Practices and Procedures to Improve Grading Reliability on Essay Examinations

A Guide to the Care and Feeding of Graders

By Kellie R. Early

Beginning with the July 2003 bar examination, the Missouri Board of Law Examiners instituted a mandatory grading conference during which our twenty-two graders gather for two and a half days to calibrate and grade exams. Initially, there were predictions that we would never be able to schedule a weekend when all of the busy attorneys who serve as graders would be available and willing to give up time that might otherwise be devoted to family or work. There also were a few complaints about the prospect of having to grade away from the comforts of home or office. There were concerns by some graders that the grading conference was instituted because the board was dissatisfied with the quality or reliability of the grading. In fact, even before we began our conference tradition, the reliability of the essay portion of our exam was strong. The board believed, however, that a grading conference could improve reliability by ensuring that the grading was done uniformly in a structured setting over a limited period of time.[1] A grading conference would also allow the graders to work more closely with each other and would allow the Board and staff members to more effectively provide training or retraining of graders if necessary.

This past August, we completed our third grading conference. We have encountered surprisingly few scheduling conflicts, perhaps because the dates are confirmed nine to twelve months in advance. The graders now overwhelmingly prefer grading in the controlled environment of the conference where they can more easily devote their full attention to grading, free from other distractions and demands on their time. The Board's confidence in the quality of the grading is even higher. For those interested in instituting a grading conference, this article describes one jurisdiction's approach and may be helpful in deciding whether to move forward. For others, the ideas offered can be applied outside the context of a grading conference.

In Missouri, we administer ten essay questions[2] and one MPT item to more than 800 applicants in July and more than 400 in February. We use two graders for each essay question and the MPT, dividing the answers equally between the two graders, so that each grader grades half of the answers for that question. Our pool of graders is stable, with infrequent changes in the pairing of grading partners or the subject matter to which each pair of graders is assigned.

When the staff divide the answers for distribution to the graders, we make sure that each of the two graders assigned to a question receives the same mix of first-time and repeat applicants. Thus, half of the answers from the first-timers are assigned to one grader and half to the other; similarly, each grader receives half of the answers from the repeaters. This division of answers is done without regard to the identity of the individual applicants. None of the graders, including board members, have access to the database of applicants. The graders are not given any information about the number or percentage of first-time takers versus repeaters in the applicant pool, nor is there any way they can determine that information from the answer booklets they are given to grade. Providing comparable sets of answers to each of the two graders, however, allows more meaningful comparisons to be made between the "Group A" and "Group B" graders' means, standard deviations, and score distributions for each question, as will be discussed in greater detail below.

The grading conference is held no earlier than the second weekend following the bar examination. This allows us to distribute to the graders in advance of the conference the revised model analyses for the MEE questions and the revised point sheets for the MPT item prepared by the National Conference of Bar Examiners. Before they come to the conference the graders are instructed to thoroughly familiarize themselves with the model analysis for their assigned essay and to conduct any necessary legal research on issues unique to Missouri not addressed in the model analysis.[3] We also send out to the graders copies of ten randomly selected answers for their review in connection with the model analysis so they can familiarize themselves with how the applicants are handling the question.

At the beginning of the grading conference, each pair of graders engages in a calibration session before doing any actual grading. During the calibration session, each pair of grading partners reads and grades together a set of 30 sample answers, discussing the standards by which points should be awarded. They reach a common understanding of how to score the answers so that by the end of the process they are awarding independently the same, or nearly the same, score to any answer they read.

