PR-85-4

LARGE DISCREPANCIES BETWEEN READERS: THEIR SOURCE AND POLICY IMPLICATIONS

Stephen P. Klein, Ph.D.
GANSK & ASSOCIATES

September 25, 1985

SUMMARY

Applicants taking California's General Bar Examination (GBX) have all of their Essay and Performance Test (PT) answers read a second time if after the first reading, their total GBX scores come close to the pass/fail line. The first and second readings are done by different graders. The second reader does not know the score assigned by the first reader. Under the current rules, an applicant's final score on an answer that is read twice is the average of these two scores.

Analyses of past examinations indicate there is generally a very high degree of agreement between the scores two readers independently assign to an answer. The difference is usually no more than 10 points on 95 percent of the answers that are read twice. The study described in this report focused on the remaining 5 percent. It investigated whether large (15 points or more) discrepancies between the scores different readers assign to an answer are most likely due to chance variation around the score the answer really deserves (i.e. its true score) or to one of the two readers making a mistake (such as the reader making a clerical error in recording the score).

This issue was explored by having a third and fourth independent reading of 269 answers. These answers were drawn from those written to three essay questions and one PT problem on the February 1985 exam. On all 269 answers, there was a large (15 point or more) discrepancy between the scores assigned by the first and second reader.

A "difference score" was computed for each answer. This score was the difference between the mean of the scores assigned to the answer on the first two readings and the mean assigned to it on the next two readings. For example, the difference score was -5 if the mean of the first and second readers' scores was 65 and the mean of the third and fourth readers' scores was 70.

Difference Score = (Mean of 1st and 2nd) - (Mean on 3rd and 4th)

The "absolute value" of a difference score is that score without its algebraic (+ or -) sign. The "average absolute value" is the average difference between the two means regardless of which one was larger.

If large discrepancies between the first two readers are due to chance variation, then: (1) the average of the third and fourth reader's scores will fall directly between the scores assigned by the first two readers; (2) the distribution of the difference scores will take the form of a normal, bell-shaped curve; (3) the average absolute value of the difference scores will be small; and (4) the mean of the first two scores assigned will be the best estimate of the answer's quality.

If large discrepancies between the first two readers are due to one of them making a mistake, then: (1) a third and fourth reader's grades on an answer will agree more with one of the first two reader's scores than with the average of the first two reader's scores, (2) the distribution of the difference scores will take the form of a U-shaped curve, (3) the average absolute value of the difference scores will be large, and (4) the best estimate of an answer's quality would not necessarily be the average of the first two scores assigned to it.

Results from analyses of the four readings of each of the 269 answers strongly support the chance variation hypothesis. They show that:

1) The scores assigned by the third and fourth reader to an answer were far more likely to agree with the mean of the first two readers' grades than they were with the grade assigned by just one of the first two readers.

2) There was a bell-shaped (rather than U-shaped) distribution of the difference scores. On 51 percent of the answers, there was a 2.5 or smaller difference between the means. Only 14 percent of the answers had an absolute difference score that was greater than 7.5 points.

3) There was a strong correlation ($r = .72$) between how the first two and the last two readers rank ordered the 269 answers, especially given the large differences between the first two readers in the scores they assigned to these answers.

4) The average absolute difference on an answer was 4.5 points, which is about one half the size it would have been if the third and fourth readers had tended to agree with one of the first two readers rather than with the mean of the first two readers' scores.

In addition, the third and fourth readers' scores were not significantly higher or lower than the mean of the first two readers' score. Thus, even if some of the discrepancies were due to one of the first two readers making a mistake, such errors did not systematically increase or decrease applicant scores. Moreover, all of the findings above were consistent across the three essay questions and one PT problem studied. They also held when the discrepancy between the first two readers was 15 points and when it was more than 15 points.

Taken together, these findings demonstrate that large and even very large discrepancies between the first two readers of an answer stem from chance variation around that answer's true score. There is no sign that the discrepancies come from one of the first two readers making a clear mistake. Thus, even when there is a large discrepancy between the grades assigned by different readers to an answer, the best estimate of the score that answer really deserves is still the mean of the grades assigned to it.

