80-2F

INTRA- AND INTER-READER AGREEMENT ON THE ESSAY SECTION OF THE CALIFORNIA STATE BAR EXAMINATION.

A Report Submitted to the
Committee of Bar Examiners
of the State Bar of California

Prepared by
Stephen P. Klein, Ph.D.

June 10, 1980

80-2P

INTRA- AND INTER-READER AGREEMENT ON THE ESSAY SECTION OF THE CALIFORNIA STATE BAR EXAMINATION.

PURPOSE

This study was conducted to determine the degree to which readers agreed with themselves and each other in the grades they assigned to essay anwers on the California State Bar Examination. This study also investigated whether there was any systematic relationship between the number of readers assigned to a question and the degree of agreement between these readers.

PROCEDURES

The essay section on the July, 1979 California State Bar Examination had three sessions. In each session, an applicant was required to answer three of the four questions presented.

Each reader was assigned to grade the answers to one and only one question. Because of the large number of applicants taking the examination, each question had 7 to 15 readers (depending upon the number of applicants who answered it). A total of 145 readers participated in the grading process.

The score assigned to an answer could theoretically range from 0 to 100 points in five point increments. A score of 70 or higher was considered as "passing" the question.

Five answers per question were selected from among the first batch of answers that were graded on each question. The grades assigned to these answers were: 60, 65, 70, 75, and 80. Copies of these answers were then distributed among the remaining set of answers each reader was supposed to grade; i.e., all the readers assigned to a given question independently read the same set of five answers. The answers in this common set were disguised so that the readers could not distinguish them from the other answers they graded or know the scores other readers assigned to them.

RESULTS

The results of this study are presented in Table 1. This table is divided into three sections: inter-reader agreement, intra-reader agreement, and agreement between the average and actual grade assigned.

An inspection of the inter-reader data indicated the following:

- The readers agreed with each other about 82 percent of the time as to the pass/fail status of any given answer; i.e., whether or not it deserved a score of 70 or higher.

- On the average, the grade assigned by one reader to an answer deviated 4 points from the grade assigned to that same answer by another reader. The largest deviation between any two readers of the same answer was 30 points.

Table 1

INDICIES OF INTER- AND INTRA-READER AGREEMENT ON THE JULY, 1979 EXAMINATION.

		Inter-Reader Agreement				Intra-Reader Agreement		Average vs. Actual Score	
Quest.	No. of Readers	Pass Fail	Mean Dev.	Stan Dev.	Corr Coef	Pass Fail	Mean Dev.	Pass Fail	Mean Dev.
1	10	88.0	3.7*	4.9	.35	80	6.0	80	6.1
2	15	90.7	3.5*	5.2	.61	100	4.0	100	3.7
3	13	76.9	4.3	5.1	.43	80	4.0	60	4.2
4	10	72.0	4.0	5.4	.26	80	4.0	60	4.2
5	7	82.8	5.4	6.6	.52	80	5.0	80	5.1
6	15	78.7	4.5*	6.0	.43	100	4.0	40	5.6
7	15	82.7	3.8*	5.3	.45	100	5.0	60	4.6
8	11	83.6	3.6*	5.0	.48	80	7.0	60	5.5
9	15	77.5	5.2*	6.6	.51	80	5.0	80	2.9
10	7	88.6	2.4	3.2	.70	60	5.0	100	4.4
11	15	82.7	3.6	5.1	.46	80	5.0	80	3.1
12	12	76.7	3.4*	4.3	.51	80	6.0	80	4.1
Average	12	81.7	4.0	5.2	.48	83	5.3	77	4.5

Pass/Fail = This the average percent of applicants whose pass/fail status was classified the same way by two readers.

Mean Dev. = The average number of points that the grade assigned to an answer by one reader deviated from the grade assigned to that same answer by another reader (in the case of intra-reader agreement, this is the average difference in the grade assigned to the same answer by the same reader on two different occasions).

The differences in average scores between readers on the questions marked by an asterisk (*) were not random; i.e., one or more readers was systematically more or less lenient than one or more other readers. The probability that a difference marked by an asterisk was due to chance is less than 1 in 100.

- About two thirds of the scores assigned to a given answer fell within 5.2 points of the average score assigned to that answer.

