPT items cannot be pretested live to gather performance data in the way that MBE questions can because they're too few and too memorable. (MBE pretest questions are indistinguishable among the scored MBE items on each MBE exam form.) Without live pretesting, we must find some other fair way to take into account differences in MEE and MPT difficulty across exam forms.

With relative grading, it doesn't matter if an exam form represents the exact degree of difficulty as past (or future) MEEs or MPTs. Relative grading means that an examinee who sits for a harder exam is not penalized and an examinee who sits for an easier one is not rewarded, because it focuses only on how examinees do in comparison to one another on the same exam. For example, suppose that February 2016 examinees were given more difficult MEE or MPT items than those administered in, say, July 2015. That would be unfair to the February 2016 examinees or, alternatively, would seem like a windfall to the July 2015 examinees *if* MEE and MPT items were graded according to an absolute standard. The July 2015 examinees would get overall higher scores because the items were easier. In the world of high-stakes tests like the bar exam, this is a situation to avoid, and relative grading helps do that. It focuses on *comparing* answer quality according to other answers to the same items. Answers to easy items are still rank-ordered, as are answers to harder ones.

Scaling the total raw score on the written portion of the bar exam to the MBE, which is equated across administrations and accounts for differences in exam difficulty, means that it doesn't matter whether the written portion on one administration is harder than on another. As long as graders are able to rank-order answers, they can fairly and consistently grade the MEE and MPT from administration to administration regardless of differences in exam form difficulty.

Examinee Proficiency Varies from One Administration to the Next

Examinee proficiency may vary across administrations. For example, in the February administration, examinee proficiency tends to be lower due to a larger proportion of repeat test takers. We see this lower performance reflected on the MBE in February and expect to see lower scores on the MEE and MPT as well. However, asking graders to maintain consistent grading standards across administrations, examinees, and items would be extremely difficult, if not impossible. There are simply too many moving parts across test administrations to make such a grading task reasonable for maintaining score meaning across administrations. But relative grading—comparing answers among the current pool of examinees and then scaling those raw scores to the MBE—is manageable for graders and fair to examinees.

It is also important to note that using a relative grading system rather than an absolute grading system does not mean that graders are artificially inflating or deflating grades in a way that allows more examinees to pass or causes more examinees to fail. All relative grading does is help graders make rank-ordering decisions, which are critical to having the question "count" in an overall bar exam score, as discussed below. Scaling to the MBE lines up an examinee's overall written score to a statistically accurate corresponding point on the MBE score distribution. Scaling standardizes rank-ordering decisions across time and exams.

Likewise, relative grading does not benefit or penalize examinees who sit in jurisdictions that have a weaker or stronger examinee pool. Relative grading practices work in tandem with the process of scaling to make the appropriate offset for each examinee's position relative to his or her own jurisdiction's examinee group and the position of that examinee group relative to other jurisdictions' examinee groups. To make meaningful and fair comparisons across time and jurisdictions, we need to know what absolute level of performance is represented by a particular group's average. We don't have that absolute performance information for essays, but we do have average performance on the MBE for the relevant groups. By virtue of the equating process, those scores are on an absolute scale.

Because the data have consistently shown across groups and time that the total MBE scaled score is strongly correlated with overall performance on the written components (correlation above .80 when reliability of the two measures is taken into account), we can use MBE performance information as a proxy indicator of the groups' general ability levels. As a result, an examinee whose total raw essay score is ranked at the top of a weak group will have, after scaling, a total scaled essay score that reflects that differential, and an examinee who is more toward the bottom of a strong group will have a total scaled essay score that accounts for that positioning as well. Similarly, offsets are made (via scaling) to account for an examinee who sits for an administration with easier essay questions or one who sits for an administration with harder essay questions. The scaling process is critical to ensure that scores have a consistent meaning and also to ameliorate any efforts at gaming the system by attempting to pick a group or an administration that is anticipated to behave a certain way (e.g., sitting for a test that is anticipated to be easy or sitting with a group that is anticipated to be particularly skilled).

Graders Vary in Harshness or Leniency

In addition to evening out the differences in MEE and MPT difficulty from one administration to the next, relative grading ameliorates grader harshness or leniency from one administration to the next and from grader to grader. Even a harsh grader in jurisdiction A has to distribute an array of grades, low to high, among papers if she uses relative grading—she can't give all papers a low grade because then she's not rank-ordering. A lenient grader in jurisdiction B grading the same item can't give all papers a high grade if he uses relative grading and follows instructions to use all grade values.

