#81-1PR

AN ANALYSIS OF POSSIBLE VARIATIONS IN PASS/FAIL STANDARDS ON THE CALIFORNIA BAR EXAMINATION

Stephen P. Klein, Ph.D.
January 15, 1981

PURPOSE

This study was conducted to investigate whether there has been any variation over time in the standard used for determining whether an applicant has passed California's bar examination. The impetus for this study came from the relatively low pass rate on both the February and July examinations given in 1980.

MEASURES

California's bar examination has two sections, Multistate Bar Examination and Essay.

The Multistate Bar Examination (MBE) consists of 200 multiple choice questions drawn from six content areas. About 150 of these questions are unique to each administration; i.e., they have not been used previously. The remaining 50 questions, called linkage items, are a representative sample of questions that have been used previously. The purpose of the linkage items is to adjust the scores on the MBE to take into account the relative difficulty of the 150 unique items. For example, if on different administrations the average score on the linkage items remains at 33, but the average score on the 150 unique items is 107 on one administration and 104 on a subsequent administration, then it can be inferred that the 150 questions used on the second occasion were, as a set, slightly more difficult than the unique questions used in the previous administration of the MBE. In other words, since the average score was the same for different groups of applicants on the common set of questions but the average score was not the same on their respective sets of unique items, it is assummed that the two unique sets must have differed in difficulty. The total scores of the applicants who took the second set of unique questions are, therefore, adjusted upwards by three points so as to compensate for their having taken a slighlty more difficult version of the MBE. Thus, as a consequence of this adjustment process, an MBE score obtained on one administration can be compared directly with an MBE score obtained on another administration.

The answer to a given question on the Essay portion of the examination is graded by one or more members of a team of specially trained and calibrated attorneys; i.e., there is a different team of readers for each question. Scores are assigned to answers in five point intervals on a theoretical scale of 0 to 100; however, most of the scores usually fall between 55 and 90.

Between February, 1976 and July, 1980, there has been a reduction (from 12 to 9) in the number of essay questions applicants were instructed to answer. During this same period, the average time that was allocated to answer each question was increased from 52.5 to 60 minutes; and, the rules were changed to no longer allow applicants some choice in which questions they answered.

An applicant's Total score on the examination is a weighted composite of the MBE and Essay scores. The current policy is to assign 40 percent of the theoretical maximum number of points an applicant can earn on the examination to the MBE and the remaining 60 percent to the Essay. For example, an applicant can currently earn a maximum score of 1500 points (9 essay questions at 100 points each plus three times the applicant's adjusted score on the MBE). When the Essay portion contained 12 questions and given 70 percent of the weight, then the MBE scores were multiplied by 2.57 to yield a theoretical maximum of 1714 points; i.e., (12)(100) + (2.57)(200) = 1714.

The nominal weights of 70:30 or 60:40 of Essay to MBE do not necessarily reflect their actual effective weights. The reason for this is that the relative influence of a section score on a total score is a function of the relative standard deviations of the two scores. It is not a function of the number of points assigned to each section or even the average score on each section. For example, previous research on California's examination has indicated that when the nominal weights of 70:30 were in force, the actual weights were closer to 65:35.

An applicant can pass the examination in either or both of two ways: (1) earning 70 percent or more of the theoretical total score or (2) earning 70 percent or more on each section (i.e., as a result of taking the examination more than once).

A screening procedure (involving an applicant's score on the MBE and the scores on a subset of essay questions) was implemented in 1978 to estimate whether the applicant was likely to pass by the first method if the remainder of that applicant's essay answers were read. Applicants with extremely high probabilities of passing (i.e., greater than 99.5) were passed by this screening process without any further evaluation of their answers. The remaining applicants had all of their essay answers read at least once. Applicants who came close to the pass/fail line after a single reading of all of their essay answers had their essay answers graded again by a different set of readers. If the average of these two readings plus the applicant's MBE score resulted in the applicant falling just below the pass/fail line, then the applicant was placed in the reappraisal category.

