

SUMMARY OF STUDIES CONDUCTED FOR
THE COMMITTEE OF BAR EXAMINERS
AND THE STATISTICAL RATIONALE
UNDERLYING PROPOSED CHANGES

Prepared by
Stephen P. Klein
August 25, 1978

This report summarizes the results of a series of studies conducted on the State of California General Bar Examination. These studies have provided information regarding the statistical properties of the present examination process and the effects of certain modifications that might be made to improve this process. A summary of the changes that have been adopted and the rationale for them also are discussed.

BACKGROUND

The General Bar Examination has two sections. The Essay portion consists of three sets of five questions each. An applicant is instructed to select four questions per set to answer in the three-and-one-half hours allocated to each Essay test session. A given question involves several issues from one or more content areas, so that the Essay Test as a whole covers the major subject areas taught in law school. These questions are designed to assess an applicant's ability to analyze legal problems and are usually initially developed by non-California law professors. The questions are then revised and edited by the Committee of Bar Examiners, its staff, and the Board of Reappraisers.^{*/}

During the past few years, four to five readers (all of whom are attorneys who have passed the examination) are assigned to each essay question. This team, under the direction of a

^{*/} The Board of Reappraisers is composed of nine attorneys who have had extensive experience as readers.

member of the Board of Reappraisers, researches the appropriate ways of responding to the question, reviews an analysis of it that was prepared by the professor who wrote it initially, and reads a small sample of the answers that were written to it. In this way, the team develops mutually agreed upon scoring criteria and standards. The answers written by a group of 75 applicants are then independently graded by all the members of the team in order to assess the extent to which they are indeed calibrated to one another. Discrepancies in the interpretation of grading procedures identified by this process are resolved via a conference with the Committee of Bar Examiners, its staff, and the Board of Reappraisers. Once these steps are completed, all the answers to the question are divided among the team members assigned to it, although periodically certain answers are graded and discussed by all the members of the team in order to help maintain interreader consistency.

The second section of the examination consists of a six-hour multiple choice test, called the Multistate Bar Examination (MBE). This test contains 200 questions that are divided among six content areas: Constitutional Law, Contracts, Criminal Law, Evidence, Real Property, and Torts. The MBE is developed by the National Conference of Bar Examiners and is scored by the Educational Testing Service of Princeton, New Jersey.

Each of the 12 essay questions answered by an applicant is worth up to 100 points. Scores on the MBE are weighted to a maximum of 514 points so as to reflect the Committee of Bar

Examiner's policy that this section account for 30% of an applicant's total score (i.e., the maximum total score is 1714 points).

An applicant may pass the examination in one of two ways. The first method is by receiving a combined Essay and MBE score of 70% (1200 points) or higher. The second method is by passing each section of the examination separately with minimum scores of 840 and 360 for the Essay and MBE, respectively. In other words, an applicant may pass one portion on one administration of the examination and the other portion on a different administration. Since an applicant's passing status on one section cannot be reversed by subsequently failing that section, it is clearly in an applicant's best interests to take both portions each time he or she attempts to pass the examination as a whole. In that way, the applicant may be able to pass by either of the two methods.

In the past, if an applicant came close to passing but still failed the examination (or the Essay portion of it), that applicant's Essay answers were turned over to the Board of Reappraisers. The reappraisers assigned to a particular applicant reviewed all 12 of the applicant's answers as a set so as to make an overall pass/fail decision. This meant that an applicant's efforts were evaluated as a whole by readers who had extensive experience in grading bar examinations and who had also been intimately involved in the question development and scorer calibration processes.

COMPARISON OF MBE AND ESSAY TEST CHARACTERISTICS

Test Difficulty

Using the criterion that 70% or higher of the total possible score on a test constitutes passing, then an applicant has about a 16% better chance of passing the MBE than the Essay. In other words, about 16% more applicants pass the MBE than pass the Essay section of the examination. And, this differential remains relatively constant across racial/ethnic groups.

