The Testing Column

Quality Control for Developing and Grading Written Bar Exam Components

by Susan M. Case, Ph.D.

In the May 2010 issue of the *Bar Examiner*,[1] I discussed the concept of best practice and outlined the following 10 best practices in testing for admission to the bar. These cover three main categories:

A. Best practices for exam development

1. Each exam component must have a stated purpose.
2. Each exam component must be developed using professional standards of test development and with strictest adherence to security.
3. Grading criteria must reflect the exam purpose, and the grading processes must adhere to professional standards.
4. Each exam question must be reviewed and pretested to ensure the quality of the test development and grading criteria.

B. Best practices for test administration

5. Test administration practices must ensure that each examinee is authorized to take the test.
6. Test administration practices must ensure that examinees do not have access to testing aids.
7. Test administration practices must ensure that examinees cannot copy from one another.
8. Test administration practices must ensure that examinees cannot take test material or information out of the testing room.

C. Best practices for grading individual essays and combining scores

9. Grading practices must follow professional standards, with emphasis on grader training, calibration, grading consistency, and monitoring.
10. Scores must be equated, scaled, and weighted to ensure appropriate score meaning.

Because of continuing questions from jurisdictions, I have devoted this column to addressing issues related to exam development, as well as issues related to grading individual essays and combining scores.[2]

Exam Development

Each Exam Component Must Have a Purpose Statement

Every high-stakes examination such as the bar examination must have a written purpose statement that explicitly states what skill and knowledge set each component is designed to assess.[3] For example, the purpose of the MEE is to test the examinee's ability to (1) identify legal issues raised by a hypothetical factual situation; (2) separate material which is relevant from that which is not; (3) present a reasoned analysis of the relevant issues in a clear, concise, and well-organized composition; and (4) demonstrate an understanding of the fundamental legal principles relevant to the probable solution of the issues raised by the factual situation.[4]

Jurisdictions that develop their own essay or performance tests should have similar statements to define the content and scoring of their specific exam components. As NCBE has done, the jurisdiction should make sure that this purpose statement is widely distributed, preferably by displaying it prominently on the jurisdiction's website.

Every Question Must Conform to the Exam Purpose Statement

Every question must conform to the purpose of the exam. For example, the bar exam is developed to assess the extent to which each examinee has the knowledge and skills that are required of newly licensed lawyers. Each question should be framed within a context of a case that might be seen by a new lawyer and that a new lawyer would be expected to handle.

The Quality of Each Question and Its Supplementary Materials Must Be Ensured

Jurisdictions that develop their own exam questions need to ensure the quality of each question. Each question's author must be familiar with the purpose of the exam and with the content specifications of the topic area. The author of the question should prepare not only the question but also other materials such as the grading guidelines, analysis (scoring rubric), and model answer. The author should be sufficiently knowledgeable about the topic to be sure that the supplementary material is accurate and that the question addresses the most important aspect of the topic from the perspective of what a newly licensed lawyer would deal with.

Each question and the grading materials should be reviewed by independent content experts. To ensure that the difficulty is appropriate for the examinee group, each question should be pretested, using recent admittees who write responses to the questions under secure, timed conditions. Obviously, in selecting the expert reviewers and pretesters, care should be taken to ensure that they will not disclose the contents of the exam.

Jurisdictions developing their own questions should ensure that each question assesses competence in key areas of the law—areas that are either seen frequently by newly licensed lawyers or that are so critically important that every new lawyer should be competent in the area being assessed. NCBE has recently completed a job analysis of what the newly licensed lawyer does and what knowledge, skills, and abilities newly licensed lawyers believe they need to carry out their work. This job analysis, available on the NCBE website,[5] provides valuable information that could be useful in deciding which topics should be covered in a jurisdiction's test component. If a jurisdiction believes that the national job analysis is not entirely relevant to practice in its jurisdiction, the jurisdiction should perform its own job analysis sampling newly licensed lawyers practicing locally.

The questions, analyses, and grading guidelines should be of publishable quality. They should be made available to the public by posting them on the jurisdiction's website after test administration.

