2018

An Evaluation of Alternative Versions of a Texas Bar Examination

A comparison of passing rates, decision consistency and reliability for simulated "2-Day" and Uniform Bar configurations using actual results from the 2013 through 2017 administrations.

Roger Bolus, Ph.D.
Research Solutions Group
6/29/18

SUMMARY

The Texas Supreme Court is currently considering modifications to the Texas Bar Examination (TBE), including adoption of the National Conference of Bar Examiners (NCBE) Uniform Bar Exam and, alternatively, a reduction in the duration of testing (a 2-Day TBE) that would involve shortening the Essay portion of the current exam.

To assist the Court in making its decision, the Texas Board of Law Examiners (TBLE) requested that a quantitative assessment of the potential impact of each of these alternatives be undertaken. Using actual examination results from 21,229 examinees that sat for the TBE between 2013 and 2017 (a total of ten administrations), simulated scores were calculated for three potential alternative configurations of the TBE and compared to actual results.

The three alternative configurations are: (1) A 2-Day version of the existing TBE consisting of only 6 essay questions covering Texas-relevant topics, with the remaining parts of the current exams and scoring protocols remaining the same (2) an alternative form of the 2-Day option with a shortened Essay section, but in which weighting of the MBE section is increased from 40% to 50% and weighting of the Essay decreased from 40% to 30% and (3) a modeled version of the UBE consisting of six "simulated" Multistate Essay questions covering similar content areas used by the NCBE, two Multistate Performance Tests, and the MBE. Similar to the UBE, a 50% weighting is given to the MBE and a 50% weighting is given to the written sections (30% for Essays and 20% for MPT). Results from these three simulated tests are compared to actual test results. Metrics used for the comparisons include overall bar passage results, consistency in pass-fail decisions, and overall test reliability.

Key findings from the simulations included the following:

- Both the 2-Day and UBE configurations of the exam were predicted to result in roughly the same *overall bar passage rates* as the current examination structure. The simulated alternative configurations were predicted to change the historical passing rate by no more than 1%. The results were consistent for February and July administrations.

- In terms of the *outcome for an individual examinee*, the pass/fail decisions from the alternative models matched the actual decision for 94% to 95% of all applicants. Where inconsistencies in the final status were observed, generally an equal proportion of actual passing and failing applicants were reclassified. The results tended to be consistent across the ten administrations, varying by no more than a couple of percent.

- As reported previously, the reliability of the TBE consistently meets expected standards for a high-stakes licensing test. The estimated reliabilities for the simulated UBE examination and the shortened TBE that reweighted the MBE to 50% of the total score, however, also exceeded the .85 reliability threshold for every administration, and on average were only .02 and .03 reliability points below the current examination.

With respect to this final finding, we found that overall test reliability was impacted in all three alternatives considered, due in part to the reduction in reliability of the shortened written sections. Shortening the existing TBE essay section from 12 to 6 questions reduced its reliability by an average of .17 points in February and .21 points in July. As a result, a shortened examination *that did not increase the weighting of the MBE section*, failed to achieve a reliability of .85

In conclusion, the report offers some cautionary notes on the use and interpretations of these simulations, but also points to examples where states that have modified their examinations have subsequently validated their early simulation studies.

I. INTRODUCTION

In its May 2018 report to the Supreme Court of Texas, the Task Force on the Texas Bar Examination recommended that the Texas Board of Law Examiners (TBLE) consider adoption of the National Conference of Bar Examiners (NCBE) Uniform Bar Examination (UBE) as its semi-annual bar examination. The committee also suggested that if the Court decided against transition to the UBE, then other changes should be made in the current structure of the TBE in order to expedite release of results to bar applicants. To this end, the TBLE had previously suggested that the current examination could possibly be shortened from three to two days of testing.

