PR-01-01

RELATIONSHIP OF SCORE RELIABILITY TO ESSAY TEST LENGTH, QUESTION LENGTH, AND SECTION WEIGHTING

Stephen P. Klein, Ph.D. and Roger Bolus, Ph.D.

GANSK & ASSOCIATES

October 1, 2001

RELATIONSHIP OF SCORE RELIABILITY TO ESSAY TEST LENGTH, QUESTION LENGTH, AND SECTION WEIGHTING

California's General Bar Examination (GBX) has two parts. One part is the Multistate Bar Exam (MBE). This one-day test consists of 200 multiple-choice items. The other part of the GBX consists of a two-day written exam that contains six 60-minute essay questions and two 180-minute performance test (PT) problems. The written section carries about twice as much weight as the MBE in determining a candidate's total score and thereby pass/fail status. Licensed attorneys who have been admitted to practice for at least four years in another state are eligible for licensure in California if they just take and pass the written section. This alternate route to licensure is called the "Attorneys" exam.

We investigated the effects on score reliability of reducing the written portion of the California exam to six 30-minute questions plus two 90-minute Multistate Performance Test (MPT) problems. This change would keep the number of questions the same, but eliminate one full day of testing. The major findings of our research (which included analyzing one to three recent July bar exams in 20 other jurisdictions) were as follows:

- When question length is held constant, longer tests (i.e., those with more questions) generally have higher reliabilities and correlations with the MBE than shorter ones.

- When the number of essay questions is held constant, longer questions (as indicated by the amount of testing time that is allocated to them) generally produce more reliable scores than shorter questions.

- Essay and total score reliability varies across states and within states across administrations even when the number of questions asked and the time allocated to answer them is held constant. Several factors may account for this variation.

- Total score reliability on the California exam is about .85 under the current policy of weighting the written section twice as much as the MBE.

- Reducing the written section to one day would *not* reduce its total score reliability *provided* the written and MBE sections are weighted equally.

- Giving the written section twice as much weight as the MBE with a shortened test would reduce total score reliability, but only slightly (from about .85 to .80).

- Changing the weights given to MBE and written sections from 65/35 to 50/50 will increase the passing rate for males (by about 1.7 percentage points) but decrease it for females (by about 1.2 points). Changing the weights assigned to these sections will not systematically increase or decrease the passing rate within a racial/ethnic group.

- Reducing the written section to one day would lower the score reliability of California's Attorneys exam. This is likely to be an important concern.

Overview

The next portion of this report provides background information regarding score reliability and how it interacts with a jurisdiction's bar passage rate. Understanding the principles discussed in this section is necessary to fully appreciate what follows. We then describe the procedures in our study and the results we obtained as well as some important caveats and some possible concomitant effects of shortening the written test.

Score Reliability

The bar exam assesses a candidate's knowledge, skills, and abilities in various areas of the law. It is obviously not feasible to ask all the questions that could be posed to candidates in these areas on every administration. Hence, any given version (or "form") of the exam (such as July 2000) contains only a *sample* of these questions. Score reliability indicates the degree to which the scores on this sample are likely to be *representative* of the scores these candidates would earn if they answered other sets of questions from the same pool of questions that could have been asked.

Score reliability also can be thought of as a measure the confidence that can be placed in the likelihood that the scores the candidates earned on the test that was given would correspond to their scores on an equivalent test that could have been given (such as the one slated for the next administration of the exam). Thus, score reliability refers to the *consistency* or *stability* of a candidate's score across different samples of questions that could be asked. Put another way, score reliability indicates how confident we can be in *generalizing* from the results with the particular sample of questions that were asked to the larger domain of all the relevant questions that could have been asked.

Several factors affect score reliability. For example, most candidates have only partial knowledge of a subject. Consequently, some candidates will be able to answer certain questions but not others while the reverse is true for other candidates. Similarly, some candidates may be "morning" people while others may function better later in the day. On essay tests, there also is the problem that different readers (or the same reader on different occasions) may assign different scores to the same answer.