Weighting

Relative to the total possible points an applicant can earn on each section of the examination, MBE scores spread out somewhat more from the average MBE score than Essay scores spread out from their average. The effect of this difference in the shapes of the two distributions of scores has been to let the MBE have about a 35%, rather than just a 30% influence on determining an applicant's performance level on the examination as a whole. Since the two tests differ in difficulty, assigning more weight to the MBE than intended has resulted in a slightly higher passing rate than would have occurred if the 70:30 policy was strictly enforced.

Reliability

The term "reliability" in the present context refers to the degree to which an applicant performed in a consistent manner

across the questions asked within each test. Two requirements have to be met in order for such consistency to occur. First, the general skills and knowledge that are needed to answer one question are also needed to answer the other questions on the test. And second, the applicant's performance on these questions is not influenced by chance events, such as being asked a question related to an esoteric aspect of the law with which the applicant was particularly familiar. In short, the more a score on a question is influenced by isolated skills, highly unique knowledge, scoring inconsistencies, and random factors (such as luck), the lower the reliability of a test containing such questions.

The MBE portion of the examination has a reliability that meets the generally accepted minimum standards for a test that is used in making important decisions about individuals. Although the Essay test by itself falls short of this criterion, the total examination score (i.e., MBE plus Essay) is consistently able to achieve the necessary level of reliability. Analyses of the past several examinations have further indicated that this standard would not be met if there was a reduction in the number of Essay questions answered per applicant unless other changes were made in the examination process, such as having more than one reader grade each answer.

Abilities Assessed

There is a strong positive correlation between an applicant's performance on the MBE and Essay sections of the test. In other

words, applicants who tend to do well on the MBE also do well on the Essay and vice versa. For example, one can predict with about 75% accuracy an applicant's pass/fail status on the Essay from knowledge of that applicant's MBE score. An even higher level of predictive accuracy probably would be obtained if the Essay scores were as reliable as the MBE. It appears, therefore, that the two tests are measuring very similar or at least highly related skills and knowledge even though their formats (multiple choice and essay) are quite different.

It is also evident that these skills and knowledge are closely associated with the abilities required to perform well in law school. This conclusion is based on the strong relationship between bar scores and law school grade point averages. For example, after adjusting law school grade point averages for variations in the general ability levels of students attending different schools, these averages are able to predict bar scores about four times better than the combination of undergraduate grades and law school admissions test scores are able to predict first year grades in law school. In other words, the relative performance levels of students in law school correspond closely with their performance on the bar examination, especially when one takes into consideration the marked differences in the overall caliber of the students attending different schools.

Equivalency of Examinations Across Administrations

The procedure used to equate scores on different administrations of the MBE involves embedding in each new form of the test several questions that have been used in previous forms. An analysis of how applicants taking the new form perform on these questions relative to how well applicants taking previous forms did on them is then used in scaling the total MBE scores. For instance, if the applicants taking a new form of the MBE do better on the scaling questions than applicants taking previous forms of the test, then it is assumed that those taking the new form are a more able group, and their scores on the new form are adjusted accordingly.

The foregoing procedures cannot be employed in assessing the equivalency of different administrations of the Essay test because none of its questions are ever used twice. However, the strong correlation between the Essay and MBE, as well as the equivalency of the MBE scores across administrations, permitted using this test to examine whether applicants earning similar MBE scores on different administrations were indeed exhibiting comparable levels of Essay performance. The results of this analysis indicated that about the same Essay score was assigned across test administrations for a given level of performance on the MBE. For example, applicants with MBE scores of 351 to 375 had average Essay scores of 820, 811, 818, 824, and 820 on

the six bar examinations administered between February 1975 and July 1977. It appears, therefore, that scoring standards have remained pretty much the same across different administrations of the test. This means that an applicant's chances of passing the examination are not appreciably influenced by taking it on any particular occasion.

The foregoing findings also indicate that the higher average scores and correspondingly higher passing rates on the July versus the February examinations are due to differences in the general level of ability of the applicants on these two testing dates rather than any variation in grading standards. This conclusion is supported further by the substantially larger percentage of applicants who take the examination for the first time during its July administration.