LARGE DISCREPANCIES BETWEEN READERS: THEIR SOURCE AND POLICY IMPLICATIONS

INTRODUCTION

About 25 to 30 percent of the applicants who take California's General Bar Examination (GBX) have all of their Essay and Performance Test (PT) answers read twice. The applicants in this group came close to the pass/fail line of 1260 after the scores on the first reading of all of their answers were combined with their scores on the MBE (the multiple choice portion of exam). For example, on the February 1985 exam, applicants had their answers read twice if their total scores (Essay + PT + MBE) fell between 1225 and 1278 after the first reading of their Essay and PT answers.

The second reader knows the answer has been read once, but does not know the score assigned to it by the first reader. Under the current rules, readers assign grades in 5-point intervals and the final score on an answer that is read twice is the average of the two grades. For example, the final score on an answer would be 67.5 if one reader gave it a 65 and another gave it a 70.

The assignment of the second reader to an answer is not random. If the first reader tended to grade high or low (as indicated by the scores that reader assigned during the calibration sessions), then a reader with the opposite tendency was appointed as the second grader. This procedure increases the fairness of the scoring process. However, it also increases the likelihood of obtaining large differences between the first and second reader's grades.

Analyses of past exams indicate that even with the bias introduced by the procedures described above, the two readers on an answer generally agree closely on the score they assign to it. For example, the score assigned by one reader differed by no more than 10 points from the score assigned by the other reader on over 95 percent of the answers that were read twice on the July 1983, February 1984, and July 1984 exams. On the average, the scores assigned by two readers to an answer differ by 4.5 points (this statistic is called the "average absolute difference"). A study with a random sample of 865 applicants who took the July 1983 exam also showed that the total essay score on the first reading correlated 0.91 with the total essay score on the second reading.

Despite the generally high degree of agreement between readers, there are large differences (15 points or more) on about 3 to 5 percent of answers that are read twice. It is not known whether these large discrepancies represent chance variation around the true score the answer deserves or whether they stem from one of the two readers making a mistake that unduly lowers or raises the assigned score (such as the reader recording the wrong grade, skipping a page in the answer book where an important issue was discussed, or giving an applicant credit for discussing an issue that was not really covered by the answer).

POLICY IMPLICATIONS

If large discrepancies between the scores different readers assign to an answer stem from chance variation around the score that answer really deserves, then the Committee's policy of using the average of the two scores makes the most sense. Chance errors tend to cancel each other out. And when this happens, the average score of the first two readings is more likely to come closer to an answer's "true" score (i.e., the score it really deserves and would receive if there were no chance errors) than would either one of the first two scores by itself.

If large discrepancies between the first two readers are most likely due to one of them making a mistake, then the true score would correspond closer to one of the two grades assigned rather than to the average of these grades. And, if large discrepancies usually stem from mistakes, the Committee might consider adopting special procedures for handling these discrepancies, such as having a third reader resolve them.

PURPOSE

The study described below investigated whether large (15 point or more) discrepancies between two readers in the scores they assign to an answer are most likely due to chance variation around an answer's true score or to a mistake made by one of the two readers.

RATIONALE FOR STUDY DESIGN

If the difference between two scores is just a chance error around its true score, then the best estimate of the true score is the mean of the first two scores. In other words, if the answer was read again and again, then the third and subsequent scores assigned to it would tend to fall exactly half-way between the first two scores. If this happened, (1) the difference between the mean of the first two scores and the mean of all subsequent scores would be small and (2) the distribution of these difference scores would form a normal, bell-shaped curve.

If, on other hand, large discrepancies stem from one of the first two readers making a mistake, then the mean of the scores on subsequent readings of an answer would fall close to one of the first two scores assigned to it rather than half-way between these scores. And if this happened, (1) there would be large positive and negative differences between the average of the first two scores and the average of any subsequent scores and (2) the distribution of these difference scores would form a U-shaped rather than a bell-shaped curve.

In this context, a "difference score" between the average of the first two readings of Answer X and the next two readings of Answer X is defined by the formula below. Difference scores also can be computed for the difference between the average of the first two readings and the score on the third (or fourth) reading by itself.

Difference Score = (Mean of 1st and 2nd readings of Answer X) − (Mean of 3rd and 4th readings on Answer X)
on Answer X

The study described below examined what happened when the score on a third and/or fourth independent reading of a answer was subtracted from the average of the scores on the first two readings. Would the resulting difference scores tend to be small and distributed normally in a bell-shaped curve as would be expected by the "chance variation" hypothesis OR would the difference scores be large and distributed in a U-shaped curve as would be expected by the "reader mistake" hypothesis?