All of these and many other extraneous factors can affect a candidate's score. Hence, a candidate's score has two components: (1) the systematic and consistent part (which is called "*true*" variance) and (2) the noise or chance part (which is called "*error*" variance). The true part is estimated from the degree to which a candidate's performance is consistent across different questions, such as the extent to which their scores on the even numbered MBE items corresponds to their scores on the odd numbered items. The *total* variance (i.e., the true + error variance) is a function of the standard deviation of the total scores (i.e., how much they spread out around the mean total score).

The *reliability coefficient* equals the true variance's share of the total variance; i.e., score reliability equals true variance divided by total variance. Reliability coefficients can therefore range from 0.00 to 1.00. A 0 coefficient indicates that we would have no confidence in the score being indicative of how a candidate would perform relative to others if we had asked a different sample of comparable questions. A 1.00 coefficient indicates that the relative standings of the candidates on one form of the test would be identical to their standings on another form of it.

Score Reliability and Passing Rates

Almost all jurisdictions base their overall pass/fail decisions on a weighted combination of a candidate's essay and MBE scores. The reliability of this total score is a function of the reliabilities of the MBE and essay scores plus the correlation between these scores. The higher this correlation and the higher the reliabilities of the MBE and essay sections, the higher the reliability of the total scores. Thus, when state boards consider changing the structure of the essay portions of their bar exams, they will need to consider the effects of these changes on the correlation between MBE and essay scores as well as on the reliability of the essay section itself. Unfortunately, a change that lowers the score reliability of the essay section will also tend to lower its correlation with the MBE.

To illustrate, suppose a state's essay test consisted of six 60-minute questions. Our data suggest this will likely lead to an essay score reliability of about .71 and about a .62 correlation between essay and MBE scores. The MBE's score reliability is typically about .88. If the MBE and essay were weighted equally in computing a candidate's total score, then the reliability of these total scores would be about .87. It would slip slightly to about .83 if the essay carried twice as much weight as the MBE.

Now suppose this state shortened its essay test to six 30-minute questions; i.e., from a six-hour test to a three-hour test. Our data indicate this will lower the reliability of the essay section from .71 to about .54. It also will reduce the essay's correlation with the MBE to about .55. Taken together, these changes will lower the reliability of the total scores to about .81 if the sections are weighted equally and to .73 if the essay carries twice as much weight as the MBE. However, this loss in total score reliability could be offset somewhat by using the three hours that were taken away from the essay to add two Multistate Performance Test (MPT) problems to the exam.

Whether a change in the reliability of the total scores is policy relevant or not depends heavily on a state's passing rate. Specifically, the closer that rate is to 50%, the more important it is to have high total score reliability. Conversely, if almost everyone passes (or everyone fails), score reliability becomes much less critical because it is unlikely that using a different form of the test would result in changing the pass/fail status of many candidates.

Table 1 shows the relationship between score reliability, passing rate, and the likelihood of changing a candidate's pass/fail status as a result of simply administering a different form of the test. For example, if the reliability of the total scores is .90 and 80% of the applicants pass, then about 10% of them would have a different pass/fail status if they took another form of the test. In other words, some would go from passing to failing while others would move in the opposite direction. In contrast, if reliability slips to .80 and only 60% pass, then 20% of the candidates would have their pass/fail status affected by just taking a different form of the test. Hence, whether a small or even a fairly large reduction in score reliability matters depends heavily on a jurisdiction's passing rate. The closer that rate is to 50% for all takers, the more important score reliability becomes.

The overall passing rates on the February and July 2000 California GBX were 40% and 55%, respectively. The corresponding rates on the Attorneys exam were 54% and 55%. Because these rates are so close to 50%, score reliability is an especially important consideration for California.

Table 1

Percentage of Candidates Who Would Have a Different Pass/Fail Status if They Took a Different Form of the Test as a Function of Passing Rate and Total Score Reliability

Percent	Score Reliability									
Passing	.00	.10	.20	.30	.40	.50	.60	.70	.80	.90
90	19	17	16	16	15	14	12	11	9	6
80	32	30	28	26	24	22	20	17	14	10
70	42	39	37	34	32	29	25	22	18	13
60	48	45	42	38	35	32	28	24	20	14
50	50	47	44	40	37	33	29	25	21	15
40	48	45	42	38	35	32	28	24	20	14
30	42	39	37	34	32	29	25	22	18	13
20	32	30	28	26	24	22	20	17	14	10
10	19	17	16	16	15	14	12	11	9	6

Research Methods

We examined recent July bar exam results from 21 jurisdictions. These jurisdictions ranged from having only 65 to nearly 8,000 July candidates. Together, these jurisdictions account for roughly 70% of all the bar exam takers annually. We restricted our study to states whose exams yielded scores on six or more independently graded essay questions. Almost all the states from which we requested scores were able to provide these data.