The consequences of a decision to modify the current structure and scoring procedures for any high stakes exam could be far reaching and should be based on both administrative and psychometric considerations. Fortunately, other states considering adoption of the UBE or other changes to their current bar exams have provided a roadmap for how to evaluate the impacts of such proposals. Statistical "simulations" or "models" have been used to predict the effects that proposed changes could have upon applicants' scores, passing rates and other variables. These analyses have used data from past examinations to recreate conditions that would represent (as closely as possible) the alternative structures and/or scoring protocols that are under consideration.[1] The studies have examined potential changes to overall bar passage rates, possible impacts on the performance of specific subpopulations, and the degree to which various statistical and psychometric properties of the test (e.g., test reliability) could be affected, among other topics.

To aid the Court in deciding whether to make changes to the TBE, the TBLE requested that similar modeling be performed on the historical data for the Texas examination. This report presents the results from this exercise, along with details of the analyses conducted. We examine each of the alternatives (i.e., the "2-Day" and "UBE") separately, comparing them to the actual results from prior administrations, and then provide a combined summary.

The analyses were conducted on test result data for a full five-year period covering the ten February and July administrations of the TBE from 2013 to 2017. This same data set was previously used for a technical review of the statistical and psychometric properties of the State's current examination during the identical period.[2] The reader is referred to this prior report for a statistical assessment of how well the TBE has functioned, and as a reference point for the consideration of the potential modifications to the examination discussed in the remainder of this report.

In simulating each exam option, we focused on three metrics:

Passage Rates. We first calculated overall bar passage rates by applying the current scaling procedures and passing scores to the simulated scores produced for each exam alternative. Additionally, predicted passage rates were calculated under revised weighting schemes proposed

[1] The author has conducted studies for five of those states.
[2] See Bolus (2018)

to compensate for potential impacts on test reliability arising from changes to the exam structure.

Pass/Fail Decision Consistency. It is inevitable that individual scores will change as well as the eventual decisions that are made based upon those scores. What is not known is whether the change in the actual vs. simulated scores will lead to different decisions; and if so, for how many examinees would be impacted and in which directions.

This report provides estimates on the degree of decision inconsistency likely to result from the use of proposed alternative versions of the TBE.

Score Reliability. The Bolus 2018 report documented the high statistical reliability of the current TBE. Several factors are associated with reliability. Generally speaking, reductions in test length will lead to reduced reliability, as can adding or deleting sections of an examination, Changes in section weighting can sometimes counteract those reductions.

To assess the impact of exam alternatives on reliability, we recalculated the test section and overall reliability of the alternative configurations and compared them to the original reliability results from the Bolus 2018 report.

II. SIMULATION OF A 2-DAY TEXAS BAR EXAMINATION

Prior to any consideration of the UBE, the Texas Supreme Court had requested that the TBLE consider strategies to provide examination results to applicants more quickly. In response, the TBLE's Executive Director proposed an option that would simultaneously shorten the actual test administration period from three days to two days, speed up the grading of the constructed response section of the exam, and reduce the burden on applicants sitting for the examination. The Director's recommendation centered around reducing the Essay portion of the examination from 12 to 6 questions, and from 6 to 3 hours in duration. The essay questions would continue to be developed by the TBLE, while the remaining sections of the exam would stay the same. All other testing and scoring protocols would also remain the same as the current TBE.[3]

Methodology

To model the outcomes of this exam alternative, the historical scores for the MBE, P&E and MPT sections were used in their entirety in the simulation analysis. For the Essay section, however, scores for only six of the 12 questions from each of the administrations were considered. The questions that were included were selected by the Executive Director to reflect Texas law subjects including Oil & Gas and Texas Consumer Law, that are not covered on the MEE or MBE.

For the three test sections that were used in their entirety, examinees' raw converted and scaled scores were included in the simulation. For each of the six selected Essay questions, the original raw and converted scores were used. Those converted scores were then totaled to form a revised total converted Essay score. For each administration, the revised essay scores were then re-scaled to the MBE using a new set of scaling equations based on the newly-calculated converted total scores.