Grading

The Quality of the Grading Materials Must Be Ensured

The grading materials are prepared to help the grader score the written answers accurately and also to help each grader maintain consistency in the grading of the essays. As noted earlier, the grading materials should be reviewed carefully by content experts to ensure that they accurately reflect the law. In addition, someone other than the question's author should review the grading materials to make sure they are consistent with the purpose of the exam. For example, if the purpose of the exam includes the assessment of writing quality, writing quality should be part of the scoring rubric.

Grader Calibration Must Be Achieved

If the responses to each question are graded by a single grader, the grader should grade approximately 30 papers (selected at random), place them in grading buckets, and then review each bucket to ensure that the papers within the bucket are consistent in quality. These 30 papers, referred to as calibration papers, should then be spread throughout the remaining papers to be graded, with their grades masked, and checks should be made to ensure that the calibration papers' grades remain consistent across the grading period.

If the responses to each question are graded by more than one grader, 30 or so papers should be randomly selected to be graded as part of the calibration. Each grader should read several papers and assign grades to them. Then the graders should discuss the grades that have been awarded and resolve any differences. A second group of papers should then be graded and discussed. This process should be continued until the graders are in sync.

Almost all jurisdictions scale their essay grades to the MBE. Under this condition, the graders should rank-order the papers instead of deciding which papers are passing and which are failing. The top grade does not necessarily indicate an excellent paper; it just indicates a paper that is better than the other papers. The calibration process should continue until all the grading points have been used—that is, if a jurisdiction has a 1 to 6 grading scale, some of the graded papers should be assigned to each of the available points.[6]

For jurisdictions not scaling essay grades to the MBE, there are additional hurdles that must be met. First, the graders must have a consistent definition of what constitutes a grade of passing, as well as a consistent definition of what each score on the grading scale represents. Discussions should be held to ensure a common understanding of the characteristics of the just-passing examinee and how these characteristics would manifest themselves on the papers being graded. Obviously, it is very difficult to maintain consistent standards from one administration to the next, but this is required in order to ensure fairness.

The Importance of Ensuring Score Reliability Must Be Recognized

Jurisdictions must ensure that any score that is used for decision making is sufficiently reliable for high-stakes testing. High reliability is essential to ensure that the pass/fail status of examinees would not flip-flop from one administration to the next or if different questions were asked, if different graders were grading the papers, or if the examinees were testing with a more or less able group of examinees.

Jurisdictions that scale the essays to the MBE scores for their jurisdiction, that weight the MBE at least 50%, and that make the pass/fail decision on the total score are assured of a sufficiently high reliability and high decision consistency. Jurisdictions that make a separate pass/fail decision based on the written exam need to undergo separate psychometric analyses to ensure that they are meeting standards for high-stakes examinations.

Conclusion

The best practices described in this article are required for high-stakes standardized tests used for licensure of professionals. Using non-standardized components, such as locally developed essay questions and performance tests, requires quality control procedures unlike those that are followed for the standardized multiple-choice component.

Notes

1. Susan M. Case, Ph.D., *The Testing Column: Top 10 List of Best Practices in Testing for Admission to the Bar*, 79(2) The Bar Examiner 36–39 (May 2010), *available at* http://www.ncbex.org/assets/media_files/Bar-Examiner/articles/2010/790210_TestingColumn.pdf.
2. The best practices described in this article are not typically followed for classroom tests but are followed for high-stakes standardized tests used for licensure of professionals.
3. *See* American Educational Research Association, American Psychological Association, and National Council on Measurement in Education, Standards for Educational and Psychological Testing (American Educational Research Association 1999), Standards 3.2, 3.6, 14.14 (regarding purpose); Standard 3.22 (regarding scoring); Standard 4.21 (regarding standard setting).
4. National Conference of Bar Examiners, *The Multistate Essay Examination*, http://www.ncbex.org/multistate-tests/mee/.
5. National Conference of Bar Examiners, *NCBE Job Analysis: A Study of the Newly Licensed Lawyer*, http://www.ncbex.org/publications/ncbe-job-analysis/.
6. For additional details regarding the calibration process, *see* Susan M. Case, Ph.D., *The Testing Column: Procedure for Grading Essays and Performance Tests*, 79(4) The Bar Examiner 36–38 (November 2010), *available at* http://www.ncbex.org/assets/media_files/Bar-Examiner/articles/2010/790410_TestingColumn.pdf.

Susan M. Case, Ph.D., is the Director of Testing for the National Conference of Bar Examiners.