R E P O R T R E S U M E S

ED 018 824 CG 001 497

"TEST-WISENESS" ON PERSONALITY SCALES.

BY- STRICKER, LAWRENCE J.

AMERICAN PSYCHOLOGICAL ASSN., WASHINGTON, D.C.
EDUCATIONAL TESTING SERVICE, PRINCETON, N.J.
REPORT NUMBER RM-67-22 PUB DATE OCT 67
EDRS PRICE MF-$0.25 HC-$0.68 15P.

DESCRIPTORS- *PERSONALITY TESTS, *TESTING PROBLEMS, *RESEARCH, *TEST VALIDITY, PERFORMANCE,

TEST WISENESS ABILITIES ON PERSONALITY SCALES WERE MEASURED BY SPECIALLY DEVELOPED INSTRUMENTS. THE FOUR TEST-WISENESS VARIABLES WERE--(1) ESTIMATING DESIRABILITY (ACCURACY IN ESTIMATING THE DESIRABILITY OF PERSONALITY ITEMS), (2) ESTIMATING COMMUNALITY (ACCURACY IN ESTIMATING THE COMMUNALITY OF ITEMS), AND (3) ABILITY TO IDENTIFY ITEMS (CORRECTNESS IN IDENTIFYING AND "KEYING" ITEMS). SEVERAL ROLE PLAYING MEASURES BASED ON THE DIFFERENCE BETWEEN SCORES OBTAINED WITH FAKE-GOOD AND FAKE-BAD INSTRUCTIONS ON EACH OF FOUR SCALES FROM THE GUILFORD-ZIMMERMAN TEMPERAMENT SURVEY WERE ALSO ADMINISTERED. TESTS OF VERBAL COMPREHENSION, GENERAL REASONING, AND INDUCTION, MEASURES OF CATEGORY WIDTH AND EQUIVALENCE RANGE ON AN INVENTORY OF SOCIAL DESIRABILITY REPONSE STYLE, AND DEFENSIVENESS MEASURES WERE USED IN THIS STUDY. THE SUBJECTS WERE NINETY-ONE UNDERGRADUATE WOMEN. THE RESULTS INDICATED A HIGH LEVEL OF TEST-WISENESS. HOWEVER, TEST-WISENESS WAS NOT A BROAD GENERAL ABILITY BUT CONSISTS OF A SET OF DISTINCT AND UNRELATED ABILITIES. TEST-WISENESS WAS NOT FOUND TO BE THE MAJOR CAUSE OF DISSEMBLING. THIS PAPER WAS PRESENTED AT THE AMERICAN PSYCHOLOGICAL ASSOCIATION MEETING, (WASHINGTON, D.C., SEPTEMBER 1967). (CG)

RM-67-22

RESEARCH MEMORANDUM

"TEST-WISENESS" ON PERSONALITY SCALES

Lawrence J. Stricker

Paper presented at the American Psychological Association Meeting, Washington, D.C., September, 1967.

Educational Testing Service
Princeton, New Jersey
October, 1967

"Test-Wiseness" on Personality Scales[1]

According to test lore, people vary in their knowledge about tests, and this "test-wiseness" affects their performance on these devices. It is believed that test-wise individuals obtain higher scores on ability and aptitude tests, and distort their scores on personality inventories (Anastasi, 1961; Cronbach, 1960; Ebel & Damrin, 1960; Fishman, Deutsch, Kogan, North, & Whiteman, 1964; Goslin, 1963; Guilford, 1959; Pauck, 1950; Thorndike, 1949; Vernon, 1958, 1962). Despite the prevalence of these notions, the relevant data are sparse (Millman, Bishop, & Ebel, 1965).

Several abilities may be involved in test-wiseness on personality inventories. One is the ability to respond to an inventory in a way that is consistent with a prescribed role. This particular kind of "impression management" (Goffman, 1959) is displayed in role-playing studies (Ellis, 1953; Waters, 1965), which administer personality scales with instructions to "fake" "good" or "bad" roles.