In preparation for the conference, we select and copy the 30 answers to each question to be used for calibration purposes. We do not copy the front cover of the answer books so the graders are not able to identify the exam numbers of the answers in the calibration set. Instead, a generic cover page is attached that is labeled "Calibration Example #___," and the numbers "1" through "30" are marked on the covers to identify the calibration answers. The graders record the agreed-upon score assigned to each calibration answer on a calibration grade sheet and also on the front cover of the answer booklet copy. When they finish calibrating, they turn in the calibration grade sheet but keep the set of calibration answers to refer to during actual grading as "benchmark" answers that reflect the standards agreed upon during the calibration session.[4]

The answer booklets that were copied for the calibration set are mixed back in with the other answer booklets and graded again during actual grading. Because the graders cannot identify the calibration answers by exam number, we are able to compare the calibration score with the score assigned during actual grading. This allows us to discover and correct "renegade" graders who stray significantly from the standards agreed to during calibration.[5]

The set of calibration answers is selected to reflect the characteristics of the applicant pool in terms of the percentage of first-time takers and repeat takers. For example, if 75 percent of the applicants are first-time takers, then 75 percent of the calibration answers are selected randomly from first-time takers. For the reasons described in the paragraph above, we also draw the calibration answers equally from the answers assigned to each grader so that fifteen are drawn from the Group A answers and fifteen from the Group B. The graders typically spend several hours calibrating before they begin grading.

To maintain calibration over the course of grading, we designate a few randomly selected answers as "grading exchange answers" to be graded by both graders. The grading exchange answers are designated as such by an asterisk marked beside the exam number on the grading sheet. When a grader reaches a grading exchange answer, she reads it and decides on a score but waits to record the score until her grading partner also grades the answer. After both graders have independently graded the answer, they compare scores. If their scores are not the same, they discuss the answer to arrive at an agreed-upon score and determine if further calibration is needed.

In general, when a grader is reading a series of essay answers, the order in which answers are read might influence how leniently or harshly the answers are scored. An average answer might be scored more harshly if it is read after a very good answer or more leniently if it is read after a very bad answer. This is referred to by psychometricians as the "context effect."[6] In addition, there is some evidence that answers read at the beginning of the grading process might be scored more leniently than those read at the end of the process.[7] To reduce the effect of such variables to the extent practicable, we stagger the starting point of each grader to prevent the answers from being read in the same order by every grader.[8]

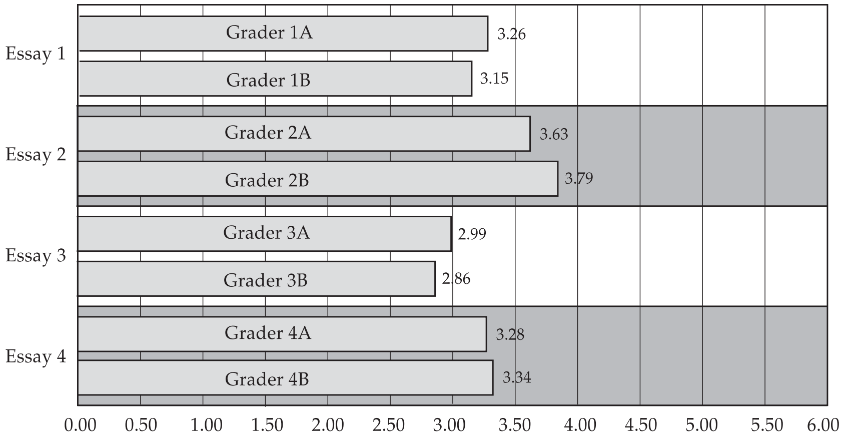
The calibrating and grading is done in a large meeting room at the hotel. Each pair of grading partners is seated together at a table to allow them to converse as needed while grading independently. Meals are served in a room separate from the grading room. We open the grading room by 7 a.m. and keep it open as late as any graders wish to keep working. We do not enforce a strict grading schedule and allow graders to take breaks when needed. Graders are expected, however, to work diligently to complete as much grading as possible before the end of the conference. Those who do not finish by the end of the conference are required to complete their grading within the following week while standards are still fresh in their minds, and are reminded to refer to their calibration set of benchmark answers. Our graders generally are able to complete all initial grading of a February examination and about 80 percent of a July examination before the end of the conference.