To derive a Total Scale Score for each applicant, their scores on the four sections of the exam were first weighted using the current protocols: 40% for the MBE, 40% for the Revised Essay, 10% for the MPT and 10% for the P&E test. Recognizing that a shortened Essay section could have reduced reliability, a second Total Scale Score was also calculated in which the MBE weight was increased to 50% and the Essay section weight was decreased to 30%. The purpose in computing the two alternative scores was to compare the net impact on the outcomes of interest (i.e., passage rate, decision consistency and reliability.)

The Total Scale Score(s) was next evaluated relative to the same decision protocols as used in current practice. Examinees with scores 675 or above were considered to have passed outright. Those with simulated scores that fell below 669 failed outright. If applicants' scores fell within those two points, they were considered to have gone into re-grade. Since a second grading was not possible in this simulation exercise,[4] it was necessary to estimate what would have occurred had this group of examinees had a second reading. Appendix 1 discusses the procedure for making those estimations.

[3] See Bolus (2018) for a full description of those protocols

[4] In the simulation, some applicants that originally passed or failed would not have second read scores.

Through the simulation, we were able to compare the actual Pass/Fail outcomes for all applicants tested during the five-year study time frame with the outcome they would have achieved under a 2-Day exam alternative.

Findings

Passage Rates. Table 1 presents the actual TBE passage rates for the ten administrations under study along with the estimated rates for the simulated 2-day examinations under 1) the current weighting scheme (40% for MBE and Essay); and 2) a revised weighting scheme (50% for MBE, 30% Essay).

Table 1

Actual vs. Simulated 2-Day
Texas Bar Examination
Passage Rates
2013 through 2017

	Total Applicants	Actual		Simulated: Current Weighting		Simulated: Revised Weighting*	
		N	%	N	%	N	%
February							
2013	1,185	879	74%	900	76%	886	75%
2014	1,152	781	68%	776	67%	770	67%
2015	1,333	806	60%	821	62%	815	61%
2016	1,433	806	56%	800	56%	788	55%
2017	1,253	606	48%	613	49%	616	49%
5-Year	6,356	3,878	61%	3,910	62%	3,875	61%
July							
2013	3,023	2,474	82%	2,505	83%	2,509	83%
2014	2,929	2,091	71%	2,101	72%	2,083	71%
2015	2,987	1,985	66%	2,002	67%	1,974	66%
2016	2,975	2,098	71%	2,098	71%	2,092	70%
2017	2,959	2,124	72%	2,145	72%	2,149	73%
5-Year	14,873	10,772	72%	10,851	73%	10,807	73%

* MBE and Essay given 50% and 30% weighting, respectively

The results from Table 1 indicate *that the proposed 2-Day exam would have minimal impact on overall passage rates.* Additionally, providing more or less weight to the MBE and Essay sections also appeared to have little impact. Differences in the simulated and actual score passage rates across the five February administrations varied by 1% to a maximum of 2%, and no more than 1% on the July administrations. Across the 6,356 February examinees, only 31 more applicants would have been predicted to pass a 2-Day exam with similar weighting, while

only three fewer examinees would have passed under revised weighting. For the five July administrations, only 79 and 35 more examinees out of 14,873 would have been predicted to pass under the two alternative weighting schemes. A reduction in the number of essay questions, then, appears to have little impact on the overall bar passage rates. In most cases, the number of passing scores would increase slightly.

These results are not unexpected for a couple of reasons. First, there is relatively high correlation between the essay scores on the subset selected for the reduced examination and the full set (correlations range from .91 to .93 for the ten administrations), suggesting that a reduced set of essay question results would rank order applicants in virtually the same manner as the full set. Secondly, all scores continued to be scaled to the MBE, thereby helping to maintain similar shapes of the Total Score distributions under each alternative examined.