Another potentially relevant ability is accuracy in estimating the desirability of personality items. Edwards' (1957) social desirability paradigm suggests that accurate knowledge of desirability is needed for socially desirable responding and, more generally, for dissembling on personality inventories. Partially supporting such a conception, Wiggins (1966) found that subjects' accuracy in estimating the average desirability ratings made by others did correlate with skill in faking-good on MMPI (Hathaway & McKinley, 1951) clinical scales. However, this accuracy did not correlate with success in faking-good on MMPI scales measuring test-taking attitudes.[2]

A similar ability is accuracy in estimating the "communality" (Wiggins, 1962) or frequency of endorsement of personality items. In Wiggins' (1966) study, this accuracy did not correlate with ability to fake-good on any of the MMPI scales.

Accuracy in analyzing a personality scale and determining the nature of the traits it is intended to measure may also be a pertinent ability. This ability has not been investigated in the context of test-wiseness.

The present study was designed to explore the role of these test-wiseness abilities on personality scales, using specially developed measures of these skills.

Method

Four kinds of test-wiseness variables were employed.[3] One, called Estimating Desirability, assessed accuracy in estimating the desirability of personality items. Estimating Communality, which was a similar instrument, gauged accuracy in estimating items' communality. Ability to Identify Items was a measure of correctness in identifying and "keying" items loading the same factor. There were several role-playing measures. They were based on the difference between scores obtained with fake-good and fake-bad instructions on each of four scales from the Guilford-Zimmerman Temperament Survey (GZTS, Guilford & Zimmerman, 1949). The scales were General Activity, Sociability, Emotional Stability, and Personal Relations. They were given with standard instructions, as well as with fake-good and fake-bad instructions. Several other instruments were administered. These were tests of verbal comprehension, general reasoning, and induction; measures of category width and equivalence range; and an inventory containing social

desirability (SD) response style and defensiveness measures. The subjects were 91 undergraduate women.

Results

In Table 1, comparisons of the mean scores on the test-wiseness measures with their theoretical limits indicate that the subjects were highly test-wise in absolute terms. The means for Ability to Identify Items and for the role-playing measures were particularly close to their theoretical limits.

Insert Table 1 about here

In the analyses that follow, the correlations for the role-playing measures have been reflected in sign so that, in effect, high scores on all test-wiseness measures represent high ability. In Table 2, none of the correlations between the four kinds of test-wiseness measures was significant ($p > .05$), except for the .[illegible] correlation between Estimating Desirability and Estimating Communality. The role-playing measures generally correlated significantly with each other, but the correlations were moderate.

The test-wiseness measures had some significant but moderate correlations with the standard GZTS scales in this table. All but one of the correlations of Estimating Desirability, Estimating Communality, and Ability to Identify Items were with the Emotional Stability scale. In general, the role-playing measures correlated with the standard version of their <u>own</u> scale, but they did not correlate with the <u>other</u> GZTS scales.

Insert Table 2 about here

In Table 3, Estimating Desirability, Estimating Communality, and Ability to Identify Items consistently had significant correlations with the SD measures. The two accuracy instruments were positively related to socially desirable responding, but Ability to Identify Items was negatively related to this response style. In contrast to these extensive correlations, only one correlation between these test-wiseness measures and the defensiveness scales was significant.

Virtually all the significant correlations for the role-playing measures in this table involved the Emotional Stability and Personal Relations scales, on the one hand, and the SD scales, on the other. Role-playing success on these GZTS scales was positively related to socially desirable responding. Like the other test-wiseness instruments, the role-playing measures had few correlations with the defensiveness scales.

Incidentally, none of the test-wiseness measures correlated significantly with any of the ability tests, and they had only scattered correlations with the cognitive style measures.

Among the standard GZTS scales, Sociability, Emotional Stability, and Personal Relations correlated significantly with all the SD measures in this same table. These GZTS scales were positively related to socially desirable responding. They also had relatively low and less consistent correlations with the defensiveness scales, the GZTS scales being positively related to defensiveness.

