The Testing Column

Relationships Among Bar Examination Component Scores: Do They Measure Anything Different?

by Susan M. Case, Ph.D.

The written portion of the bar examination in many jurisdictions includes multiple components to evaluate candidates' competencies. The three main options for written components are (1) locally developed essay questions, (2) the Multistate Essay Examination (the MEE), and (3) the Multistate Performance Test (the MPT). Jurisdictions' choices regarding the use of multiple components and the number of "cases," or questions, to include for each component are based on multiple factors, one of which is often a desire to measure various aspects of competency. One would expect that the various written methods assess some shared aspects of competency, and that each method also assesses some unique aspect of competency.

We have analyzed the data from nine jurisdictions that used two MPTs in the February 2008 bar examination to determine the extent to which the statistical relationships among the scores reflect the relationships that would be expected on a rational basis. Five of these jurisdictions used only local essay questions in addition to the MPTs, and four also used the MEE.

Why Use Written Components on a Bar Examination?

We expect that the written components measure, at least in part, something different from what is measured by the MBE; that's why all jurisdictions use them. And the data support this expectation. If two components measured exactly the same thing, the correlation between the two would be 1.00 (perfectly related). A correlation this strong is rarely if ever seen in real life—think about the strong correlation between height and weight. The correlation is strong (taller people tend to be heavier), but there are short fat people and tall skinny people who make the correlation less than perfect, even if everyone's weight and height were measured with total precision. In our data set, the correlation with the MBE is 0.55 for local essay questions, 0.58 for the MEE, and 0.38 for the MPT. This shows a moderate correlation for both the locally developed essay questions and the MEE, but a weaker correlation for the MPT, indicating that the MPT is measuring different skills than the MBE, and that the MPT skills are less like those measured by the MBE than are the skills measured by the MEE and local essay questions.

An additional consideration in these relationships is that we know we are not measuring these skills with perfect precision. The MBE is long enough (i.e., contains sufficient questions) so that MBE scores are very precise (i.e., reliable). Scores on the MPT, with only two cases, and scores on the essays, with only a few questions, are less precise. However, there are statistical techniques that may be applied to estimate what the relationships would be if the scores were perfectly reliable. Our data show that, when corrected for the lack of perfect reliability, the correlation with the MBE is 0.76 for local essay questions, 0.78 for the MEE, and 0.58 for the MPT. These results show a moderately strong relationship between the MBE and the local essay questions, as well as between the MBE and the MEE, and a weaker relationship between the MBE and the MPT, as expected.

Table 1
Correlations Among Written Components and the MBE: Observed and Adjusted for Reliability

	Local Essays	MEE	MPT	MBE
Local Essays		0.44	0.43	0.55
MEE	0.83		0.38	0.58
MPT	0.81	0.76		0.38
MBE	0.76	0.78	0.58	

Table 1 shows the pattern of correlations, with those above the shaded diagonal showing the observed relationships and those below the shaded diagonal showing the relationships adjusted for reliability. The data show that, even when corrected for the lack of precision in the scores, the essays and the MPTs are measuring something somewhat different from what the MBE is measuring.

Maintaining Adequate Reliability Numbers When Using Written Components

These same statistical techniques may be used to estimate the level of reliability that would be expected for a given number of questions (cases) or the level of reliability that would be expected for a given amount of testing time. Table 2 shows these values.

Table 2
Estimated Reliability of Written Bar Examination Components

Components	Reliability with six cases	Reliability for three hours of testing time
Local Essays	0.63	0.63 (6 cases)
MEE	0.67	0.67 (6 cases)
MPT	0.69	0.46 (2 cases)

The second column shows the reliability that would be expected if each component included six cases. In general, an overall MPT component score based on six cases would have a higher estimated reliability (0.69) than an overall MEE component score (0.67) or an overall local essay score (0.63) based on the same number of cases. However, because each MPT case takes 90 minutes, it would be difficult to include more than two in a bar examination, and as the chart also shows, the reliability for three hours of testing time is much lower for the MPT (0.46) than for the local essays (0.63) or the MEE (0.67). To put these reliability values into context, the reliability of the MBE scaled score is approximately 0.81 for three hours of testing time (half the length of the normal exam).

The industry standard is that any score used for decision making should have a reliability of at least 0.90. This means that if a jurisdiction scales its written scores to the MBE, combines scores, and makes a pass/fail decision based on this total score, the total score should have a reliability of at least 0.90. Because the reliability of the MBE itself is 0.90, combining written scores with the MBE and making a decision on that total score is a good procedure because the combined score will achieve the appropriate level of reliability. On the other hand, if a jurisdiction requires applicants to pass the written portion separately from their performance on the MBE, the reliability of that written portion should be 0.90 by itself, an almost impossible feat, given the values in Table 3.

Table 3
Testing Time and Number of Cases Needed to Achieve Reliability Over 0.90

Components	Testing time (in hours)	Number of cases
Local Essays	16.0	32
MEE	13.5	27
MPT	33.0	22

Table 3 shows that testing time requirements to achieve reliability values over 0.90 would be 16 hours for local essays, 13.5 hours for the MEE, and 33 hours for the MPT. The number of cases would be 32 local essays, 27 MEE questions, and 22 MPT items. If decisions were made based on one of these components alone, without combining it with the MBE, these are the values that would be required to meet reliability standards.

CONCLUSIONS

The pattern in the data showing moderate but less-than-perfect correlations among written scores is consistent with expectations that written components are designed to assess some shared aspects of competency but also to assess unique aspects of that same domain. The fact that these correlations are less than 1.00 is consistent with the expectation that the written components measure aspects of competency that are different from the aspects of competency measured by the MBE.

Based on these data, MPT scores and MEE scores tend to be more reliable than local essay scores based on the same number of cases. Both local essays and the MEE produce more reliable scores per unit of testing time than does the MPT, but the MPT measures skills that are more distinct from MBE skills than are the skills measured by local essays or the MEE.

Both the MPT and the MEE (and local essays) add something unique (as they were designed to do) to the holistic assessment of competence. As long as these component scores are scaled to the MBE, the scores are added together (with the MBE weighted at least 50%), and the pass/fail decision is based on the total score, the score is sufficiently precise and reflects the applicant's proficiency with respect to a wide range of knowledge and skills that are relevant to the tasks that are important for the newly licensed lawyer.

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