After the initial grading is complete, the essay examination scores are scaled to the MBE and preliminary total scores are calculated. Review grading then begins.[9] All ten of the essay answers and the MPT answers of those applicants whose preliminary total scores are within a particular range of the pass/fail line are subjected to a "blind" review grading by the partner of the original grader for each question. The review is blind in the sense that the review grader does not know the original score assigned to the answer being reviewed nor the applicant's preliminary total score. In fact, the graders are not even told what range of preliminary total scores qualify for review grading. The original and the review scores for each essay are averaged except in those instances when the two scores are considered too far apart.[10] In that case, the two graders are asked to discuss the answer and arrive at an agreed-upon score.

After the examination results are released, grading statistics are prepared to monitor and provide feedback and training to the graders. We distribute grading statistics from the previous exam at the start of each grading conference. The conference setting provides the opportunity to thoroughly explain any grading issues raised by a review of the grader statistics or to reinforce grading issues previously discussed. For each grader, the mean, standard deviation, and frequency distribution of scores are calculated. All graders' means and standard deviations are set out in one table so that comparisons can be made at a glance. See Figure 1, above. Similarly, the frequency distributions of all graders are set out in one table. In addition, the frequency distributions are charted graphically for each grader and for each pair of grading partners. See Figure 2, below.

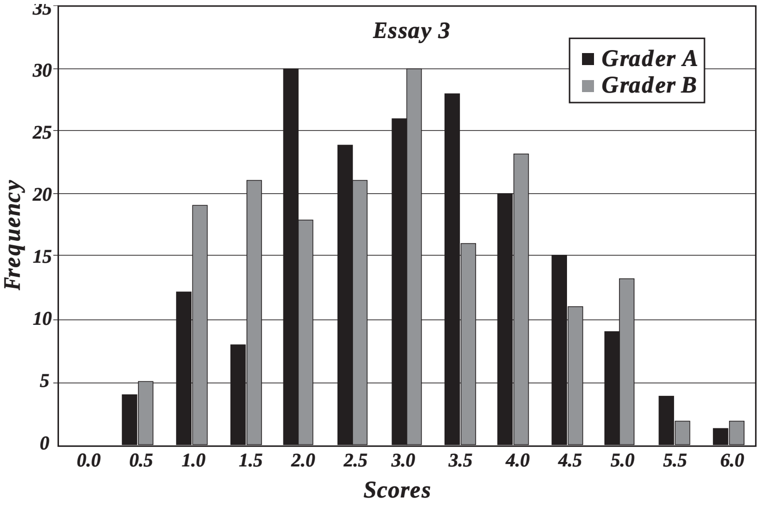
Figure 1. Means and Standard Deviations for Graders

	Grader 1A Essay 1	Grader 1B Essay 1	Grader 2A Essay 2	Grader 2B Essay 2	Grader 3A Essay 3	Grader 3B Essay 3	Grader 4A Essay 4	Grader 4B Essay 4
Mean	3.26	3.15	3.63	3.79	2.99	2.86	3.28	3.34
Std. Dev.	1.20	1.25	1.21	1.26	1.19	1.31	1.21	1.10



Comparative review of the grader statistics allows us to notice if a grader is assigning much harsher or more lenient scores than her grading partner. This phenomenon is generally noted by a significant difference in the means of the grader and her grading partner. We also can spot graders who tend to use only part of the grading scale by monitoring each grader's standard deviation and distribution of scores. Good grading generally should result in a reasonable spread of scores across the grading scale. If a grader assigns the same score to a large proportion of answers, it indicates a possible lack of differentiation by the grader in assessing the relative quality of answers. The standard deviation indicates how much the scores spread around the mean. A small standard deviation signals that the grader is not spreading scores across the scale.[11] The distribution of scores

Figure 2. Frequency Distribution for One Essay



further indicates which particular scores, if any, are being over- or underused by the grader.