Decision Consistency. In addition to examining impacts on overall passage rates, we also investigated the extent to which the *same pass/fail decisions* would be given to individual examinees. Table 2 presents data comparing pass/fail decisions on the simulated 2-Day examination scenarios under both the current weighting and the alternative weighting schemes

Table 2

Actual vs. Simulated
2-Day Texas Bar Examination
Decision Consistency Rates
2013 through 2017

	Current Weighting						Revised Weighting*					
	Agreement: Actual & Simulated		Actual Pass/ Simulated Fail		Actual Fail/ Simulated Pass		Agreement: Actual & Simulated		Actual Pass/ Simulated Fail		Actual Fail/ Simulated Pass	
	N	%	N	%	N	%	N	%	N	%	N	%
Feb.												
2013	1,116	94%	24	2%	45	4%	1,130	95%	24	2%	31	3%
2014	1,081	93%	38	3%	33	3%	1,079	94%	42	4%	31	3%
2015	1,242	94%	38	3%	53	4%	1,246	94%	39	3%	48	4%
2016	1,353	94%	43	3%	37	3%	1,345	94%	53	4%	35	2%
2017	1,164	93%	41	3%	48	4%	1,179	94%	32	3%	42	3%
5-Yr	5,956	94%	184	3%	216	3%	5,979	94%	190	3%	187	3%
July												
2013	2,892	95%	50	2%	81	3%	2,928	97%	30	1%	65	2%
2014	2,777	95%	71	2%	81	3%	2,789	95%	74	3%	66	2%
2015	2,802	94%	84	3%	101	3%	2,828	95%	85	3%	74	2%
2016	2,833	95%	71	2%	71	2%	2,841	95%	70	2%	64	2%
2017	2,818	95%	60	2%	81	3%	2,838	96%	48	2%	73	2%
5-Yr	14,122	95%	336	2%	415	3%	14,224	95%	307	2%	342	2%

* MBE and Essay given 50% and 30% weighting, respectively

with actual historical results. The table presents the agreement rate (defined as the percentage of examinees who either fail under the current and simulated test, or pass under both), as well as

the disagreement rates (defined as the percentage of applicants who actually passed or failed the exam but had the reverse outcome in the simulation).

The results from Table 2 show *a high degree of decision consistency between the actual and simulated examinations*. On the February examinations, 93% to 94% of all examinees would have been expected to earn the same final pass or fail status, while the rates on the July examinations were 95% or more in all but one year when it was 94%.

Among all 21,229 examinees included in the simulations, 1151 examinees (5%) would have their pass/fail status reversed under the current weighting scheme, while 1,026 (5%) examinees would have been similarly impacted under the revised 50% weighting scheme. As shown in Table 2, we also observed symmetry in these results. That is to say, a roughly equivalent percentage of applicants who had failed were predicted to have passed and vice-versa. As an example, for all February examinees, 3% of applicants who actually passed were expected to fail under both simulated weighting schemes, while 3% who actually failed were predicted to have passed under the alternative testing models. Closer examination also revealed that the majority of the examinees whose status changed in the simulation had scores quite close to the pass/fail line. The mean differences in their actual and expected revised scores were relatively small, averaging only about 2.5 scale score points.

Examination Reliability. As the format of any exam is changed, it is reasonable to assume that the reliability of the examination could change as well, particularly if an exam is shortened. Under both weighting schemes for the 2-Day exam option, the length of the Essay section is cut in half, which in theory should result in reduced reliability and possibly a lowered correlational relationship with the other sections. The question at hand is whether the alternative that provides greater weight to the MBE (the exam section with the highest reliability) would compensate for this reduction.

To investigate these questions, we re-calculated the reliability of the revised essay section and the overall test under the current and revised weighting schemes. The results are summarized in Table 3.

On the essay portion of the examination, for the simulated February exams with six essays, we calculated an average decrease of .17 reliability points (from .79 to .62; a 22% reduction), while the simulated July exams showed a .21 reliability point reduction (from .79 to .58; 27% reduction). This is a significant decrease from the current levels, which according to Bolus 2018**,** contributed heavily to an overall examination reliability that met and/or exceeded commonly accepted standards for high stakes licensing tests. These overall test reliabilities during the five-year period examined averaged .89 and .90 for the February and July examinations, respectively, and actually met or exceeded .90 for most of the July administrations.

