Originally appearing in THE BAR EXAMINER***, Vol. 72, No. 1, February 2003. Reprinted with permission of the National Conference of Bar Examiners. All rights reserved.***

THE TESTING COLUMN

BAR EXAMINING AND RELIABILITY

by Susan M. Case, Ph.D.

On January 2, several of us from the Conference had the privilege of speaking at the annual meeting of the Association of American Law Schools. Erica Moeser, Mike Kane, and I joined law professors Shelly Kurtz, from the University of Iowa, and Charles Daye, from the University of North Carolina, to present a three-hour program entitled "How to Construct an Exam." The event was organized by Dale Whitman, the outgoing president of AALS and a law professor from the University of Missouri-Columbia. We had been asked to discuss basic principles of assessment (reliability and validity), the advantages and disadvantages of various assessment methods, how to write and review multiple-choice questions, how to write and grade essays, and how to combine scores from different essays and other exam components to create a total test score.

We were surprised to be faced with a packed room of more than 100 attendees, and we were even more surprised that the room was just as packed three hours later. From the first couple of minutes, there were dozens of questions from both senior and junior faculty members from a variety of law schools. It quickly became clear that the audience was full of people with many testing questions, and it also became clear that the issues of concern to law school faculty are very similar to issues of concern to those of us involved with bar admissions. Most of the questions related to grading essay examinations, which the group clearly believed was a painful part of their jobs. As Shelly Kurtz put it, "all of us teach for nothing; our salaries are for grading."

We began with a discussion of score reliability, with an emphasis on the importance of reliability whenever scores are used to make high-stakes decisions (such as determining final course grades and deciding who has passed the bar examination). Some aspects of scoring are highly relevant for bar examinations but are of less relevance for most course examinations. For example, comparability across graders is less relevant in classroom situations where a single professor typically grades all the papers, but highly relevant in bar examinations where one grader may grade the answers for an essay question written by some of the examinees while a second grader grades answers written by other examinees for the same question.

Other aspects of scoring that affect reliability, such as how stable examinee scores would be across other hypothetical forms of the test, are as important for course examinations as they are for bar examinations. In both high-stakes end-of-course final examinations and bar examinations, it is important to be assured that the grade reflects proficiency in the topic areas covered by the examination, beyond the specific questions actually asked on the examination. If we were to administer two completely different tests covering the same topics, under ideal circumstances, an examinee's scores on the two tests would be identical. This would indicate perfect reliability, which of course is not possible in real life. But we should strive to make scores as reliable as is possible and feasible.

This issue raised many practical questions that can be discussed within the context of the following example. Consider a hypothetical situation where a final exam in a torts course includes three essays on three separate topics: negligence, intentional torts, and strict liability. You would imagine that examinees who score well on one topic would score fairly well on the other topics (and vice versa), but you would not expect the rank-ordering of examinees from topic to topic to be identical. And if a second question were to be asked on negligence, for example, you would expect the rank-ordering of examinees on the two negligence questions to also be similar, but not identical.

On a global level, one would hope that total scores on a second final exam covering the same three topics would yield a very similar rank-ordering of students. The test scores are supposed to represent proficiency in torts, and more specifically in negligence, intentional torts, and strict liability. Each question is a sample of the questions that might have been asked to test proficiency in the topic area, and the test as a whole is a sample of the examinations that might have been developed to reflect proficiency in the course. It should be irrelevant to the students which exam they took; each exam should include a fair representation of the important topics covered in the course.

In the case of final examinations, and in the case of bar examinations, pass/fail and other types of decisions are made based on something analogous to these total exam scores. We need to be fair in making these decisions, but we also want to make the right decision: a crapshoot is fair but not good enough; we can do better. Some scoring strategies work better than others, even given the limitations of resources such as testing time and grader time.

Here at NCBE, we have been working to develop some specific advice that will tend to enhance the accuracy of scores for examinations used in bar admissions. One strategy relates to the scoring scheme that is used. All else being equal, more score gradations work better than fewer score gradations. The key is to make sure that the scale reflects the level of judgments the grader can make. For example, with a 200-item multiple-choice test, the scoring machine can grade on a 200-point scale, one point for each question. Because of limits in differentiating so finely, many professors and jurisdictions grade essays using scales with a range of 4 to 12 points (using either a numeric scale or an alphabetic one such as A, B, C, D, with perhaps pluses and minuses). A six-point grading scale tends to work better than a four-point grading scale. Something much broader, like a 20-point grading scale, would work better than a six-point scale, but only if the grader could make reasonable, consistent, meaningful decisions along that scale; it would not work as well if the grader could not distinguish at so fine a level (that is, if the grader had to make arbitrary decisions about whether a paper should get a 3 or a 4, then those scale points should be collapsed).

TABLE 1: Distribution of Grades on Essay 1, Essay 2, and Essay 3.