RESULTS OF RECENT RESEARCH ON BAR EXAMINATION PRACTICES

Differences in Passing Rates Between Groups

Various studies have been conducted on the bar examination to investigate whether the scores on it are related to an applicant's sex or racial/ethnic group. While some of these studies are still underway, the initial results are as follows:

1. The magnitude of the difference in average performance levels between racial/ethnic groups remains relatively constant across the examination's essay questions.

The same is true for the multiple choice questions and subtests. Thus, certain questions or subject matter areas are not especially more or less difficult for a given group relative to the performance of other groups taking the test. These results indicate that individual questions or subtests do not differentially favor or discriminate against a particular group.

2. The magnitude of the difference in average Essay scores between groups does not change when applicants have their answers graded by readers whose racial/ethnic group is the same versus different than their own. Thus, the racial/ethnic group of the reader does not diminish or enhance a minority or majority group applicant's chances of passing the examination.
3. Controlling for potentially relevant attitudinal and background characteristics of the applicants, such as their socioeconomic status and whether they took an in-depth bar review course, does not change the between-group disparities in their respective passing rates. For instance, the passing rates for minority and majority applicants who feel that the examination is "not a fair test" of their legal skills and knowledge are essentially the same as the passing rates for all the members of their respective groups.

4. Controlling for an applicant's performance in law school as well as the overall caliber of the students attending different schools substantially reduces between-group differences in their passing rates. This is especially true for those minority applicants who have above-average law school grades. In other words, the overall differences in the passing rates between groups appears to be primarily a function of differences in their relative academic achievement levels in law school. And, the differences in passing rates that still remain after this factor has been controlled are due to some factor or set of factors that is not applicable to all minority group applicants.

Effect of Time Limits on the Essay Test

While no study has been done on the bar examination's time limits per se, an experiment has been conducted with the same type of question that is used on its Essay test. This study involved varying the time allowed to answer essay questions that were administered as part of the First Year Law Students Examination. This test was chosen in part because it was felt that the performance of those taking it would be especially sensitive to any benefit that might be derived from an increased time allocation.

The study's design varied the time limits by giving one half of the examinees 65, rather than the regular 52 1/2 minutes

to answer one question and then giving the other one half of the applicants the additional 12 1/2 minutes to answer another question. The two sets of answers to each question were inter-mixed prior to scoring so that the graders did not know which ones had been written under the expanded time limit.

The results of this investigation indicated that, on the average, an examinee's score on a question improved only about one point with the additional time allocation. Moreover, the relative performance of the examinees across questions was unaffected by how much time they had to answer them. It appears, therefore, that increasing the Essay test's time limits would not appreciably affect the quality of the answers written or who was able to write a passing answer.

Essay Grading Process

It was noted above that the reliability of the Essay test has not been as high as the MBE even though far more applicant time and a great deal more of the bar's financial and personnel resources have been devoted to it. The Essay test also carries substantially more weight in determining an applicant's pass/fail status than does the MBE. This situation led to a series of studies that were designed to determine the cause of this discrepancy in reliability and the efficacy of alternative means of dealing with it.

Scorer Reliability. The first of these studies involved an analysis of the extent to which readers agreed with themselves and with one another on the scores that should be assigned to a given answer. Thirty answers to each of four essay questions appearing on the February 1977 examination were selected for this purpose. These answers were then embedded into the total batch of answers that were graded by each reader during the normal scoring process; i.e., all four readers assigned to a given question graded the same set of thirty answers along with the several hundred other answers that each reader was separately responsible for scoring. This insertion of the special set of thirty was done twice, once near the beginning of the six-week reading period and again near the end.

The results of this analysis indicated that a substantial portion of the variation in an applicant's scores across Essay questions was due to the grading process rather than to the uniqueness of the skills and knowledge required to answer different questions. For example, two readers of the same answer usually only agreed with one another about 67% of the time as to whether that answer was or was not "passing" (i.e., deserving a score of 70 or higher). The average intrareader agreement rate was slightly better (75%) but still low enough to indicate that systematic differences in grading standards or criteria between readers were not the only reasons for the disagreements between them in the scores they assigned.