All the essay answers written by an applicant in the reappraisal category were evaluated as a set by one or more members of the Board of Reappraisors. The individuals who serve on this board are attorneys with several years of experience in grading California essay answers. Although the reappraisal process also has varied somewhat over time, it has always involved reviewing an applicant's essay answers to determine whether enough extra points can be found to pass the applicant. Thus, reappraisors are aware of how many points an applicant needs to pass. Previous research has indicated that with few exceptions, the reappraisal process results in passing applicants who fell just below the pass/fail line and failing applicants who fell well below it.

SAMPLE

As a result of previous research, data were readily available on nine of the ten bar examinations given between February, 1976 and July, 1980. Data on the February, 1978 administration had not been used in these studies. Moreover, the time and cost required to obtain the February, 1978 data did not seem warrented given the overall purpose of the present research. Thus, only MBE scores, obtained from the National Conference of Bar Examiners, will be reported for the February, 1978 administration.

With the exception of the February, 1978 test, the sample used in the analysis of each examination consisted of those applicants who took both the MBE and Essay portions. Attorney applicants and others who took only one section (i.e., as a result of passing one section on a previous administration, illness, etc.) were excluded from the applicant pool. Thus, the results presented in this report may not conform exactly with previously published statistics on each examination.

ANALYSIS PLAN

Differences or similiarities in the percent passing different administrations of an examination do not in themselves indicate whether a consistent pass/fail standard has been used over time. This is because the average level of legal skills and knowledge of the applicants taking one examination can differ from the average level of those taking another examination. For example, there is a larger proportion of repeaters (i.e., applicants who failed one or both sections previously) in February than in July. The February applicants also tend to have lower law school grades and admissions scores. Thus, one would not expect the passing rate in February to be as high as it is in July.

Moreover, setting a passing score of "70 Percent of the Theoretical Maximum Score" may not result in a consistent standard because the difficulty of the Essay portion of the examination (and the leniency with which it is graded) may vary across administrations. There also can be variation across administrations in the "nominal" and "effective" weighting of the Essay section relative to the MBE.

The foregoing considerations led to the development of a bar examination score that would not be affected by possible variations across administrations in the the difficulty of the MBE and Essay portions. The procedure for deriving this score was as follows:

1) The average score on an essay question was computed for each applicant so as to adjust for the variation both across and within examinations in the number of questions answered and graded.

2) California MBE scores were divided by either 2.57 or 3.00 (depending upon the particular administration of the examination) so as to put them back into scale score form. It will be recalled that the MBE scores in this form had already been adjusted, through linkage items, for possible differences in test difficulty across administrations.

3) The applicants' average essay question scores on a given administration of the examination were transformed to the same mean and standard deviation as their MBE scores. Because of the strong correlation between MBE and essay scores, this transformation resulted in adjusting the essay scores for possible differences across administrations in the difficulty of the questions asked and/or in the leniency with which the answers to these questions were graded.

4) A Scale score was computed for each applicant that reflected the current policy of weighting the Essay and MBE portions 60:40, respectively. The formula for computing this score was:

Scale score = (.60)(Transformed Average Essay Score) + (.40)(MBE)

If the same level of performance is required for passing across administrations, then the Scale score corresponding to the pass/fail line on one administration should be the same as the Scale score needed for passing on every other administration. For example, if 38 percent of the applicants pass on one occasion and 60 percent on another, then the Scale score above which there were 38 percent on the first occasion should be identical to the Scale score above which there were 60 percent on the second occasion. If less than 60 percent of the applicants who took the second administration had Scale scores above the one that was needed to pass 38 percent on the first administration, then it is apparent that the level of performance required for passing the second administration was higher than the level needed for passing the first administration.

In summary, similiarities or differences in passing rates across administrations do not necessarily reflect consistent or inconsistent pass/fail standards because of possible variations in the legal skills and knowledge of the applicants taking those examinations. Similarly, setting a certain score as the pass/fail line, such as 70 percent of the maximum possible points, also does not guarantee consistency because the examination can vary in difficulty across administrations. The foregoing considerations led to the computation of a Scale score for each applicant that reflected that applicant's performance level in a way that was independent of the difficulty of the particular version of the examination on which the score was earned.