A reasonable spread of scores generally is portrayed graphically as a bell-shaped curve. We have found that providing statistical feedback to the graders is best accomplished by the use of graphs. A visual depiction of a grader's score distribution is more meaningful to the grader than merely being told the frequency with which she assigned each score. Further, when the score distributions of the grader and her partner are plotted on the same graph, the graders each get a visual feel for their consistency with each other.

A final word about the care and feeding of graders. All of our graders are highly qualified attorneys who grade bar exams out of a desire to serve the profession. They strive to be as objective and fair as possible and to give every single answer due consideration, whether it be the first or the last in a stack of hundreds of answers. Their legal knowledge and good intentions are put to best use when they receive as much feedback and support as possible.[12]

Endnotes

1. ***See*** Susan M. Case, Ph.D., The Testing Column: Quick Responses to Issues in Testing, 72 Bar Examiner 4:30, 31 (Nov. 2003).
2. In addition to six MEE questions, Missouri administers four essays drafted by our board members. The board members draft a model answer for each question they write. Generally, a board member is assigned primary drafting responsibility for a particular subject matter, but all members participate in review and editing of the questions and model answers. The board member with primary drafting responsibility for a question also grades that question, along with a grading partner.
3. Because the graders are assigned to subject matters in which they practice and are familiar with Missouri law, they generally recognize from reading the question if there is an issue on which Missouri law differs from that set out in the model analysis. We also take a laptop to the grading conference and can access Westlaw if further legal research is necessary.
4. ***See*** Jean C. Gaskill, A Model for Grading Bar Examination Essay Questions, 65 Bar Examiner 1:30, 35 (Feb. 1996).
5. ***See generally*** Julia C. Lenel, Ph.D., The Essay Examination Part III: Grading the Essay Examination, 59 Bar Examiner 16, 23 (Aug. 1990) (recommending that to measure the consistency of a grader's scores over time, the grader must read the same answers twice, preferably without recognizing the answers and remembering what score the grader initially assigned).
6. ***See*** Stephen P. Klein, Ph.D., Options for Assigning Essay Scores, 65 Bar Examiner 1:24, 26 (Feb. 1996); Lenel, ***supra*** note 3, at 20.
7. ***See*** Lenel, ***supra*** note 3, at 18 and 20.
8. To make recording, entry and proofing of grades as efficient and accurate as possible, the grade sheet on which the scores are marked lists the answers in exam number order. Likewise, to make it easier to account for all the answer books, the books are sorted and provided to the graders in exam number order. Grader 1A, however, might be told to start grading with exam number 33, while grader 2A is told to start with exam number 42, and so on.
9. Graders are reminded again to first review the calibration set of benchmark answers before they start review grading.
10. Until July 2004, we used a 12-point grading scale that ranged from zero to six but included half-point decimals, i.e., 0, .5, 1, 1.5, 2, 2.5, 3, 3.5, 4, 4.5, 5, 5.5, and 6. Beginning in July 2004, we switched to a 12-point whole number scale to eliminate decimals from our raw scores. When we used the decimal scale, we required graders to discuss answers on which the original and review scores differed by more than one point. On the whole number scale, we require them to discuss any differences of more than two points. The sample statistics provided are from an exam graded using the decimal scale.
11. ***See*** Lenel, ***supra*** note 3, at 23.
12. Plus, it helps to keep fresh, hot coffee and snacks available at all times!

Kellie R. Early is the Executive Director of the Missouri Board of Law Examiners. She currently serves on the Editorial Advisory Committee of the National Conference of Bar Examiners and is vice chair of the Council of Bar Admission Administrators' Survey Committee. She received her J.D. in 1985 from the University of Missouri-Columbia and is admitted to the bar in Missouri and California. Before beginning her work in bar admissions, she was the Director of Continuing Legal Education and an adjunct professor in legal research and writing at the University of Missouri-Columbia; she also worked in private practice in St. Louis and Los Angeles.