A recalculation of the total test reliability using a shortened Essay section and no change to the section weighting resulted *in decreases that failed to achieve the minimal target reliability of .85.* The estimated test reliabilities dropped to values ranging from .817 to .844 across the 10 examinations, with the February and July averages estimated to be .827 and .831, respectively.

However, *when adjustments were made in the weighting of the MBE and Essay sections (50% for the MBE and only 30% for the Essay), all total test reliabilities exceeded the .85 criteria*, averaging .865 for February administrations and .871 for July.

Table 3

Reliability Estimates of Original & Revised
Texas Bar Examination Essay and Total Scores
2013 through 2017

	Actual Essay	Rvs'd Essay	Actual Total Score	Simulated Current Weighting	Simulated Revised Weighting*
February					
2013	.80	.63	.892	.831	.866
2014	.82	.68	.898	.839	.869
2015	.76	.60	.881	.820	.860
2016	.77	.60	.892	.825	.866
2017	.78	.59	.894	.820	.865
5-Yr Ave.	.79	.62	.891	.827	.865
July					
2013	.79	.63	.868	.838	.866
2014	.79	.54	.900	.844	.875
2015	.78	.54	.896	.817	.867
2016	.79	.59	.903	.824	.871
2017	.82	.59	.912	.831	.876
5-Yr Ave.	.79	.58	.896	.831	.871

* MBE and Essay given 50% and 30% weighting, respectively

This finding is consistent with both psychometric theory and the experiences of several other jurisdictions that have conducted similar studies. For example, California considered reducing the number of essay questions and shortening the time given for its performance tests to accommodate a 2-day test. To accomplish this, while maintaining the traditionally high levels of test reliability, their modeling exercises determined that it was necessary to change the weighting scheme on the written and MBE sections from .65/.35 to .50/.50. Other states had similar findings, and this is one of the rationales for why the UBE employs the .50/.50 weighting.[5]

[5] This line of reasoning speaks only to the reliability criteria. Some argue that written or constructed-response tests such as the Essay or MPT, are a more valid measure of legal skills, and as such should be given more weight.

III. SIMULATION OF A TEXAS UNIFORM BAR EXAMINATION

The UBE, gradually being adopted by many jurisdictions across the U.S., has a standardized format that differs from the TBE in several ways. Implementation of the exam would require: 1) elimination of Texas's Procedure and Evidence test section[6]; 2) reduction of the Essay section from twelve questions constructed by the TBLE and scored on a 25-point scale to six Multistate Essay questions (MEE) developed by NCBE and scored on a 7-point scale (i.e., 0 to 6); 3) an increase in the Performance Test section from one Multistate Performance Test (MPT) to two; and 4) application of a new weighting scheme in which the MBE would carry a 50% weighting in the calculation of the Total Scale Score, (compared to the current 40%), the Essay section would carry a 30% weighting (currently 40%), and the Performance Test Section would carry a 20% weighting (currently 10%). Additionally, the UBE would be administered over a two-day period, as opposed to three.

In regards to scoring of the exam, raw scores on the different sections of the exam would continue to be scaled to the MBE. Rather than standardizing constructed scores beforehand, however, all written scores would be added together and scaled directly to the MBE. Scaled written scores would be weighted and added to the applicant's MBE score reported on the existing scale of measurement (i.e. 0 to 200). Total scores above 135 on the MBE scale (270 overall) would be considered passing.[7] This score is mathematically equivalent to Texas' current passing standard of 675.

NCBE has no official policy on re-grading, allowing states to implement their own procedures.[8]

Methodology

For our simulation of the UBE, only scores from the MBE, Essay and MPT sections of the TBE were used; the P&E was dropped. To simulate the UBE Essay section, the Executive Director reviewed the topic areas covered in the UBE Multistate Essay (MEE) portion of the UBE and identified six questions from each of the 10 examinations that most closely covered one or more of those topics. Those six questions made up the simulated MEE, and the remaining essay questions were excluded. Finally, since the UBE uses two 90-minute MPT tasks and Texas administers only one, we simply used the existing Texas MPT twice in the simulation.