Insert Table 3 about here

Discussion

The high level of test-wiseness was striking. Equally important was the specificity of the test-wiseness measures. The different kinds of instruments were unrelated, with one minor exception, though there was some generality among the role-playing variables. This finding implies that test-wiseness is not a broad, general ability, but consists of a set of distinct and largely unrelated skills.

The subjects had the ability to distort their scores on personality scales, as gauged from the extent of test-wiseness that was observed. And considerable dissembling apparently did occur, as suggested by the correlations of the response style measures with the GZTS personality scales. But the limited relationships between the test-wiseness measures and the personality scales imply that test-wiseness was not a major cause of this dissembling. Why wasn't test-wiseness closely linked with the distortion on these scales? If test-wiseness didn't produce the dissembling, what did? One plausible answer is that test-wiseness did, indeed, produce the distortion, but this relationship was obscured by the heterogeneity of the subjects. They may have consisted of those unmotivated to distort, those motivated to distort in a favorable direction, and those motivated to distort in an unfavorable direction. The test-wiseness measures would be expected to correlate with the personality scales in the two motivated groups, but in opposite directions, and no correlation would be anticipated in the unmotivated group. Hence, the overall correlations for the total group would be relatively small. Research on this issue is under way. If, in fact, test-wiseness was not the major source of the distortion, no compelling explanation can be offered for

why this should be so or for the real source of the distortion. One speculation is that dissembling is largely unconscious, and, thus, not dependent on the skills involved in the test-wiseness measures.

The consistent correlations of the test-wiseness measures with the SD scales contrast sharply with the scanty correlations of the test-wiseness measures with the defensiveness scales. The explanation may lie in differences in the item composition of the two kinds of scales. Unlike most personality items, including those on the SD scales, the items on the defensiveness scales typically have discrepant desirability and communality (Jackson & Messick, 1962). The desirable items are seldom endorsed (e.g., "I read in the Bible several times a week") and the undesirable ones are often endorsed (e.g., "At times I feel like swearing"). The complex and unusual interaction of desirability and communality on the defensiveness items, their rareness, and their underrepresentation on the test-wiseness measures should attenuate the correlations of the defensiveness scales with the test-wiseness variables. A different explanation of these findings is that the defensiveness scales may tap a characteristic that is under less conscious control than the trait measured by the SD scales. Socially desirable responding may entail a relatively systematic consideration of the items' desirability and communality, accounting for its link with the accuracy measures. Defensiveness, because it may not involve such considerations, is unrelated to knowledge of these item characteristics.

References

Anastasi, Anne. Psychological testing. (2nd ed.) New York: Macmillan, 1961.

Comrey, A. L., & Soufi, A. Further investigation of some factors found in MMPI items. Educational and Psychological Measurement, 1960, 20, 777-786.

Comrey, A. L., & Soufi, A. Attempted verification of certain personality factors. Educational and Psychological Measurement, 1961, 21, 113-127.

Cronbach, L. J. Essentials of psychological testing. (2nd ed.) New York: Harper & Row, 1960.

Ebel, R. L., & Damrin, Dora E. Tests and examinations. In C. W. Harris (Ed.), Encyclopedia of educational research, (3rd ed.), 1960. Pp. 1502-1517.

Edwards, A. L. The social desirability variable in personality assessment and research. New York: Dryden, 1957.

Ellis, A. Recent research with personality questionnaires. Journal of Consulting Psychology, 1953, 17, 45-49.

Fishman, J. A., Deutsch, M., Kogan, L., North, R., & Whiteman, M. Guidelines for testing minority group children. Journal of Social Issues, 1964, 20, 129-145.

Goffman, E. The presentation of self in everyday life. Garden City, N. Y.: Doubleday Anchor Books, 1959.

Goslin, D. A. The search for ability. New York: Russell Sage Foundation, 1963.

Green, B. F., Jr. A note on the calculation of weights for maximum battery reliability. Psychometrika, 1950, 15, 57-61.

Guilford, J. P. Personality. New York: McGraw-Hill, 1959.