The major implication of this finding was that it would be necessary to have the applicant answer several essay questions and/or have several independent readings per answer in order to balance out the effects of chance factors in the scoring process unless some means were found to increase the consistency with which the grades were assigned.

Scorecard Grading. It was thought that one way of improving the reliability of the grading process would be to have the readers grade the answers in terms of how well they addressed each of the major issues within a question as well as other general features of the applicants' answers, such as the quality of the logical analysis and reasoning that was exhibited. Each of the dimensions that was unique to a given question, as well as those that were common to all questions, was graded separately by an additional reader per question.

The analyses of the data obtained by this process indicated that this scorecard approach did not improve the reliability of the grading process enough to warrant the almost 100% increase in time it took to grade an answer with it. Further, the regular and the scorecard procedures tended to result in the same rank ordering of applicants in terms of the quality of their performance. In short, no new information about an applicant was obtained by means of employing scorecard grading, and it did not prove to be a cost effective means for increasing reader consistency.

Analysis of Artifacts. The foregoing studies also provided an opportunity to examine whether Essay scores may have been influenced by certain systematic factors that were not associated with answer quality. For instance, it was observed that there was no relationship between handwriting clarity and grades, although unedited typewritten answers received somewhat lower grades than handwritten versions of these same answers. Whether an answer was graded toward the beginning, middle, or end of the six-week scoring process did not appear to affect the grades assigned. It was noted, however, that longer answers tended to get somewhat higher grades than shorter ones. While part of this relationship may stem from the more knowledgeable applicant writing more, this does not seem to be the whole story, in that answer length had absolutely no relationship to MBE scores even though the Essay and MBE scores were closely associated with one another.

Reappraisal Process

In the past, the reappraisal process involved having an applicant's answers (and the scores assigned to them by the initial readers) reviewed by two independent reappraisers if that applicant's total score fell between 1140 and 1199 points. If the two reappraisers agreed on what the pass/fail decision should be for an applicant, then that decision stood. If they disagreed, the applicant's set of 12 answers was

evaluated by a third reappraiser so that a final decision could be made.

An analysis of this process indicated that reappraisers assigned somewhat higher total Essay scores than did the regular readers when the reappraisers were aware of the grades assigned by the initial readers. Despite this difference in the level of the overall grades assigned, the reappraisers and the regular readers agreed quite closely in how they assessed the relative performance of the applicants. For instance, better than 85% of the applicants with initial total scores between 1190 and 1199 passed the examination as a result of the reappraisal process, whereas if an applicant's score was below 1190, the applicant had less than an 8% chance of passing. Essentially no applicant during the past two years with an initial total score below 1170 has passed as a result of reappraisal.

CONCLUSIONS

The results of the research studies presented in this report have indicated that California's State Bar Examination is achieving its objective of assessing applicant skills that are considered necessary (albeit not necessarily sufficient) for the practice of law. The examination scores correspond

closely to performance in law school, they are relatively free from possible contaminating influences, and they are sufficiently reliable for making pass/fail decisions about individual applicants. It is apparent, however, that the resources required to maintain these procedures, especially in light of the large number of applicants taking the examination, might be reallocated so as to improve the appropriateness with which the pass/fail decisions are made. For instance, it does not make a great deal of sense to have two reappraisers review the answers of applicants with total scores below 1170 when experience has shown that these applicants have essentially no chance of passing as a result of the reappraisal process. Further, the reappraisers just evaluate the applicants who came close to passing but failed. What about the applicants who came close to failing but passed? Isn't it just as likely that the initial readers inadvertantly passed as many applicants who should have failed as vice versa?