PROCEDURES

The scores on the answers to three essay questions (2, 4, and 5) and one PT problem (Question 7) on the February 1985 exam were searched in order to identify all them that were read twice and had a "large" (15 points or more) discrepancy between the grades assigned by the first and second reader. This selection criterion identified 269 answers.

Each selected answer was read two more times. The third and fourth readers on an answer did not know each other's scores, the scores assigned by the first two readers, or that their grades would be compared to those assigned by other readers. The third and fourth readers were drawn from the same pool of readers that conducted the first and second readings. However, neither the third or fourth reader on an answer was one of that answer's original two readers.

RESULTS

Table 1 lists the number of answers with large discrepancies on each question and overall. It also presents the mean score on: each reading, the mean of first two readings, the mean of the last two readings, and the difference between the means of the first and last two readings. These data indicate that the third and fourth readers' average score (68.3) was almost the same as the first two readers' mean (67.7). Thus, differences between the score on the third (or fourth) reading of an answer and the average of the scores on the first two readings of that answer cannot be due to the third (or fourth) reader being generally more or less lenient than the first two readers.

Table 2 shows there was a very low and usually negative correlation between the scores assigned by the first two readers in the set of 269 answers. This is to be expected given the way these answers were selected; i.e., they were the ones on which the first two readers were

known to disagree. Thus, the correlations between the first and second readings do not reflect the typical average correlations (of .64 to .77) between two readers on the thousands of answers that are read twice.

Table 2 also shows that the mean of the first two readers' scores on an answer correlated .72 with the mean of the scores assigned by the next two readers. In other words, both sets of readings rank ordered the answers in about the same way, even though the first two readers differed markedly with each other in the scores they assigned to each answer. In addition, the correlation between the two means was higher than the .52 correlation between the grades assigned by the third and fourth reader.

Table 1

MEAN SCORES ON EACH READING AND EACH PAIR OF READINGS

Question Number	Number of Answers	Mean on Each Reading				Mean of 1st and 2nd	Mean of 3rd and 4th	Mean of 1st & 2nd minus 3rd & 4th
		1st	2nd	3rd	4th			
2	104	70.4	64.9	68.6	69.4	67.6	69.0	-1.35
4	46	73.6	67.0	68.5	69.9	70.3	69.2	1.09
5	78	68.8	62.9	66.8	66.3	65.9	66.6	-0.71
7	41	68.0	68.9	68.2	68.9	68.5	68.5	-0.06
Combined	269	70.1	65.3	68.2	68.3	67.7	68.3	-0.55

Table 2

CORRELATIONS BETWEEN READINGS

Question Number	1st versus 2nd	3rd versus 4th	1st versus 3rd + 4th	2nd versus 3rd + 4th	1st + 2nd versus 3rd + 4th
2	.06	.56	.49	.54	.70
4	-.30	.40	.59	.39	.84
5	-.09	.52	.36	.61	.73
7	-.17	.36	.35	.47	.64
Combined	-.05	.52	.45	.53	.72

Table 3 shows that on 51 percent of all of the 269 answers, the average score assigned by the first two readers was within 2.5 points of the average of the third and fourth reader's scores on that answer. This table also shows that on every question, the difference scores had a relatively normal, bell-shaped distribution. In other words, most of the difference scores clustered around 0 and larger differences tapered off quickly and symmetrically from 0.

The foregoing pattern indicates that on a typical answer in the set of 269 answers, the grades assigned to it by both the third and fourth readers agreed far more with the average of the first two readers' grades than they did with the grade assigned by just one of the first two readers. In addition, even when there was a large difference between the first and last two readers, the last two readers did not give consistently higher or lower grades than the first two readers. Thus, even if a few of the large differences were due to one of the first two readers making a mistake, such errors increased applicant scores about as often as they decreased scores.

An analysis of the absolute difference scores revealed that, on the average, the mean of the first two scores on an answer differed by only 4.49 points from the mean of the second two scores on that answer (i.e., regardless of the algebraic sign of that difference). This average absolute difference is considered "small" because it is no larger than the average absolute difference that is usually obtained between a pair of readers on all of the thousands of answers that are graded twice.