The MPT sections of both the UBE and the current TBE are scored on a six-point scale. A six-point scale is also used for the MEE section of the UBE, but Texas currently scores essays on a 25-point scale. To simulate the reduced grading scale of the MEE, we recoded the actual scores to the 0 to 6-point scale by equally dividing the original score by four. Thus, a score of 1 to 4 on the original scale was assigned a value of "1," a score of 5 to 8 was assigned a score of "2," and

[6] See Bolus 2018 for a technical report on the Texas Bar Examination for a full description of the current test configuration, grading and scoring rules and statistical results from five years of recent testing.
[7] If scale scores are simply added, as opposed to weighted (the same mathematical effect) then a score of 270 would be considered passing.
[8] Early on in the development of the UBE, NCBE advocated a no re-grade policy on the basis of psychometric principles. This stance was subsequently modified and a subcommittee of participating states was created to evaluate whether a standard could be achieved.

so on. A "0" score was not changed. Since the UBE does not convert the individual Essay or MPT scores to a scale with a common mean and standard deviation, Texas' current process of score standardization was not applied for the simulation.

To derive a simulated UBE Total Scale Score for each applicant, a raw written score was first calculated for each examinee based upon the formula that assigns 30% weight to the Essay section and 20% weight to the MPT section. That written score was then linearly scaled to the same mean and standard deviation of the MBE distribution for each of the ten administrations. To arrive at a simulated Total Scale Score, the MBE and Written Total Scale Scores were both multiplied by 50% (i.e., given equal weighting) and added together, thus placing the Final Total Scale Score on the original 200-point MBE scale of measurement.

The Total Scale Score(s) were evaluated relative to the same pass/fail decision protocols as currently used. If an examinee's score was 135 or above (equivalent to 675 on Texas' current scale), they passed out right. If they scored 133 or below, they failed. If an applicant's score was 134 (roughly equivalent to the current 6-point Texas re-grade range) they were considered to have gone into re-grade. The same procedures described for the re-grading process in the simulation of the 2-Day exam were used to determine a final status for re-graded examinees.

Findings

Passage Rates. Table 4 presents the actual TBE passage rates for the ten administrations under study along with the estimated rates for the UBE simulated examinations (50% for MBE, 50% Written [30% Essay/20% MPT]).

Under the simulated UBE exam (which excluded the P&E), the average overall passage rates across the five February and five July administrations were 60% and 71%, respectively. These rates are almost identical to the actual average passage rates for those administrations (61% and 72%). For most administrations, the UBE simulated passing rates were 1% to 2% lower than actual. On two administrations., Feb. 2015 and Feb 2017, however, they were the same or higher. Across all five years, the simulation revealed that the UBE would have resulted in only 33 fewer passes in February and 233 fewer passes in July.

Table 4

Actual vs. Simulated UBE
Texas Bar Examination
Passage Rates
2013 through 2017

	Total Applicants	Actual		Simulated: UBE	
		N	%	N	%
February					
2013	1,185	879	74%	856	72%
2014	1,152	781	68%	767	67%
2015	1,333	806	60%	804	60%
2016	1,433	806	56%	793	55%
2017	1,253	606	48%	625	50%
5-Year	6,356	3,878	61%	3,845	60%
July					
2013	3,023	2,474	82%	2,419	80%
2014	2,929	2,091	71%	2,047	70%
2015	2,987	1,985	66%	1,949	65%
2016	2,975	2,098	71%	2,054	69%
2017	2,959	2,124	72%	2,072	70%
5-Year	14,873	10,772	72%	10,541	71%

Decision Consistency. Table 5 on the following page presents statistics on the consistency between the actual and simulated UBE examination pass/fail decisions. Inspection of the table shows that across the five-year period, 93% and 94% of examinees of the February and July examinations were predicted to have the same outcome under the UBE. These results were fairly consistent across examination years, varying by no more than 2%.