The major value of the Scale score is that it permits measuring the degree to which California has been using a consistent performance standard in determining an applicant's pass/fail status. This is done by assessing whether the Scale score that was apparently needed for passing on one administration has been essentially the same as that needed for passing on another administration. If these Scale scores are the same, then: (1) a consistent standard has been used and (2) any differences in passing rates between administrations can be attributable to general differences in the legal skills and knowledge of the applicants taking these examinations.

RESULTS

Tables 1 and 2 contain summary statistical data on the February and July examinations, respectively. The percent passing in these tables includes applicants who passed as a result of reappraisal; i.e., it reflects final pass/fail decisions. It does not reflect the additional (but small) percent of applicants who passed the July, 1980 examination as a result of their participation in that examination's Special Session and/or Assessment Center.

The numbers in the column headed "Passing Score" are the Scale scores that would have passed the same percent of applicants as actually did pass. For example, 38 percent of the applicants passed the February, 1976 examination. And, an inspection of the distribution of Scale scores on this examination indicated that 38 percent of the applicants had Scale scores of 141 or higher.

The data in the CA column under the heading "Average MBE" are the average MBE scores of the California applicants. The Non-CA column contains the average MBE scores of the applicants in all the other jurisdictions that gave the MBE.

Table 1

SUMMARY STATISTICS ON FEBRUARY EXAMINATIONS.

Exam Date	Number Taking	% Pass	Passing Score	% 1st Timers	% ABA	Average MBE CA	Average MBE Non-CA	ABA First Timers % of Scale	ABA First Timers Average MBE*	ABA First Timers % Pass
2/76	3088	38	141	38	49	137	134	16	144	61
2/77	3399	44	141	35	43	139	134	14	146	72
2/78						139	134			
2/79	4166	45	141	32	43	139	134	13	146	65
2/80	3758	34	141	36	42	136	134	14	145	59
Mean	3603	40	141	35	44	138	134	14	145	64

Table 2

SUMMARY STATISTICS ON JULY EXAMINATIONS.

Exam Date	Number Taking	% Pass	Passing Score	% 1st Timers	% ABA	Average MBE CA	Average MBE Non-CA	ABA First Timers % of Scale	ABA First Timers Average MBE*	ABA First Timers % Pass
7/76	6709	60	143	77	64	145	142	54	152	80
7/77	7191	55	142	77	61	143	140	52	150	76
7/78	6835	55	143	80	62	145	141	55	151	73
7/79	7152	55	142	74	62	144	140	53	150	75
7/80	7379	49	142	69	62	141	140	50	148	73
Mean	7053	55	142	75	62	144	141	53	150	75

An inspection of the data in Tables 1 and 2 indicated the following:

- The performance standard required for passing the examination has remained essentially constant over the nine examinations studied. It was observed, however, that the standard used in February (141) has been slightly lower than the one used in July (142). Thus, even though more applicants pass in July, the February test is very slightly easier.

- The February, 1980 applicants had the lowest average MBE score (136) of all the tests studied. On the only other examination which had a mean MBE almost as low as 136 (i.e., February, 1976), the passing rate (61%) was just two percentage points higher than the February, 1980 rate (59%). Moreover, a higher percentage of the February, 1976 than the February, 1980 applicants were first time takers from ABA schools.

- The July, 1980 applicants had the lowest average MBE score (141) of all the July groups studied. This average also was only one point higher than the national non-California average on this examination (140) rather than the usual three to five points higher. One factor that probably contributed to California's low average MBE score in July, 1980 was the especially low percentage of applicants who were taking the examination for the first time.

- The unusually low average MBE score among all July, 1980 examinees also was present in the subset of applicants from ABA schools who were taking the examination for the first time. The reason that the ABA first timers in July, 1978 and 1980 had the same passing rate (despite their different average MBE scores) is that the standard for passing in July, 1978 was slightly higher (143) than it was in July, 1980 (142).

SUMMARY AND CONCLUSIONS

A Scale score was computed for each applicant. This score adjusted for differences across administrations in the difficulty of the questions asked, the leniency with which essay answers were graded, the number of essay questions answered and graded, and the weight attached to the MBE and Essay sections in arriving at a total score.