- The grades assigned by one reader to a given set of five answers correlated about .48 with the grades assigned by another reader to these same answers. In other words, there was a moderate positive relationship in how any two readers rank ordered the answers.

- There did not appear to be any systematic relationship between the number of readers assigned to a question and the degree of agreement between them. For example, even though question #5 had the highest average deviation (5.4) and question #10 had the lowest average deviation, they both had the same number of readers.

- On seven of the 12 questions, there were statistically significant differences between readers in the scores they assigned; i.e., one or more readers was systematically more or less lenient than one or more of the other readers.

The second section of Table 1 provides information about the degree of intra-reader agreement. This information was gathered when a reader encountered in the common set one of the answers that reader had graded previously. The results of this analysis indicated the following:

- A reader agreed with himself (or herself) about 83.3 percent of the time on the pass/fail status of an answer.

- The score assigned by a reader on one ocassion usually deviated about 5 points from the score that same reader assigned to this answer on another ocassion.

The final section of Table 1 provides information with respect to the degree of agreement between the average grade assigned to an answer (i.e., across all the readers who scored it) and the actual grade on that answer (i.e., the one that was used in computing the applicant's pass/fail status). These data indicate the following:

- The average percent agreement on the pass/fail status of an answer across all 60 answers was 76.7 percent. It was noted, however, that this overall rate was markedly affected by the unusually low degree of agreement on question #6.

- The average degree of discrepancy between the actual and average score on an answer was 4.5 points; i.e., the actual grade usually deviated less than five points from the best estimate of that answer's quality.

CONCLUSIONS

This study investigated the extent to which readers agreed with themselves and each other in how they graded the answers on the essay portion of the California State Bar Examination. One major finding of this research was that readers agreed with themselves to about the same degree as they agreed with each other in terms of both the average score assigned to an answer and their evaluations of the relative quality of the answers. For example, the average agreement rate on the pass/fail decision was 83 percent within readers and 82 percent between readers. And, the usual discrepancy between two readers on an answer (or between two grades assigned by the same reader) was only four to five points. It was further observed that the degree of agreement on a question was not systematically related to the number of readers assigned to grade the answers to it.

In a previous study of essay grading practices on the California examination (Klein, 1977), it was found that two readers agreed with one other about 67 percent of the time on the pass/fail status of an answer as compared to the 82 percent in the present study. Thus, there has been a substantial increase in the overall consistency with which grades are assigned. This improvement is most probably attributable to the extra effort that is now taken to establish and maintain reader agreement through the use of benchmark answers, periodic calibration sessions, etc.

It should be noted, however, that there is still not perfect agreement between (or even within) readers in the grades they assign. Thus, on any given question, an applicant might receive a slightly higher or lower score than the applicant deserves. Nevertheless, such discrepancies are usually quite small and are likely to get balanced out since an applicant's essay score is based on the sum of the scores on nine questions.

As a result of this balancing process, the total essay score provides a reasonably reliable estimate of an applicant's true performance level. This is especially so for those applicants who fall near the pass/fail line since their answers are read twice by different sets of readers. In other words, the double reading essentially eliminates the effect of an applicant having drawn a particular group of readers for the first grading of his/her answers. Moreover, a failing applicant, who after the second reading is still close to passing, has his/her answers read again (as a set) by a member of the Board of Reappraisors.

Finally, the overall pass/fail decision on an applicant is made on the basis of a combined score (i.e., Essay plus Multistate Bar Examination). Data from the July, 1979 examination (Klein, 1980) indicates that this total score is sufficiently reliable for making pass/fail decisions about individual applicants. Thus, although two readers may not always agree on the score that should be assigned to a given answer, there are enough answers per applicant to be graded, enough rereadings of answers, and enough additional information (through the MBE) about an applicant's legal skills and knowledge to be certain that the final total score is an accurate indicator of an applicant's overall performance level.

REFERENCES

Klein, Stephen P. An Analysis of Grading Practices on the California Bar Examination. A report submitted to the Committee of Bar Examiners of the State Bar of California, December 15, 1977.

Klein, Stephen P. A Comparison of the Effectiveness of a Single Versus a Multiphased Grading System. A report submitted to the Committee of Bar Examiners of the State Bar of California, May 14, 1980.