IS THERE EXAMINER BIAS ON THE WECHSLER-BELLEVUE?

EDWIN COHEN,* Vet. Adm. Mental Hygiene Clinic, Durham, N. C.

It seems to be generally assumed that results of psychometric instruments, such as the Stanford-Binet and the Wechsler-Bellevue, are not affected by the examiner, if he has attained a minimal level of competence. On the other hand, many students of projective techniques feel that somewhat different results are obtained from the same subject by different examiners. Thus Klopfer, in listing prerequisites for objectivity of psychological procedures, states," . . . *most* experimenters with the necessary skill and experience will arrive at the same *or similar* results in using the procedure." (italics mine) (6, p. 15) Bell states that, "Even these who are expert with the (Rorschach) method will not secure identical personality pictures. . . ." (1, p. 492).

Miller, Sanders, and Cleveland found a definite relationship between examiner personality and obtained Rorschach protocols. (4)

While a psychometric instrument will undobutedly be less affected by examiner influence or bias than a projective technique, there seems to be a distinct possibility that different competent examiners may affect the anxiety or motivation of the subject differently, producing changes in test results. For example, if an examiner arouses anxiety in a subject by his attitude, mood, comments, etc. the subject might do more poorly on digit span, even though the mechanics of administration are flawless.

This study aims to determine whether this examiner bias is evidenced in scores on a psychometric instrument such as the Wechsler-Bellevue. The specific problem can be phrased, "Do different examiners tend to get different subtest scores, on the average, for certain subtests, from their colleagues?"

*Now at the University of Oklahoma. The writer is indebted to Dr. Burke M. Smith for providing the raw data, and to Messrs. William Michaux and John McMillan for their helpful suggestions.

TABLE I

Corrected Average Per Cent Subtest Contributions.

EXAMINER	NUMBER OF TESTS	SUBTEST: INFORMATION	COMPREHENSION	DIGIT SPAN	ARITHMETIC	SIMILARITIES	VOCABULARY	PICTURE ARRANGEMENT	PICTURE COMPLETION	BLOCK DESIGN	OBJECT ASSEMBLY	DIGIT SYMBOL	EXAMINER MEAN	EXAMINER S. D.
A	17	10.66	11.02	9.56	13.18	8.55	10.04	7.73	11.04	10.49	8.88	8.89	10.00	1.51
B	23	10.31	11.47	8.51	9.13	10.48	10.14	10.37	10.60	10.03	10.80	8.31	"	0.97
C	25	9.98	10.58	10.54	9.44	9.53	10.04	10.19	9.86	10.83	10.22	8.83	"	0.58
D	20	10.74	10.61	10.02	9.72	9.27	9.60	10.81	9.36	11.56	8.96	8.93	"	0.85
E	25	11.03	12.35	9.09	10.38	9.96	9.75	9.49	9.18	9.65	10.18	8.69	"	1.02
F	20	10.09	10.60	10.67	10.43	9.40	9.91	10.09	8.44	11.22	10.59	8.47	"	0.89
G	29	11.70	11.95	9.44	10.07	10.29	10.29	8.97	10.63	9.17	8.95	8.82	"	1.08
H	20	10.39	11.26	9.23	8.79	10.39	9.68*	10.00	9.69	10.55	10.36	9.35	"	0.70
J	35	10.79	10.03	9.04	10.17	9.23	9.91	10.33	9.75	11.96	9.09	9.66	"	0.84
K	21	11.55	10.38	9.02	9.20	9.71	10.89	9.57	10.35	12.34	8.00	9.87	"	1.17
L	25	11.42	11.27	10.47	11.41	8.60	9.84	9.69	9.10	10.63	8.73	8.65	"	1.13
M	30	10.96	11.41	10.70	8.78	9.67	10.07	10.03	10.23	10.34	9.97	7.92	"	0.98
N	35	11.10	11.39	10.48	10.51	10.57	10.53	8.62	10.54	10.25	8.18	8.33	"	1.12
SUBTEST MEAN		10.82	11.10	9.75	10.09	9.67	10.05	9.68	9.91	10.69	9.45	8.82		
SUBTEST S. D.		0.52	0.63	0.73	1.15	0.64	0.34	0.80	0.72	0.86	0.90	0.53		
S. D. RANK		2	4	7	11	5	1	8	6	9	10	3		
VALIDITY		.667	.661	.509	.625	.727	.85	.514	.605	.714	.409	.673		
VALIDITY RANK		5	6	10	7	2	1	9	8	3	11	4		

At least seventeen test protocols from each of thirteen examiners[1], were used in this study. The scoring was done under supervision, which would tend to diminsh examiner differences in scoring.

