

THE ESTIMATED EFFECT ON EXAMINATION QUALITY AND PASSING RATES OF DIFFERENT WAYS OF MODIFYING CALIFORNIA'S BAR EXAMINATION

Stephen P. Klein, Ph.D. & Roger Bolus, Ph.D.
December 12, 2011

Overview

California's General Bar Examination (GBX) is an 18-hour (three-day) test. It consists of the MBE (which is a 6-hour 200-item multiple choice test), a set of six 1-hour essay questions, and two 3-hour Performance Test (PT) questions. This report estimates the likely effects on this exam's quality and passing rate if it was shortened to a two-day test and gave the MBE and written sections equal weight.

Samples

The population for our analyses consisted of all the applicants who took the GBX one or more times between 2001 and 2010. There were 43,832 February and 81,346 applicants in this 20-exam sample for a grand total of over 125,000 applicants. We also analyzed the essay and PT scores of the subset of 20 to 25 applicants who had their answers to each question graded independently by at least ten of the readers assigned to a question; i.e., all the applicants in this sample had their answers to each question graded ten times.¹

Purposes & Definitions

Our analyses examined how score reliability was affected by: (a) using two or more independent readers per answer, (b) giving the MBE 50% of the weight (instead of the current 35%) in determining an applicant's total exam score, and (c) shortening the written portion of the exam from a two-day 12 hour test to a one day six or seven-hour test. We also examined the percentage of applicants in different racial/ethnic and gender groups whose pass/fail status would be affected by the number and sample of essay and PT questions they answered.

The term "score reliability" in this study refers to the likelihood that applicants would receive the same score (as distinct from pass/fail decision) regardless of the particular set of California bar exam essay and PT questions they were asked or the set of readers who graded their answers. For example, an essay test with high score reliability is one where the applicants who earn relatively high scores on one question also tend to earn relatively high scores on the test's other questions. All other things being equal, the higher the score reliability the greater the confidence that can be placed in the results.

¹ Results with this sample and those who went to reread must be treated very cautiously because they are not random or representative samples of the population of all takers.

“Score reliability” (coefficient alpha) is reported on a 0.00 to 1.00 scale with 1.00 being best. Values less than 1.00 may be due to: (a) some applicants being more proficient in the skills and knowledge needed for some questions while other applicants have a different pattern of expertise, (b) differences among readers in the score they would assign to an answer, and (c) other factors, such as how much scores spread out from the mean. In California, adjusting for the typically small difference in means and standard deviations among readers on a question usually has little or no effect on the written test’s score reliability.

“Decision consistency” (which is an especially important characteristic of a licensing test) refers to the stability of pass/fail decisions, such as across different types of tests or versions of a test. Thus, it is a useful index for examining the comparability of different test designs. Decision consistency is highest when: (a) score reliability is high and (b) the passing rate is well above or well below 50%.

Effect of Number of Readings Per Answer on Score Reliability

Score reliability increases as the number of readers per answer increases, but the benefit of additional readers tapers off rapidly. For example, the first row of Table 1 shows that the reliability coefficient for a 6-question essay test in July is 0.06 points higher with two readers than it is with one reader, but adding a third reader results in only a 0.01 improvement over having two readers. In short, the marginal benefit of adding readers disappears quickly (although it seems to be greater for two 3-hour PT questions than for a set of six 1-hour essay questions).

Table 1
Increase in Score Reliability Over a Single Reader as a Function of the Number of Additional Readers per Answer, Type of Question Asked, and Test Month

Number of additional readers	6 Essay Questions		2 PT Tasks	
	February	July	February	July
1	.06	.06	.06	.07
2	.07	.07	.11	.17
3	.07	.08	.13	.17
4	.07	.08	.15	.17
5	.08	.08	.17	.17
6	.08	.08	.18	.17
7	.09	.09	.18	.17
8	.09	.09	.18	.17
9	.10	.09	.18	.17

Applicants have three hours per PT task and an average of one hour per essay question. Results in this table are based on the answers written by the 20 to 25 applicants who had all of their answers graded by all the readers assigned to each question.

The diminishing benefit to improving score reliability by having more than two readers per answer supports California's policy of having a second reading of all the answers written by all the candidates who came close to passing but failed after the initial reading of their answers. In addition, the limited benefit of additional readings suggests that the less than perfect reliability of the written test stems mainly from an interaction between applicants and questions rather than from differences among readers in the scores they would assign to an answer.

The remaining analyses in this report are based on just the first reading of an applicant's answers in the ten-year population of February and July takers. We did this because: (a) not all applicants had their answers read at least twice and (b) which applicants would have their answers read more than once was likely to vary across the different test designs we examined. Thus, the results with these models may underestimate the score reliability that is likely to occur if California continues rereading the answers of those who initially came close to passing.