It is also apparent that changes might be made in the initial reading process so as to improve the reliability of the scores obtained and/or decrease the testing burden on the applicant. For instance, an eight-answer test in which each answer is independently evaluated by two readers has about the same reliability as a test in which there is only one reader for each of 12 answers. However, eight answers may not provide a sufficiently broad sample of legal issues to give an accurate and fair assessment of the applicant's skills and knowledge. Moreover, any change

in current practices must be weighed against their costs, impact on the speed of score reporting, logistical constraints, availability of skilled readers, test security, and related concerns.

The foregoing considerations and research results have led the Committee of Bar Examiners to divide the examination's scoring procedures into the following four phases:

1. Applicants will be divided into two groups, "Pass" and "Questionable," on the basis of their total score on the MBE and one of the three Essay test sessions. This will be done by assigning applicants randomly to sessions after the test has been administered so that all 15 reader teams may begin grading simultaneously.

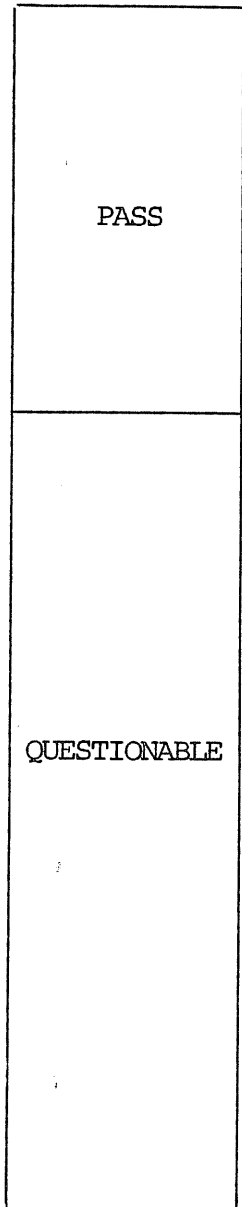
Previous research has indicated that which session is used for a particular applicant will not systematically affect his or her chances of passing in that the average scores across sessions rarely differ by more than a few points. An analysis of past examinations also has shown that a Phase I cutting score of 670 would have resulted in less than one half of one percent of the applicants unduly benefiting from the proposed procedures in the sense that they would have failed if all 12 of their answers had been graded. On the other hand, by using this cutting

PHASE I
Partial
Reading

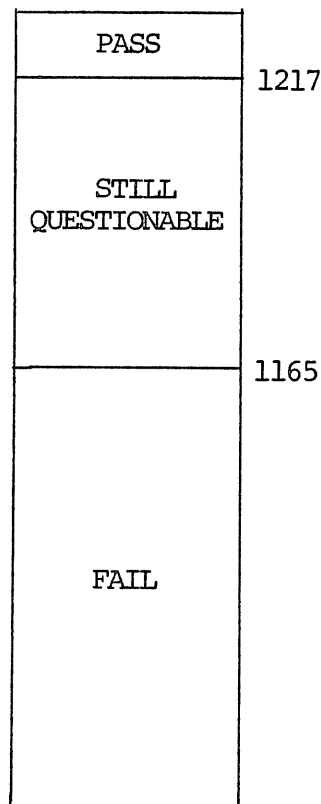
PHASE II
Complete
Reading

PHASE III
Double
Reading

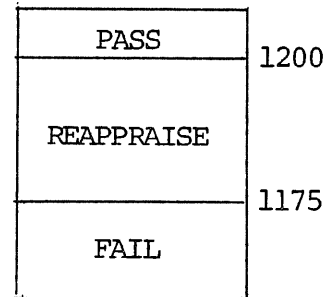
PHASE IV
Reappraisal



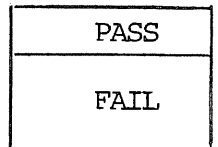
MBE + Score on
one Essay Test
Session (each
of four answers
scored once).



MBE + Score on
12 answers each
of which is
scored once.



MBE + Score on
12 answers each
of which is
scored twice.



MBE + all 12
answers and
scores are
reviewed as
a set by one
reappraiser.