Data were analyzed for one to three exams for each of the participating jurisdictions. All told, we examined data for 42 exams that together contained over 400 questions. The results with state developed essay questions did not appear to differ systematically from those obtained with Multistate Essay Exam (MEE) questions. Thus, we did not distinguish between them in this report.

Where possible, we examined the reliability of each jurisdiction's full test. We also created as many completely separate six-question and nine-question essay subtests as the data would allow. For example, we formed two subtests per exam administration for the states that asked 12 essay questions, but only one six-question subtest (usually the first six questions) for states that asked 8 questions. When we were able to create more than one subtest for a state's exam, we computed statistics separately for each subtest and then used the average of their values for that administration. Thus, the unit of analysis is an administration of an exam.

Effects of Test and Question Length

Table 2 shows the average of the score reliabilities for various combinations of test and question length. It is evident from this table that if the number of questions asked is held constant, then longer questions (i.e., those to which more testing time is allocated) provide more reliable scores than shorter questions. Similarly, if the time to answer is held constant, then the tests with more questions generally provide more reliable scores. Note also that asking 12 essay questions in six hours yields a score reliability that is roughly the same as asking 6 essay questions in six hours (.67 and .71, respectively). There did not appear to be a systematic relationship between the number of candidates in a jurisdiction and the size of its reliability coefficients.

Table 2
Mean Score Reliability for Different Combinations
of Test and Question Length

Minutes per Question	Number of Essay Questions			
	6	9	12	>12
25-30	.53	.64	.67	.76
36-45	.63	.73		
50-60	.71			

Table 3 shows the correlation of MBE scores with tests of varying lengths. These data have the same general pattern as those in Table 2; i.e., increasing the number of questions and/or their length increases the correlation of their scores with MBE scores.

Table 3
Mean Correlation with the MBE for Different
Combinations of Test and Question Length

Minutes per Question	Number of Essay Questions			
	6	9	12	>12
25-30	.54	.60	.60	.69
36-45	.63	.67		
50-60	.63			

Other States' Experiences With the Proposed Structure

California's General Bar exam consists of the MBE, six 60-minute essay questions, and two 180-minute performance test problem. The score on a PT answer is nominally given twice as much weight as the score on an essay answer. The mean score reliability over the three July exams for its written (essay plus PT) section is .74. The mean correlation of this section with the MBE is .64. Total score reliability is .88 if the sections are weighted equally and .85 if the written section has twice as much weight as the MBE.

One of the other jurisdictions we studied ("State A") gave two 90-minute MPT problems and a dozen 30-minute essay questions. Analyses of its July 2000 data indicated that a written test composed of six 30-minute essay questions and two MPTs would have a written score reliability of .63. Such a test would have a .60 correlation with the MBE. These values would lead to a total score reliability of .85 if the sections are weighted equally and a .78 if the written section is weighted twice as much as the MBE.

Another jurisdiction ("State B") also gave two 90-minute MPTs and several essay questions. Averaging results over the last three July bar exams in this jurisdiction indicated that the combination of a three-hour essay test plus two 90-minute MPTs would have a written test score reliability of .65 and a correlation of .62 with the MBE. The reliability of this state's total scores would be .85 if the written and MBE sections were weighted equally and .80 if the written section had twice as much weight as the MBE.

A third jurisdiction ("State C") had exactly the same written test structure as California, but used half the testing time by giving six 30-minute essays and two 90-minute MPTs. By averaging data over the last two July exams (and giving each PT answer twice the nominal weight assigned to an essay answer) we found that State C's written section had a score reliability of .67 and a .61 correlation with the MBE. The reliability of the total scores was .86 when the sections were weighted equally and .81 when the written section was given twice as much weight as the MBE.

Table 4 contrasts the results in States A, B, and C with those in California. These data suggest that cutting the time California currently allocates to its written section in half would not affect total score reliability *provided* California gave the written and MBE sections equal weight. However, if California cut testing time in half and continued to give the written section twice as much weight as the MBE, then there would be slight (.05) drop in total score reliability (i.e., from about .85 to about .80).