If a particular question is graded by a harsh grader or a lenient grader, as long as that grader is consistently harsh or lenient in rank-ordering, examinees are not unfairly penalized or rewarded—the rank-ordering decisions made by the grader remain; the actual raw scores assigned, whether harsh or lenient, are smoothed out to fit the MBE scaled score distribution. Examinees will not be penalized even if harsher graders have a lower mean score than lenient graders. (Note that if multiple graders are assigned to grade a single question, they must be calibrated so that they do not have different levels of harshness or leniency.)

Relative Grading Facilitates the Equal Weighting of All Items

Relative grading facilitates spreading out scores, which is critical to ensuring that all items carry the weight they should in an examinee's overall written bar exam score. The weight an item gets is strongly affected by the amount of variation that scores have on that item. The less variation, the less weight the item carries in determining the total written score value. A question that every examinee gets right doesn't discriminate or distinguish between examinees, just as a question that every examinee gets wrong doesn't discriminate. For example, a question asking an examinee to write the English alphabet wouldn't distinguish between examinees, because virtually everyone would get the answer correct. Or a question asking examinees to write the Burmese alphabet would probably stump 99% of U.S. examinees. In both instances, those questions would let us know that all examinees do know the English alphabet but don't know the Burmese alphabet, but they wouldn't provide any information to allow us to make distinctions *between* examinees based on their performance.

All MBE, MEE, and MPT items are designed to elicit information about examinee performance. Relative grading on the MEE and MPT, both of which have multiple issues per item, allows graders to gather information about examinee performance and assign a score that accurately reflects examinee performance. All MEE and MPT items are drafted, reviewed, edited, and pretested to ensure that graders will be able to spread examinee scores according to relative quality if they follow grading instructions properly.

Graders should award points or credit reflecting the spectrum of the score scale used in their jurisdiction to maximize and equalize the information provided by each MEE or MPT item. Consider the following examples of what happens when a grader fails to discriminate among answers. Suppose a jurisdiction uses a 10-point score scale and a grader is using an absolute or criterion-referenced approach (or, for that matter, a relative grading approach) and no examinees address all points raised in an item. The absolute grader won't award any papers a 10 or possibly even a 9, depending on how inadequate the answers are. And the relative grader, if not trained properly, might hold back and not award maximum points even to the best answers. So now that 10-point scale (also being used to grade other items), for purposes of *this* item, is a de facto 8-point scale, because no one is getting a 9 or a 10. Suppose a grader is even more extreme and uses only 5 points on a 10-point score scale—from 2 to 6, for example. The result is that the item has even less impact on examinees' overall written scores in comparison to other items that are being graded on all points of the jurisdiction's 1–10 scale.[5]

Another way of not spreading out scores is by bunching a large percentage of scores in the middle of the score scale. For example, on a 1–6 score scale, a grader who gives 75% of her papers a 4 (whether using absolute or relative grading), is, in effect, downgrading the weight given to that item. This particular question has elicited very little information about examinee performance and has compressed the score scale from 6 points to just a few points—and mainly to one point, a 4.

One reason why we emphasize relative grading is that it should help graders spread out scores—no examinee has to write a perfect paper to get the highest score, and no examinee has to leave the page blank to get a very low score. As long as a grader keeps that principle in mind, it should be natural to spread out scores. And the actual score distribution for each grader is easy to keep track of and need not be in equal piles. It is enough to make meaningful distinctions between relative examinee performances that reflect all or most of any given score scale.

What Is the Best Approach for Optimizing Relative Grading?

Using a Manageable Score Scale

While relative grading works the same no matter the score scale, it tends to work *best* and is easiest to manage using score scales that are relatively compressed. For example, if a grader uses a 1–100 scale, it's conceivable that the grader could make 100 piles of rank-ordered answers, but 100 separate piles representing qualitative differences between answers can be pretty hard to wrap one's brain around. And, of course, a 1–100 scale brings to mind grading as it's done in school—absolute grading. That is, a 90–100 is an A and is reserved for an answer that covers all possible issues in the item, as opposed to an answer that is the best of a possibly weak group of answers. Also, probably for many jurisdictions that use a 1–100 scale, their graders assign grades by 10's—that is, 10, 20, 30, etc.—so that the score scale is really functioning more like a 10-point scale, not a 100-point scale.

NCBE uses a 1–6 scale to train graders, in part, because six piles of answers are manageable and memorable. And we use a 6-point scale instead of a 5-point scale because a 5-point scale resembles the A, B, C, D, and F grading paradigm that makes it a bit too easy to bunch scores on the midpoint or average—a 3 or a C. Using a 1–6 scale means that graders can't just label an answer average, or a 3—they have to make a decision as to whether it's a 3 or a 4, that is, a little bit above an average paper or a little bit below. Because many graders tend to bunch their answers in the middle (rather than at the ends of the score scale), just by using a 1–6 score scale rather than a 1–5 scale, they have to make a choice around that critical midpoint, which makes bunching harder and spreading easier. Some jurisdictions use a 1–10 scale, which also works well provided that all points on that score scale are awarded.