A comparison of the distribution of Scale scores to the percent passing was made for nine of the ten bar examinations administered between February, 1976 and July, 1980. The results of this analysis indicated that the level of performance required for passing the California bar examination has remained essentially constant over the last five years. It was observed, however, that the Scale score needed for passing was slightly higher in July (142) than it was in February (141). It was further observed that the average MBE scores of the 1980 applicants were lower than usual. Thus, this difference in average score, rather than any changes in California's standards for passing, was the apparent source of the relatively low passing rate among February and July, 1980 applicants.

Appendix B

HOW THE MBE IS SCALED ACROSS ADMINISTRATIONS OF THE EXAMINATION

Each form of the MBE contains two groups of items. One set is new items; i.e., they have not appeared on previous forms of the test. Usually about 150 of the 200 items on each form fall in this group. The remaining set of about 50 items are called "scaling" or "linkage" items. These items have appeared on one or more previous forms of the test.

The hypothetical data in the table below will be used to illustrate how the applicants' performance on these two groups of items are used to scale the MBE scores across administrations.

HYPOTHETICAL AVERAGE SCORES

Examination Date	Linkage Items	Unique Items	Raw Score	Scale Score
July 1980	35	105	140	140
July 1981	35	100	135	140
February 1982	30	115	145	120
Number of Items	50	150	200	200

A linkage item is a question that is used across administrations of the examination whereas a unique item is used on only one administration of the examination.

An inspection of these data indicates that the applicants who took the July 1980 examination had the same average score on the linkage items as did the applicants who took the examination in July 1981. It may be inferred, therefore, that the two groups of applicants were generally comparable with respect to the skills and knowledge that are measured by the MBE. In other words, since they had the same average score (35) on the same set of 50 items, we assume the two groups were equal in ability.

Although the 1980 and 1981 applicants performed at the same level on the linkage items, the former group had a higher average score (105) than the latter group (100) on their respective sets of unique items. The source of this five-point discrepancy must therefore be that the 1981 unique items were more difficult than were the 1980 items since we know from the linkage items that the two groups were equal in ability. Thus, in order to make the 1981 scores comparable to the 1980 scores, all the applicants who took the 1981 examination would have five points added to their total score. In other words, the total raw scores in 1981 are adjusted or "scaled" upwards to reflect the fact the unique questions asked in 1981 were slightly more difficult than those asked in 1980.

The performance on the linkage items of the applicants who took the examination in February 1982 is below that of the applicants who took the examination in July of 1980 or 1981. It may be inferred, therefore, that the applicants who took the examination in February of 1982 were less able (in terms of the skills and knowledge measured by the MBE) than were the other two groups of applicants even though the

1982 group had the highest average score on its set of unique items (115) and on the total test (145). It is evident, therefore, that the unique questions asked in 1982 must have been substantially easier than those asked in 1980. To make the scores on the 1980 and 1982 versions of the test comparable, one must scale each 1982 raw score downwards by 25 points. In other words, if the difference between the two groups on the common set of 50 linkage questions was 5 points, and if the entire 200 item examination had been equally difficult across administrations, then there should have been a 20 point difference between the groups in their respective total scores. This difference is obtained by the scaling process.

The foregoing is presented solely to illustrate how linkage items can be used to equate two different forms of a test. The actual scaling procedures are somewhat more complicated, because they consider how the groups spread out on the linkage items (not just their respective averages) as well as the correlation between the linkage and unique items in each group. Moreover, good test construction practices generally produce forms that are already almost equivalent in difficulty; i.e., the amount of the adjustment required is usually far less than that indicated in this hypothetical example.

The foregoing discussion also points to the extreme importance of maintaining test security. In other words, the scaling process assumes that applicants who take a future examination have no advantage on the linkage items relative to the groups who took these items previously. If the linkage items' security is breached (such that they become available to a significant proportion of new applicants), then they cannot be used for scaling purposes.

Appendix D

SCALING METHODS

I. Standard Deviation Method

A. Steps

1. Compute the mean and standard deviation of the MBE Scale scores and the Raw Essay scores. The following notation may be used:

	Score on the test	Mean Score	Standard Deviation	Standard Score
MBE	mbe	Mmbe	SDmbe	Zmbe
Essay	e	Me	SDe	Ze

2. Convert each applicant's raw essay score to a standard score using the following formula:

$$Ze = \frac{e - Me}{SDe}$$

3. Insert the standard score (Ze) into the following formula to obtain an applicant's Essay Scale score:

Essay Scale Score = (Ze)(SDmbe) + Mmbe

4. Check the results. The mean and standard deviation of the Ze scores should equal 0.00 and 1.00, respectively. The mean and standard deviation of the Essay Scale scores should equal the mean and standard deviation of the MBE Scale scores, respectively.