Illustration of how the Pass, Fail, and Reappraisal Decisions
Would Have Been Made if the Proposed Procedures had Been Used
With the July, 1977 Examination.

score, over 35% of the applicants who take the examination are passed with just one reading of each of four answers. This constitutes a savings of about 21,000 readings if 7,500 applicants take the examination (i.e., $.35 \times 8 \times 7,500 = 21,000$).

2. In Phase II of the new procedure, all the applicants who were classified as "Questionables" have their remaining eight answers read. Their total score (MBE plus Essay) is then used to place each of these applicants into one of three categories: "Pass," "Still Questionable," and "Fail." The cutting scores for the "Still Questionable" category, 1165 to 1217, represent 68 to 78% of the total possible range. It also will be noted that the 1164 fail line is well below the score level from which applicants were likely to pass as a result of reappraisal.
3. All the applicants who were classified as "Still Questionable" in Phase II would have their 12 answers reinserted into the regular reading process in Phase III. In other words, all the applicants in this group (about 23% of the total taking the test) would have their 12 answers read twice. The average of these two independent sets of readings will then be combined with the MBE score to make a pass/fail or reappraisal decision for

each applicant. While this process essentially offsets the savings gained in Phase II in terms of the number of answers that have to be read (i.e., $.23 \times 12 \times 7,500 = 20,700$), it does increase the reliability with which the pass/fail decision is made for those applicants whose status is most in doubt. For instance, an analysis of the data collected on the February 1977 examination indicated that doubling the number of readers per answer would have had the same effect on the Essay Test's reliability as increasing the length of the Essay Test to 20 answers per applicant.

4. Applicants who at the end of Phase III had scores between 1175 and 1199 will be placed in reappraisal. Also placed in reappraisal will be those applicants whose total score on the first reading was between 1200 and 1217 but who dropped below 1175 as a result of the second reading. Each reappraised applicant will have his or her full set of answers and scores reviewed by one reappraiser, who will make the final pass/fail decision. Thus, although this process substantially reduces the number of booklets that have to be evaluated by the Board of Reappraisers, it still focuses on those applicants whose passing status might be affected by reappraisal, as well as preserves the procedure of having an applicant's efforts reviewed as a whole.

The foregoing procedures differ from past practices primarily in terms of focusing reader resources at the pass/fail decision point rather than spreading them out evenly across all the applicants taking the examination. On the other hand, the new procedures are still designed to provide added protection against the chances of failing an applicant who should pass. Although it is unlikely that these procedures will result in any major change in the total proportion passing, they will increase the accuracy with which the pass/fail decisions are made. Moreover, the total number of answers reappraised will be somewhat less than in the past, which will hopefully expedite the score reporting process. This savings in reappraiser time will most likely be channeled into greater use of their skills and knowledge in the test development and scorer calibration aspects of the examination process.

Finally, the changes described above are based on two years of research and analysis of bar examination practices. And, they do not encompass all the ways in which the bar examinations's procedures have been modified. For instance, as a result of the findings on Essay test grading practices, the Committee of Bar Examiners has revised its reader calibration procedures so as to devote more effort to establishing and maintaining a high level of scorer reliability. Additional research also is being conducted on the relationships between test performance and an applicant's racial/ethnic group, sex, and law school grades.

Other research studies, in progress and in the planning stages, are oriented toward gathering data regarding ways in which various aspects of the examination process might be improved, such as the nature and format of the questions asked. In short, the changes in grading procedures presented in this report are only a few of the outcomes of the Committee's directive for continued research and improvement of bar examination practices.