Reliability of Essay and PT Scores

The mean reliability of a single reading of a set of six 1-hour essay questions in our population of February and July takers was 0.64 and 0.70, respectively. The higher scorer reliability in July than in February may be due at least in part to the greater variance in July scores. In both months, the reliability of the sum of the scores on a single reading of two 3-hour PTs was about 0.52 (based on the Spearman-Brown stepped-up mean correlation of 0.35 between two PT scores).

Procedures for Computing Total Scores

MBE raw scores (i.e., the number of items answered correctly) are converted to scale scores to adjust for possible differences in the difficulty of the questions asked. Essay and PT readers assign scores to answers in 5-point intervals on a 40 to 100-point scale. PT raw scores are then multiplied by 2.00 so that the maximum possible written raw score is 1,000 points. California (like most other states) converts its written raw scores to a score distribution that has the same mean and standard deviation as its MBE scale scores. This step adjusts the reader assigned total raw scores for possible variation in essay and PT question difficulty and grading standards over time. Total scale scores are computed using the formula below. Applicants with total scores of 1440 or higher pass, those in the 1390-1439 zone have all their answers reread, and all others fail.

$$\text{Total Scale Score} = (.35 \times \text{MBE Scale}) + (.65 \times \text{Written Scale})$$

Except as noted otherwise, the same procedures were used to compute written scale and total scale scores and to determine an applicant's pass/fail status for all the models discussed in the next section of this report.

Modeling Results

Tables 2 and 3 should be used together. Table 2 lists the key features of the models we examined and Table 3 shows their impact on total (MBE + Written) score reliability on February, July, and all exams combined.² For example, the only difference between models 1a and 1b is that as per current practice, model 1a weights the written and MBE scores 65% and 35%, respectively. In contrast, model 1b weights them equally. Table 3 shows that this single difference results in a relatively large improvement in score reliability (0.06 in February and 0.05 in July). The benefits of going to 50/50 weighting are consistent with the differences in reliability between models 5a and 5b.

Models 2a and 2b have the same structure, namely: three 1-hour essay questions and one 3-hour PT with the MBE and written sections weighted equally. The only difference between these models is that they use completely different essay and PT questions. The degree of agreement between these models therefore provides an unbiased estimate of their decision consistency and shows the reliability of an exam that is limited to the MBE and a 6-hour written test composed of three 1-hour essay questions and one full 3-hour PT question when the MBE and written portions are weighted equally.

Models 4a and 4b show the results for a two-day exam consisting of five essay questions and one PT. Although these analyses had to rely on data from 3-hour PTs, the results with them are likely to be very close to what would be obtained with 90-minute PTs; i.e., a 6½ hour written test. Models 4a and 4b have higher reliabilities than the current exam (model 1a) as a result of their giving the MBE and written sections equal weight.

Tables 4 and 5 show pass/fail decisions are consistent between various pairs of models. For example, Table 5 shows that in July, 93% of the applicants had the same pass/fail status under Model 2 (a two-day exam with a written component consisting of 3 essay questions and one PT) as they had with the current exam (i.e., a test with twice as many essay and PT questions) provided both exams weighted the written and MBE sections equally.

Table 6 shows that reducing test length does not affect overall passing rates or exacerbate the differences in rates that are typically found among racial/ethnic groups. Assigning equal weights eliminates the difference in passing rates between men and women. In short, California can implement a two day exam in a way that improves test quality, maintains existing pass/fail standards, and does so without making it more difficult for minority applicants to pass.

² Total score reliability calculations used MBE score reliabilities of .89 and .91 for the February and July exams, respectively as per the mean values in the MBE's technical reports. Written test reliabilities (coefficient alphas) were based on un-standardized essay raw scores on the first reading.

Table 2
Main Features of the Models Tested

Model	Essay	PT	Written/MBE Weights	Written Time	Model Description
1a	1-6	A & B	65/35	12 hrs	Current model & 65/35 weights
1b	1-6	A & B	50/50	12 hrs	Current but 50/50 weights
2a	1-3	A	50/50	6 hrs	Half of current written exam
2b	4-6	B	50/50	6 hrs	Half of current written exam
3a	1-4	A	50/50	7 hrs	4 1-hr Essays + one 3-hr PT
3b	3-6	B	50/50	7 hrs	4 1-hr Essays + one 3-hr PT
4a	1-5	A	50/50	8 hrs	5 1-hr Essays + one 3-hr PT
4b	2-6	B	50/50	8 hrs	5 1-hr Essays + one 3-hr PT
5a	1-6	None	65/35	6 hrs	6 1-hr Essays 65/35 weights
5b	1-6	None	50/50	6 hrs	6 1-hr Essays 50/50 weights
6	None	A&B	50/50	6 hrs	PT only

Table 3
Total Score Reliability (Coefficient Alpha, decimal points omitted)

Model Number	Test Month(s)			Model Description
	February	July	All	
1a	81	85	83	Current model & 65/35 weights
1b	87	90	88	Current but 50/50 weights
2a	80	85	82	Half of current written
2b	80	83	82	Half of current written
3a	82	87	84	4 1-hr Essays + one 3-hr PT
3b	82	86	84	4 1-hr Essays + one 3-hr PT
4a	84	88	86	5 1-hr Essays + one 3-hr PT
4b	84	88	86	5 1-hr Essays + one 3-hr PT
5a	81	85	83	6 1-hr Essays 65/35 weights
5b	86	89	88	6 1-hr Essays 50/50 weights
6	78	80	79	PT only