The mean of the weighted scores for each subtest was computed for the tests administered by each examiner. Each mean was then subjected to the following operations:

1. It was multiplied by 100/average-total-weighted-score of the particular examiner. This puts all figures on the basis of per cent average subtest contribution to the total weighted score. This operation was necessary because the changing type of patient load at the Clinic, combined with the urnover of trainees, produced a disparity among the mean IQ's obtained by different examiners. In this study we are interested in the relative difference among *subtests*, rather than the relation of examiner to obtained IQ; this *step* has the effect of equalizing total weighted scores.

[1]The examiners were Clinical Psychology Trainees at the Durham, N. C. Veterans Administration Mental Hygiene Clinic, at which all these tests were administered.

2. The percent average subtest contribution was in turn corrected (for four of the subtests) for the difference in the expected contribution of a particular subtest toward the total score at different intelligence levels. For example, the Object Assembly subtest was found, in a reworking of previous data (5), to contribute 15.20 per cent of the total weighted score for the borderline intelligence group, but only 11.36 per cent for the average intelligence group. Four subtests, Digit Span, Arithmetic, Picture Completion, and Object Assembly varied systematically with intelligence level in their per cent contribution to total weighted score; the per cent average contribution for these four subtests were corrected by dividing them by the expected per cent subtest contribution at the mean intelligence level of each examiner, and then multiplying by 10.

3. Step 1 was repeated to put all figures again on the basis of per cent average subtest contribution, but this time there was allowance made for the influence of differences in intellectual level. In this way, *examiner* differences would be made to stand out more sharply.

Table I gives these per cent average subtest contributions for each examiner. If the standard deviation of 3 for each subtest, with which the Wechsler was constructed (6, p. 219), was not increased considerably by the three corrections applied (a plausible but untested hypothesis), the standard error of the examiner means can be calculated. Thus, for an examiner who administered seventeen tests, $\sigma\text{mean} = \sigma/N - 1 = 3/\sqrt{16} = .75$.

One mean, that of Examiner A for Arithmetic, is 3.09 removed from the average of examiner means, corresponding to 4.12σ; P less than .00003. Even if this P be multiplied by 143, the number of subtest means under study, the resultant P less than .005 is quite significant. This is the only significant difference to be found in Table I, but it appears to demonstrate fairly conclusively that subjects tested by Examiner A made significantly higher scores on arithmetic than would a random sample of Clinic subjects of the same intelligence level.

The data here subjected to *post hoc* analysis are admittedly more difficult to work with than would be those of a controlled study. However, they do exhibit a rather clear instance of examiner bias.

The rank order correlation coefficient between smallness of inter-examiner variation on a subtest[3] and validity of the subtest[4] is $.59 \pm .21$. This relationship suggests that examiner bias is one of the extraneous factors which reduce Wechsler subtest validity.[5]

BIBLIOGRAPHY

1. Bell, John E. 1948. Projective techniques. New York: Longmans, Green and Co.
2. Guilford, J. P. 1942. Fundamental statistics in psychology and education. New York: McGraw-Hill Book Co., Inc.
3. Klopfer, Bruno, and Douglas McGlashan Kelley. 1946. The Rorschach technique. New York: World Book Co.
4. Miller, D. R., R. Sanders, and S. E. Cleveland. 1950. "The Relationship between examiner personality and obtained Rorschach protocols," Am. Psychol. *5*, 322-3.

[3]Does not include vocabulary.
[4]As measured by the standard deviation of the 13 corrected examiner means.
[4]As measured by the correlation of the subtest with total test minus the subtest (6, p. 224 and p. 101).
[5]This suggestion could be checked by comparing correlation coefficients between subtest and total test minus subtest of (a) subjects tested by the same examiner with (b) subjects tested by any examiner. If examiner bias reduces subtest validity, the single examiner *r*'s (a) should be higher, on the average.

5. Veterans Administration Clinical Psychology Trainees. 1950. A Preliminary study of item difficulty on the Wechsler-Bellevue scale of adult intelligence. Private Communication from Dr. Burke M. Smith, Chief Clinical Psychologist, Veterans Administration Mental Hygiene Clinic, Durham, N. C.

6. WECHSLER, DAVID. 1944. The measurement of adult intelligence. Third edition. Baltimore: The Williams and Wilkins Co.