B. Numerical Example

1. Given the following data:

	Mean	Standard Deviation
MBE Scale	133.06	16.40
Raw Essay	155.69	17.78

2. An applicant with a Raw Essay score of 150 would have a standard score of:

$$Ze = \frac{150 - 155.69}{17.78} = -.32$$

3. An applicant with a Raw Essay score of 150 would have an Essay Scale score of:

Essay Scale Score = (ze)(SDmbe) + Mmbe = (-.32)(16.4) + 133.06 = 12

II. Equipercentile Method

A. Construct the cumulative frequency distribution of the Raw Essay Scores. This will be a table in which one column will list (from lowest to highest) all of the different essay scores that were assigned. The second column indicates the percent of all applicants who achieved each score or lower.

Possible Raw Essay Scores	Cumulative Percent
79	0.2
93	0.5
101	0.7
102	1.0
103	1.2
110	1.4
111	1.7
113	1.9
115	2.1
116	2.6
:	:
:	:
192	100.0

Note: The lowest and highest Essay scores on this examination were 79 and 192, respectively.

The cumulative percents do not move in equal units because there are different numbers of applicants earning each score.

B. Construct the cumulative frequency distribution of the MBE Scale Scores.

Possible MBE Scale Scores	Cumulative Percent
85	0.2
91	0.5
93	0.7
94	1.0
95	1.2
98	1.7
99	2.4
100	2.6
101	3.6
:	:
:	:
176	100.0

C. Merge the two distributions into one table of three columns so that for a given cumulative percentile value in the Raw Essay table, one can find the corresponding value in the MBE Scale score table.

Possible Raw Essay Scores	Cumulative Percent	Corresponding MBE Scale Score
79	0.2	85
93	0.5	91
101	0.7	93
102	1.0	94
103	1.2	95
110	1.4	96.2
111	1.7	98
113	1.9	98.3
115	2.1	98.6
116	2.6	100

Note that to find the corresponding MBE Scale score for a given Raw Essay score may require the mathematical procedure called interpolation. For example, the MBE Scale score below which there are only 2.1 percent of the applicants (i.e., MBE Scale score corresponding to a Raw Essay score of 115) is obtained as follows:

$$\left\{\frac{2.1 - 1.7}{2.4 - 1.7}\right\} \quad (99 - 98) + 98 = 98.57$$

Similarly, MBE equivalent of a Raw Essay score of 110 is given by the following computations:

$$\left\{\frac{1.4 - 1.2}{1.7 - 1.2}\right\} \quad (98 - 95) + 95 = 96.2$$

D. To find an applicant's Essay Scale score using this method, read down the column of Raw Essay scores until you come to the applicant's Raw Essay score. Read across this row to the column headed "Corresponding MBE Scale Score." The value in this column will be the Essay Scale score for that applicant.

III. Comparison of the Two Methods

A. In general, the two methods will yield fairly similar results (especially for the bulk of the applicants) provided that both score distributions are similar in shape.

B. The equipercentile method is usually preferred when the shapes of the two distributions are not similar. However, the standard deviation method tends to give a "better" estimate for extremely high and low Raw Essay scores. Because of the generally strong correlation between the Essay and MBE and because only a pass/fail decision is needed at a point some distance away from the extremely low score end, the equipercentile method is probably the best approach for most jurisdictions.

C. The table below compares the two scaling methods for a Southeastern state with slightly more than 400 applicants:

Raw Essay Score	Cumulative Percent	Corresponding Essay Scale Scores	
		Equipercentile	Standard Deviation
110	1.4	96.2	90.9
120	3.2	100.6	100.2
130	8.1	109.3	109.4
140	18.6	117.9	118.6
150	35.8	127.2	127.8
160	57.3	134.9	137.0
170	78.0	145.8	146.3
180	94.5	160.5	155.5
190	99.8	175.0	164.7