Table 4
Estimated Total Score Reliability as a Function of Question Length and the Weight Given to the Written Section Relative to the MBE

State	Written Test Length	Written and MBE have equal weights	Written has twice the weight as the MBE
A	6 Hours	.85	.78
B	6 Hours	.85	.80
C	6 Hours	.86	.81
California	12 Hours	.88	.85*

* California's current exam

Concomitant Effects

It is conceivable that reducing the GBX's total score reliability from its present .85 to about .80 could have some very slight downstream consequences on other outcomes. Specifically, for reasons that go beyond the scope of this report, it will tend to slightly reduce the standard deviation of the total scores. This could, in turn, lead to a very small reduction in the February passing rate and a very slight increase in the July rate (because the passing score is above the mean in February but below the mean in July). Reducing total score reliability from .85 to .80 would very slightly reduce or have no effect on the differences in passing rates among gender and racial/ethnic groups.

As noted above, reducing the written section from a two-day test to a one-day test is likely to reduce the reliability of the written scores from about .74 to about .65. This may raise concerns about the stability of the pass/fail decisions on the Attorneys exam. Specifically, about 55% of the Attorney candidates pass this test. If score reliability is .74, then according to the data in Table 1, slightly more than one-out-of-five of the Attorney exam candidates would have their pass/fail status affected by simply taking another form of the test. If score reliability dropped to the expected .65 for a one-day written test, then about one-out-of-four candidates would experience a change in status just by taking another form of the test. In short, chance would play an even larger role in determining an Attorney candidate's pass/fail status.

There would be similar concerns in states that allow candidate's to "bank" their essay scores and/or require them to surpass a certain minimum score on each section of the exam. In short, written score reliability will be an issue when these scores are used by themselves to make pass/fail decisions (i.e., rather than in conjunction with MBE scores).

Effect of Section Weighting

Changing the weight the MBE carries relative to the written section could have policy implications. To examine this issue, we recomputed every candidate's total scale score (prior to reread) on the last six July exams using a 50/50 written to MBE scale score weighting policy. We then contrasted the percent passing with these recomputed scores with the percent passing based on total scores (prior to reread) with the existing 65/35 weighting.

This analysis found that about 4.3% of the candidates who would pass under the current 65/35 weighting would fail if the weights were 50/50. Similarly, about 4.2% of those who would pass under a 50/50 weighting policy would fail under the current 65/35 policy. Thus, overall, about 8.5% of the candidates would have their pass/fail status affected by which set of weights was used, with about half of them going from a pass to a fail while the other half going in the opposite direction.

Which set of weights is used had almost no effect on the passing rates within racial/ethnic groups. For example, over the last six July exams, the passing rates prior to reread for Hispanics with the 65/35 rule and the 50/50 rule were 41.5% and 41.6%, respectively. The corresponding rates for African Americans were 27.9% and 27.8%. In short, weighting policy did not have much if any impact on a racial/ethnic group's passing rate.

The choice of weights did have a small effect on the passing rates within gender groups. Changing the weights from 65/35 to 50/50 would increase the passing rate for males by about 1.7 percentage points while decreasing it for females by about 1.2 points. These changes would occur because males generally earn higher scale scores on the MBE than on the written section while the reverse is true for females.

Caveats

Several caveats should be kept in mind when considering the results presented above. First, the score reliability of an essay test depends on several factors, including the characteristics of those tested and those who grade their answers. Consequently, the reliability of essay scores varies across jurisdictions even when they ask the same questions, use the same time limits, and endorse the same grading standards as other jurisdictions. Thus, the experience in one jurisdiction may not coincide exactly with the experience in another.

Second, score reliability may be affected by the number of different subjects for which candidates had to prepare, the degree to which these subjects overlapped those on the MBE, what candidates were told and learned in their bar review courses, the diversity of the subjects covered by the questions that actually appeared on the test, the quality of the questions that were asked, and the procedures used in grading the answers to them.