Table 4
Average Percentage of **FEBRUARY** Applicants with the
Same Pass/Fail Status Under Alternative Models

Model 1a	Model 1b	% Agree
MBE weighted 35% Essays 1-6 in 6 hours PT-A & B in 6 hours Reliability = .81	MBE weighted 50% Essays 1-6 in 6 hours PT-A & B in 6 hours Reliability = .87	95%
Shows unique effect of weighting the MBE 50%		

Model 1b	Mean of Models 2a & 2b	% Agree
MBE weighted 50% Essays 1-6 in 6 hours PT-A & B in 6 hours Reliability = .87	MBE weighted 50% 3 Essays in 3 hours 1 PT in 3 hours Reliability = .80	91%
Models 2a and 2b cut test length in half with MBE weighted 50%		

Model 2a	Model 2b	% Agree
MBE weighted 50% Essays 1-3 in 3 hours PT-A in 3 hours Reliability = .80	MBE weighted 50% Essays 4-6 in 3 hours PT-B in 3 hours Reliability = .80	82%
Same models but completely different written questions in 6 hrs		

Model 3a	Model 3b	% Agree
MBE weighted 50% Essays 1-4 in 4 hours PT-A in 3 hours Reliability = .82	MBE weighted 50% Essays 3-6 in 4 hours PT-B in 3 hours Reliability = .82	86%
Models share 2 of their 4 essay questions in 7 hrs		

Model 4a	Model 4b	% Agree
MBE weighted 50% Essays 1-5 in 5 hours PT-A in 3 hours Reliability = .84	MBE weighted 50% Essays 2-6 in 5 hours PT-B in 3 hours Reliability = .84	88%
Proxy for a 6½ hour written exam (4 essay questions in common)		

Table 5
Average Percentage of **JULY** Applicants with the
Same Pass/Fail Status Under Alternative Models

Model 1a	Model 1b	% Agree
MBE weighted 35% Essays 1-6 in 6 hours PT-A & B in 6 hours Reliability = .85	MBE weighted 50% Essays 1-6 in 6 hours PT-A & B in 6 hours Reliability = .90	96%
Unique effect of weighting the MBE 50%		

Model 1b	Mean of Models 2a & 2b	% Agree
MBE weighted 50% Essays 1-6 in 6 hours PT-A & B in 6 hours Reliability = .90	MBE weighted 50% 3 Essays in 3 hours 1 PT in 3 hours Reliability = .84	93%
Models 2a and 2b cut test length in half with MBE weighted 50%		

Model 2a	Model 2b	% Agree
MBE weighted 50% Essays 1-3 in 3 hours PT-A in 3 hours Reliability = .85	MBE weighted 50% Essays 4-6 in 3 hours PT-B in 3 hours Reliability = .83	85%
Same models but completely different written questions in 6 hrs		

Model 3a	Model 3b	% Agree
MBE weighted 50% Essays 1-4 in 4 hours PT-A in 3 hours Reliability = .87	MBE weighted 50% Essays 3-6 in 4 hours PT-B in 3 hours Reliability = .86	88%
Models share 2 of their 4 essay questions in 7 hrs		

Model 4a	Model 4b	% Agree
MBE weighted 50% Essays 1-5 in 5 hours PT-A in 3 hours Reliability = .88	MBE weighted 50% Essays 2-6 in 5 hours PT-B in 3 hours Reliability = .88	91%
Proxy for a 6½ hour written exam (4 essay questions in common)		

Table 6
Side-By-Side Model Comparison Chart

Total testing time	3 days		2 Days
Written components	6 Essays + 2 PTs		3-4 Essays + 1 PT
Model	Model 1a	Model 1b	Models 2 & 3
Written/MBE weight	65/35	50/50	50/50
Score Reliability			
All Takers	.83	.88	.82 - .84
February	.81	.87	.80 - .82
July	.85	.90	.83 - .87
February Passing Rates			
All February takers	37%	37%	37%
Females	39%	37%	37%
Males	35%	37%	37%
White	41%	42%	42%
Asian	35%	35%	35%
Hispanic	28%	28%	28%
African American	20%	20%	21%
July Passing Rates			
All July takers	53%	54%	54%
Females	55%	54%	54%
Males	52%	54%	54%
White	60%	61%	61%
Asian	49%	49%	49%
Hispanic	40%	40%	41%
African American	24%	25%	25%

Total testing time includes the MBE. Models 2a and 2b use three 1-hour essay questions. Models 3a and 3b use four 1-hour essay questions. Results are based on a single reading of answers on the 10 February and 10 July exams given between 2001 and 2010 (total N = 125,178 candidates). Model 4's February and July passing rates were consistent with Model 1's rates.