In terms of inconsistent decisions from actual and simulated procedures, we observed a pattern of a higher number of test-takers passing the actual exam but failing the simulated UBE than vice-versa. This result was detected in nine out of the ten administrations.

Table 5

Actual vs. Simulated UBE
Texas Bar Examination
Decision Consistency Rates
2013 through 2017

	Agreement: Actual & Simulated		Actual Pass/ Simulated Fail		Actual Fail/ Simulated Pass	
	N	%	N	%	N	%
Febr.						
2013	1,102	93%	53	4%	30	3%
2014	1,080	94%	43	4%	29	3%
2015	1,229	92%	53	4%	51	4%
2016	1,336	93%	55	4%	42	3%
2017	1,152	92%	41	3%	60	5%
5-Year	5,899	93%	245	4%	212	3%
July						
2013	2,868	95%	105	3%	50	2%
2014	2,779	95%	97	3%	53	2%
2015	2,761	93%	131	4%	95	3%
2016	2,797	94%	111	4%	67	2%
2017	2,803	94%	104	4%	52	2%
5-Year	14,008	94%	548	4%	317	2%

Examination Reliability. As described above, the configurations of the UBE and the current TBE differ not only in test content (i.e., no P&E, fewer subject matters covered), but also on test length (fewer essay questions but an expanded MPT) and weighting of the respective sections. To estimate the impact of these changes on overall test reliability, we calculated a total "Written Test score" based on the 6-question essay test and a 2-question "MPT" test.[9] We then estimated the simulated UBE Total Test reliability by equally weighting the written test and MBE sections. Table 6 on the following page presents the estimated reliabilities along with the actual reliabilities previously reported in the technical review of the TBLE's performance during the same five-year period.[10]

The results in Table 6 show that the simulated UBE Total Test score reliability is slightly lower than the actual reliability, but still consistently above the .85 criterion. We observed only minor variation from year to year, suggesting that the estimates are fairly stable. We acknowledge that the absence of a true 2nd MPT in the simulation would possibly impact these findings. However, since that question would make up only 10% of the overall score, and there is a generally

[9] Since simply using the current Texas MPT test twice would artificially increase reliability, we estimated a 2nd MPT score based on 3 essay questions that were not used to simulate the UBE Essay portion. We reasoned that that those three tests would be based on 90 minutes of testing, which is the length of time given to an MPT.
[10] Bolus 2018.

moderate relationship between MPT and Essay scores, we suspect that the effect would be minimal.

Table 6

Reliability Estimates of Original & UBE-Simulated Texas Bar Examination Total Scores 2013 through 2017

	Actual Total Score	Simulated UBE Total Score	Difference
February			
2013	0.892	0.860	0.032
2014	0.898	0.867	0.031
2015	0.881	0.847	0.034
2016	0.892	0.869	0.023
2017	0.894	0.862	0.032
5-Yr Ave.	0.891	0.861	0.030
July			
2013	0.868	0.862	0.006
2014	0.900	0.865	0.035
2015	0.896	0.852	0.044
2016	0.903	0.866	0.037
2017	0.912	0.891	0.021
5-Yr Ave.	0.896	0.867	0.029

IV. SUMMARY AND DISCUSSION

The Texas Supreme Court is considering whether changes to the TBE are warranted. An appointed Task Force has recommended either transitioning to the UBE or shortening the current exam to two days.

This report presents the outcomes of statistical simulations that allow for a comparison of actual exam results from a five year-period with the predicted results from several alternative configurations of the TBE under consideration. These analyses estimate the impact of structural changes on key test metrics including overall bar passage rates, ultimate pass-fail decisions for individual test-takers and the statistical reliability of test scores.

Table 7 presents a summary of the study findings.