If the third and/or fourth reader had tended to agree with just one of the first two readers on an answer (as would have been predicted by the "mistake" hypothesis), then the distribution of difference scores would have been U-shaped rather than the bell-shaped curve that was observed with all four of the questions studied. The "mistake" hypothesis also would have predicted an average absolute difference score that was almost twice as large as the 4.49 that was actually obtained.

The mistake hypothesis predicts that the average of the third and fourth readers' scores would coincide more closely with one of the first two readers' scores than with the mean of their scores. For example, if one of the first two readers gave it a 50 and the other a 65, then the mean of the third and fourth readings would be closer to 50 or 65 than it would be to 57.5. Thus, if there was a 15 point difference between the first two readers, the expected value of the absolute difference between their mean score and the mean of the third and fourth readers' scores would be 7.5. The expected average absolute difference score in the set of 269 answers under the mistake hypothesis would be 8.47 as per the following formula: $[(192)(7.5) + (54)(10) + (19)(12.5) + (4)(15)]/269$

Table 3

NUMBER OF ANSWERS ON EACH QUESTION WITH DIFFERENT SIZED DIFFERENCES BETWEEN THE AVERAGE OF THE SCORE ON THE FIRST TWO READINGS AND THE SCORE ON THE THIRD AND/OR FOURTH READING.

Question Number	Reading Number	Difference Between the Mean Score on the 1st and 2nd Reading of an Answer and the Score on the 3rd and/or 4th Reading of That Answer: <-7.5	-7.5 to -2.6	-2.5 to +2.5	+2.6 to +7.5	>7.5	Number of Answers Read
2	3	12	25	38	18	11	104
2	4	20	23	28	22	11	104
4	3	0	3	39	1	3	46
4	4	1	3	31	8	3	46
5	3	6	21	32	15	4	78
5	4	7	8	49	9	5	78
7	3	3	3	27	5	3	41
7	4	6	6	18	7	4	41
Total	3 or 4	55	92	262	85	44	538
Percent	3 or 4	10	17	49	16	8	100%
Mean of the 3rd and 4th	Number	25	49	138	43	14	269
	Percent	9	18	51	16	5	100%

The row labelled "Question 2, Reading 3" refers to the scores assigned by the third reader to the 104 answers to Question 2. In this row, the value of 12 in the "< -7.5" column indicates there were 12 answers where the average of the first and second readers' scores on these answers was more than 7.5 points lower than the score assigned by the third reader. The value of 11 in this same row indicates there were 11 answers where the mean of the scores assigned by the first two readers was more than 7.5 points higher than the score assigned the third reader.

Table 4 presents the distribution of difference scores relative to the size of the initial discrepancy between readers. These data show that regardless of the size of the discrepancy between the first two readers, the distribution of the differences between their average score on an answer and the average of the third and fourth readers' scores on that answer follow a normal, bell-shaped curve. There are slightly more cases in the categories corresponding to large negative difference than large positive differences because the mean of the third and fourth readers' grades was slightly larger than the mean of the first and second readers' grades (see Table 1).

Table 4

NUMBER OF ANSWERS WITH DIFFERENT SIZED DIFFERENCES BETWEEN THE MEAN OF THE SCORES ON THE FIRST AND SECOND READINGS AND THE MEAN OF THE SCORES ON THE THIRD OR FOURTH READING RELATIVE TO THE SIZE OF THE DISCREPANCY BETWEEN THE FIRST TWO READERS' SCORES.

Size of the Discrepancy Between the Scores on the 1st and 2nd Reading	Number of Answers	Mean of the Scores on the 1st and 2nd Reading of an Answer Minus the Mean of the Scores on the 3rd and 4th Reading of that Same Answer				
		<-7.5	-7.5 to -2.6	-2.5 to +2.5	+2.6 to +7.5	>7.5
15 points	192	17	30	108	27	10
20 points	54	5	14	21	11	3
>20 points	23	3	5	9	5	1
Combined	269	25	49	138	43	14

CONCLUSION

Large discrepancies (15 points or more) between the scores assigned by the first two readers of an answer stem from chance variation around that answer's true score. There is no sign that large or even very large discrepancies come from one of the first two readers making a clear mistake. Thus, the best estimate of an answer's true score is the mean of the grades assigned to it. Having a third or fourth reading would not produce a significant improvement in accuracy nor would it tend to generally raise or lower grades.