Participating in NCBE's Grader Training

While relative grading is fairly intuitive, it is informed and standardized by the MEE and MPT grading materials and the post-exam grader training that puts the relative grading principles into action.

That training emphasizes reading through several answers before assigning final grades to those first several papers read, as a grader won't yet have a good idea of what the overall pool of answers looks like. NCBE's grader training and materials also assign weights to subparts in a question. So an examinee who performs well on one subpart of an MEE question worth 25% of the total score that could be awarded for that question is not assured a 6 unless he performs well on the other parts of the question, too, in comparison with other examinees. In other words, there is a weighting framework for assigning points, which helps to keep graders calibrated and consistent.

NCBE also offers online support and hands-on workshops that demonstrate how to use relative grading. For most graders, it is not hard to rate answers in order of their relative quality. It might be a little more difficult to use the entire score scale, whatever that may be, but practices as simple as keeping track of the number of score piles and the number of answers in each score pile go a long way toward keeping score distribution (or lack thereof) front and center during the grading process. And consistency can be maintained by keeping benchmark papers for each score pile to illustrate what representative answers look like for each score—6's, 5's, 4's, etc. This is particularly important if grading is done over an extended period of time or by multiple graders.

Ensuring the Calibration of Graders

No grading process will be effective if graders (especially multiple graders assigned to a single item) are not calibrated. It should not matter to examinees who grades their papers or when their papers are graded. Relative grading and absolute grading both require calibration to be consistent and fair. Under either grading method, calibration requires reading through several sample answers, a thorough understanding of and facility with the grading rubric, and agreement on the standards to be applied and how to apply them. In my experience, calibration is an exercise to which jurisdictions devote substantial resources, time, and verification to ensure that graders become and remain calibrated throughout the grading process.[6]

Conclusion

The relative grading approach is employed widely in the jurisdictions that administer the MEE and the MPT. When used in conjunction with scaling to the MBE and proper training and calibration of graders, relative grading promotes consistency and fairness in grading the written portion of the bar exam. It compensates for the use of varying score scales and grading procedures among jurisdictions as well as differences in the harshness or leniency of the graders themselves. It neither artificially inflates or deflates grades but facilitates the spreading out of scores, which are then scaled to the highly reliable anchor of the equated MBE score distribution. Finally, it ensures that any variation in difficulty of items from one administration to the next does not penalize or reward examinees, while facilitating the appropriate weighting of all items so that each item provides information about each examinee's performance.

Notes

1. NCBE Testing Department staff members Dr. Mark Albanese, Dr. Joanne Kane, Dr. Andrew Mroch, and Douglas Ripkey all provided assistance with this article.
2. For a detailed explanation of how MEE and MPT items and their grading materials are prepared, see my article in the June 2015 *Bar Examiner*: Judith A. Gundersen, *MEE and MPT Test Development: A Walk-Through from First Draft to Administration,* 84(2) The Bar Examiner 29–34 (June 2015).
3. Some jurisdictions may grade on an absolute basis—awarding points according to the grading rubric—regardless of how other examinees answer the question. As long as this grading method is employed consistently and spreads out scores, it is an acceptable method of grading.

4. Precisely how the written portion of the exam is scaled to the MBE is complex; there is not enough space in this article to fully discuss it, nor am I qualified to explain how it is done from a true measurement perspective. For a complete discussion of the steps we undertake in scaling written scores to the MBE, see Susan M. Case, Ph.D., *The Testing Column: Demystifying Scaling to the MBE: How'd You Do That?*, 74(2) The Bar Examiner 45–46 (May 2005); see also 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).
5. Minimizing a question's contribution to an overall written score is not necessarily problematic; if a question does not perform as intended, so that no examinees (or all examinees) get it right, then it is appropriate that that particular question's impact is minimal. Note that graders should not artificially spread scores just for the sake of spreading them. Distinctions made between papers should be material.
6. For a full discussion of calibration, see my Testing Column in the March 2015 *Bar Examiner*: Judith A. Gundersen, *The Testing Column: Essay Grading Fundamentals*, 84(1) The Bar Examiner 54–56 (March 2015).

Judith A. Gundersen is the Director of Test Operations for the National Conference of Bar Examiners. Among her responsibilities, she is program director for the Multistate Essay Examination and the Multistate Performance Test. Gundersen received her J.D. from the University of Wisconsin Law School.