Third, the number of questions asked may influence drafting, editing, and scoring practices. For example, a board that asks six questions may devote more time to writing and grading each of them than a board that asks 12 questions. Similarly, a state with relatively few candidates can devote more time to grading answers or even grading each answer twice. As a consequence, increasing test length from 6 to 12 questions may not increase score reliability as much as would otherwise be expected by standard statistical formulas.

Fourth, the analyses presented in this report are based solely on July exams. We did this in the interests of consistency because some of the states we wanted to include only give the test in July or had so few February takers as to render the results with their exams questionable. Our experience indicates that February exams tend to have slightly lower total score reliabilities than July exams.

Fifth, score reliability may vary across jurisdictions even when they ask the same questions and the same number of questions in the same amount of time and have ostensibly similar populations of candidates. For example, we found that score reliabilities ranged from .40 to .67 across the 22 exams for which we had data on one or more six-question essay tests with exactly 30 minutes per question. Thus, there was quite a lot of variation around the median value of .54 for an essay test of this length.

Sixth, our study focused on estimating how shortening the written section would affect score reliability. We did not examine whether shorter essay and PT questions would assess the same (or more or less appropriate) skills and abilities as longer ones.

Finally, some states, such as New York, consistently obtain a higher degree of score reliability per hour of essay testing time than other jurisdictions. And, score reliability may vary considerably within a state across administrations of its exam. It is not clear why such variation occurs (e.g., the standard deviation of MBE scores is generally quite comparable across states).

Conclusions

The data in this report show that essay score reliability tends to increase as the number of questions asked increases. Increasing the amount of testing time per question (as a proxy for question length) also increases essay score reliability. Reducing test or question length will therefore tend to have the opposite effect. However, even a large reduction in question length is unlikely to have any practical effect on the reliability of total exam scores provided the MBE and written sections are weighted equally in the computation of those total scores. This outcome stems from the substantial correlation between MBE and written scores (and the MBE's high score reliability).

Nevertheless, there are some situations where a reduction in the written section's score reliability could be a serious concern. Specifically, this is likely to be a significant issue for any exam (including California's Attorneys exam) that bases pass/fail decisions on written test scores alone. This is an especially important concern when the passing rate is close to 50% (as it is on California's Attorneys exam).

Finally, how much weight is given to the written section relative to the MBE in computing a candidate's total scale score is likely to have some effect on the passing rates within gender groups. This will happen because males tend to earn higher MBE than written scores while the reverse is true for females.

STATISTICAL NOTES

Table 1 was created by conducting a Monte Carlo simulation with 10,000 replications. It assumes that both tests have equal means and variances, and that their scores are distributed normally. Deviations from these assumptions would tend to lower agreement.

The table below shows the number of tests that were used to calculate the mean values that appear in Tables 2 and 3. A given exam was often used to calculate the mean for more than one cell in these tables. For example, analyses of the data from states that asked a dozen 30-minute questions were used to estimate the reliability for tests with 6, 9, and 12 questions.

Number of Tests Used to Calculate Score Reliability and Correlation with MBE

Minutes per Question	Number of Essay Questions			
	6	9	12	>12
25-30	27	21	9	4
36-45	5	5*		
50-60	8			

* Three of the exams in this group contained 8 questions. The other two contained 10 questions apiece. However, the values in this cell were very similar across the five exams.

All the reliability coefficients in this report are unstandardized coefficient alphas. Calculations of total score reliability assumed that the MBE's reliability was .88 which is typical of that reported by ACT for July administrations.

The formula for estimating the reliability of the total scores of an equally weighted linear composite of MBE and essay (or written) scores is shown below where A = the reliability of the essay test, B = reliability of the MBE, and C = the correlation between them.

$$\text{Rtt when weighted equally} = 1 - [(2 - A - B)/(2 + 2*C)]$$

The formula for estimating the reliability of the total scores of a differentially weighted linear composite of written and MBE scores is shown below where A, B, and C are the same as above, D = standard deviation of the essay, E = the square of this standard deviation, F = standard deviation of the MBE, and G = the square of this standard deviation.

$$\text{Rtt when weights differ} = 1 - \{[(E + G) - (A*E) - (B*G)]/[(E + G) + 2(C*D*F)]\}$$

Sections are weighted by the relative sizes of their standard deviations. For example, for the analyses that gave the written section twice as much weight as the MBE, we set the written section's standard deviation at twice the size as the MBE's standard deviation.