BASIC TEST STATISTICS *

MEASURE	TEST STATISTIC	February 1976	July 1976	February 1977	July 1977
Essay	Mean	809.71	828.46	815.53	826.44
	Standard Deviation	58.35	68.79	58.96	62.12
	Standard Error	27.37	27.52	28.28	26.34
	Reliability	.78	.84	.77	.82
	Percent Passing	30.80	47.60	36.40	44.20
MBE	Mean	351.52	373.32	356.42	367.84
	Standard Deviation	36.93	40.59	35.99	38.65
	Standard Error	11.68	12.83	11.38	12.22
	Reliability	.90	.90	.90	.91
	Percent Passing	43.20	66.40	46.60	60.00
Total Score	Mean	1161.23	1201.78	1171.94	1194.27
	Standard Deviation	86.37	101.36	86.86	93.34
	Standard Error	29.92	30.41	30.09	29.52
	Reliability	.88	.91	.88	.90
	Percent Passing	37.70	54.30	38.20	49.50
MBE/Essay Correlation		.62	.70	.65	.70
Total Number of Applicants		3089	6710	3399	7191

* Passing scores on the Essay, MBE, and Total examination were 840, 360, and 1200, respectively. Reappraisal results were not included in calculating percent passing.

The Essay test's reliability is based on the average interitem correlation stepped-up to a 12 item test.

All test statistics are based on those applicants who answered 12 essay questions and who had an MBE score.

AVERAGE ESSAY SCORES FOR GIVEN LEVELS OF MBE SCORES

MBE SCORE RANGE	Febr 1975	July 1975	Febr 1976	July 1976	Febr 1977	July 1977	Average Score
Less than 351	777	766	777	767	780	777	774
351 to 375	820	811	818	819	824	820	819
376 to 395	838	836	838	842	842	847	841
396 to 415	861	863	857	867	866	869	864
More than 415	900	888	889	892	894	893	893

NUMBER OF APPLICANTS REAPPRAISED AND PERCENT PASSING
RELATIVE TO INITIAL TOTAL SCORE: 1976 THROUGH 1977

Initial Total Score Range	Percent Passing				Number of Applicants
	Febr 1976	July 1976	Febr 1977	July 1977	
1190-1199	86	88	85	89	821
1180-1189	32	43	34	42	860
1170-1179	5	10	5	6	823
1160-1169	0	*	0	0	814
1150-1159	0	*	0	0	812
1140-1149	0	*	0	0	795
Number of Applicants	914	1332	991	1688	4925
Percent Passing	19	27	20	23	23

* In July of 1976, there was a problem in the typing room during the first Essay test session in Los Angeles. As a result, the scores of certain applicants on this part of the examination did not reflect accurately their true performance level. This situation led to a special reappraisal of the applicants involved if they failed the examination which in turn led to passing 11 applicants whose initial total score fell below 1170. In July of 1977, one applicant with an initial total score between 1165 and 1169 passed as a result of reappraisal. In no other instance did an applicant with an initial total score below 1170 pass as a result of reappraisal on these four examinations.

AVERAGE TOTAL ESSAY SCORE BY TEST SESSION

Essay Test Session	Febr 1976	July 1976	Febr 1977	July 1977	Overall Average
First	264	275	271	274	271
Second	268	277	270	272	272
Third	277	277	274	280	277
Average	270	276	272	275	273

EFFECT OF VARIOUS PHASE I CUTTING SCORES ON PASS/FAIL DECISIONS

Possible Cutting Score	% Passing At End of Phase I				% Passing At End of Phase I Who Failed the Total Examination			
	Febr 1976	July 1976	Febr 1977	July 1977	Febr 1976	July 1976	Febr 1977	July 1977
660	22.3	46.3	27.0	40.5	0.8	0.8	0.6	0.7
665	19.4	43.0	23.7	37.6	0.4	0.5	0.3	0.5
670	17.0	39.7	20.8	34.5	0.3	0.2	0.1	0.3

* Based on final Pass/Fail decision; i.e., it includes reappraisal results.

NUMBER AND PERCENT OF APPLICANTS IN THE "STILL QUESTIONABLE"
CATEGORY AT THE END OF PHASE II OF THE PROPOSED PROCEDURES *

	Febr 1976	July 1976	Febr 1977	July 1977	Total
Number of Applicants in the Category	767	1306	866	1521	4460
Percent of all Applicants	25	20	26	21	22

* The "still questionable" category includes those applicants who had initial total scores between 1165 and 1217 after a single reading of all 12 of their Essay answers.