This table shows the percentage of examinees who were assigned each grade (0 to 6) for the essays. For example, on Essay 1, 3% of the examinees received a grade of 1 and 16% received a grade of 2; no one received a grade of 0, 5 or 6.

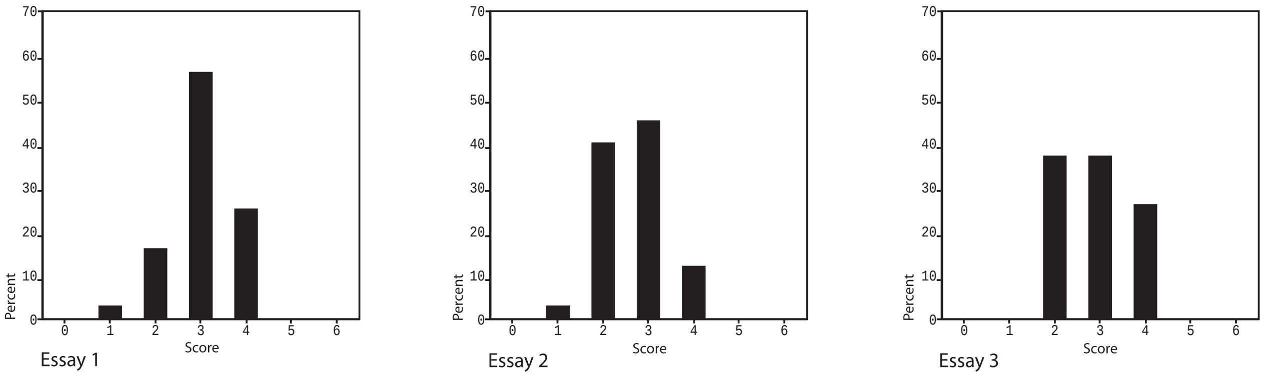
	Scores Assigned						
	0	1	2	3	4	5	6
Essay 1	-	3	16	56	25	-	-
Essay 2	-	3	40	45	12	-	-
Essay 3	-	-	37	37	26	-	-

A problem arises when the actual scale used is markedly narrower than the purported scale. As an example, consider the scores from one professor's class. He used three essays on his final exam; each was scored on a 0 to 6 scale. However, when his scores were analyzed, it became apparent that most of his students received scores of 2, 3, or 4—what began as a 7-point scale had been reduced to a 3-point scale. Table 1 and Figure 1 show these data. No one scored as low as 0; no one scored as high as 5 or 6. If the example had been slightly more extreme, all the students would have received the same score.

The professor could argue that the proficiency of the students was so similar that no one excelled on the topic and no one performed badly. If this is true, then the question had limited utility in determining course grades. Spreading scores out and using as many score points as possible is even more important for bar examinations where the only purpose is to determine proficiency (in contrast to course examinations where other purposes such as providing feedback that directs student learning could be envisioned). It is also even more important in situations where one grader grades some of the

FIGURE 1: Distribution of Grades on Essay 1, Essay 2, and Essay 3.

This figure shows graphically the same data that are included in Table 1.



examination answers and another grader grades others; it is unfair to have variation in the use of the scale from one grader to the next, particularly on the same question.

So, the underlying questions are these: How can we assess the proficiency of examinees, using a reasonable amount of resources, especially testing time and grading time? Given limited resources, what is the best combination of essay questions, multiple-choice questions, and performance test questions? These issues have been researched for decades in all areas of testing, but we are working on better and more specific answers as they apply to bar examinations. Based on what we know so far, it seems wise to encourage graders to use the entire range of scores on the scale if they can. If they are unable to use the entire range, perhaps one strategy is to ask them if the 3s (for example) could be divided into groups of 3+, 3, and 3-. This would serve to generate a broader range of scores without forcing graders to use the extremes.

The reliability of the total score also depends on the questions asked. As noted above, we want the scores to be reliable in the sense that two tests covering the same content areas would yield similar scores. In this vein, test developers should ask themselves the following question:

> Does the question represent the topic (or subtopic) area, so that if a second question were developed to assess proficiency in the same area, the content would be close enough to the first question that experts on one question are likely to be experts on the second, and those deficient on the first will be deficient on the second?

Test developers can also help to facilitate the use of the full scale score in subsequent scoring and should ask themselves the following question:

> Is the question broad enough to generate a range in the quality of the responses (more than just right/wrong), but also clear enough so that examinees understand what is asked of them?

A question that does not allow for variation in the quality of answers (beyond simply right and wrong) does not provide a good basis for reliably assigning a range of scores.

Grading is difficult (it's what faculty and bar examiners get the big bucks for!). But scoring that differentiates the various levels of performance is essential in any high-stakes context and deserves serious effort.

SUSAN M. CASE, PH.D., is the Director of Testing for the National Conference of Bar Examiners.