Table 7

A Summary of Statistical Findings
Comparing the Current Configuration of the
Texas Bar Examination
To Three Alternatives
2013 through 2017

	February				July				Overall			
	Current	"2-Day" 40/40	"2-Day" 50/30	"UBE"	Current	"2-Day" 40/40	"2-Day" 50/30	"UBE"	Current	"2-Day" 40/40	"2-Day" 50/30	"UBE"
Metric												
Pass Rate	61%	62%	61%	60%	72%	72%	73%	71%	69%	70%	69%	68%
Decision Consistency		94%	95%	93%		94%	95%	94%		94%	95%	94%
Test Reliability	.89	.83	.87	.86	.89	.83	.87	.87	.89	.83	.87	.86

Table 7 substantiates that, statistically, any of the three alternatives considered in the report would be viable options. Reducing the current examination to a 2-Day examination, by reducing the number of essay questions, is projected to have a negligible impact on either February or July bar passage rates. A UBE-like configuration, eliminating the P&E resulted in only a 1% decrease. The simulated total scores for each of the models resulted in highly consistent pass-fail decisions as well.

The only difference of concern emerged in relation to the overall test reliabilities. As expected, shortening the length of the essay examination reduced the reliability of this section of the exam. As a result, the 2-Day model that maintained the same MBE weighting of 40% as in the current

configuration, reduced the overall test reliability below the commonly adopted .85 threshold although only by a few points. A 2-Day TBE that increased the MBE weighting to 50%, and a UBE configuration that dropped the P&E but also increased the MBE weight both achieved estimated reliabilities that met the .85 criterion.

Certain caveats must be considered in evaluating these findings. Given the lack of available demographic data, we were unable to determine whether some subgroups would be impacted differently. For example, similar studies conducted for other states in which demographic data has been included, have shown that women generally perform better than men on the written sections, while the converse is true for the MBE. Therefore, configurations that gave more weight to the MBE tended to predict slight increases in the passing rates of men.

Other cautions relate to assumptions regarding test content that were necessary for the purposes of our simulations. One of the models that we formed required us to construct a 2nd MPT score based upon the available MPT score. While the logic that we used for this calculation was based on procedures from prior studies, it is possible that when faced with two different performance tests, some test-takers may perform quite differently, and simply doubling a score may have led to a faulty estimate. Similarly, our simulations were based on examinees' performance on Texas-constructed essays rather than the Multistate Essay questions constructed by NCBE used in the UBE. It is unclear as to how similar the content would be.

It is also important to recognize that an actual change in exam format can have ripple effects that cannot be anticipated in a statistical simulation. Examination preparation practices have been known to change when the importance of one section of the examination is weighted more heavily than another. While a source of extraneous error, examinees have been known to perform both better and worse under shortened or extended testing periods.

We also note that applicants' native legal abilities may change in the future. As they become better (or less) prepared, there is no way of predicting whether the relationships observed on historical examinations would carry forward. Finally, the behavior of graders in the assignment of grades could be subject to change as well.

Having mentioned these cautions, the results that we present in this report tend to be quite consistent with those we have seen in modeling results in other states. Additionally, in the two states that have implemented changes in their exams, subsequent analysis of examinee's live performance on the modified examinations have validated the results predicted by the modeling.

APPENDIX 1

Estimating Examinee Final Status in Re-grade

Since examinees who fell within the re-grade range in the simulations may not actually have a real 2nd grading, we estimated their final outcome using the known behavior of graders during the actual re-grade process. First, for each of the ten examinations, we calculated the percentage of applicants in re-grade that were given enough additional points to pass. We assumed that graders would have used the same standards with the revised scores. We then randomly assigned a fractional value between 0 and 1 to every examinee who went into re-grade based on their simulated scores. If the fractional value for an examinee was below the target percentage, then examinee was assigned a passing status; if not, then the applicant was given a failing status.

Based on the Bolus 2018 Technical Report, it was shown that 94% of examinees who went to re-grade passed.

REFERENCES

American Educational Research Association, American Psychological Association, and the National Council on Measurement in Education, (2014 revised.). *Standards for Educational and Psychological Testing.*

Task Force on the Texas Bar Examination (2018). *Recommendations and Report: Final Report to the Texas Supreme Court*.

Bolus R. (2018). *A Technical Report on the Texas Bar Examination*. A report prepared for the Texas Board of Law Examiners