Guilford, J. P., & Zimmerman, W. S. Manual of instructions and interpretations, The Guilford-Zimmerman Temperament Survey. Beverly Hills, Calif.: Sheridan, 1949.

Hathaway, S. R., & McKinley, J. C. Manual, Minnesota Multiphasic Personality Inventory. (Rev. ed.) New York: Psychological Corporation, 1951.

Jackson, D. N., & Messick, S. Response styles on the MMPI: Comparison of clinical and normal samples. Journal of Abnormal and Social Psychology, 1962, 65, 285-299.

Layman, Emma M. An item analysis of the adjustment questionnaire. Journal of Psychology, 1940, 10, 87-106.

McNemar, Q. On growth measurement. Educational and Psychological Measurement, 1958, 18, 47-55.

Messick, S., & Jackson, D. N. Desirability scale values and dispersions for MMPI items. Psychological Reports, 1961, 8, 409-414.

Millman, J., Bishop, Carol H., & Ebel, R. An analysis of test-wiseness. Educational and Psychological Measurement, 1965, 25, 707-726.

Pauck, C. E. A square deal for the examination. Education, 1950, 71, 222-225.

Thorndike, R. L. Personnel selection. New York: Wiley, 1949.

Vernon, P. E. Educational testing and test-form factors. Research Bulletin 58-3. Princeton, N. J.: Educational Testing Service, 1958.

Vernon, P. E. The determinants of reading comprehension. Educational and Psychological Measurement, 1962, 22, 269-286.

Waters, L. K. A note on the "fakability" of forced-choice scales. Personnel Psychology, 1965, 18, 187-191.

Wiggins, J. S. Interrelationships among MMPI measures of dissimulation under standard and social desirability instructions. Journal of Consulting Psychology, 1959, 23, 419-427.

Wiggins, J. S. Strategic, method, and stylistic variance in the MMPI. Psychological Bulletin, 1962, 224-242.

Wiggins, J. S. Social desirability estimation and "faking good" well. Educational and Psychological Measurement, 1966, 26, 329-341.

Yonge, G. D., & Heist, P. A. The influence of suggested contexts on faking a personality test. American Educational Research Journal, 1965, 2, 137-144.

Footnotes

[1]This study was supported by the National Institute of Child Health and Human Development, under Research Grant 1 P01 HD 01762-01. Thanks are due to Henrietta Gallagher for supervising the statistical analyses.

[2]These results were obtained with an "absolute accuracy" score, which is roughly analogous to the scores used in the present study.

[3]Estimating Desirability employed 95 heterogeneous MMPI items. The subject judged the social desirability of each on a nine-point scale. The score was the product-moment correlation, transformed to Fisher's z, between his judgments for the items and the items' social desirability scale values (Messick & Jackson, 1961).

Estimating Communality employed 95 other MMPI items. Using a nine-point scale, the subject judged how frequently college students responded "true" to each item. The score was the transformed correlation between his judgments and the items' actual endorsement frequencies for 190 Stanford undergraduates (Wiggins, 1959).

Ability to Identify items consisted of three similarly constructed subtests, each based on a different published factor analysis of personality items (Comrey & Soufi, 1960, 1961; Layman, 1940). A subtest consisted of 15 items: the eight with the highest loading on the same factor, and seven others--each with the highest loading on one of seven other factors. The subject was instructed to identify the items that refer to the same personality trait and to "key" them ("true" or "false"). The score was the number of items from the main factor that were identified as involving the same trait and correctly keyed, plus the number of items not

from that factor which the subject indicated did not refer to the same trait. A total score for the instrument was obtained, weighting the subtest scores for maximal reliability (Green, 1950).

The role-playing measures employed a difference score (McNemar, 1958, p. 48), based on the fake-bad score minus the fake-good score. The role-playing instructions were adapted from those for the college admission situation used by Yonge and Heist (1965).

Table 1

Means, Standard Deviations, and Theoretical Score

Limits of Test-Wiseness Measures

Measure	Mean	S.D.	Theoretical Score Limits: Minimum Test-Wiseness	Theoretical Score Limits: Maximum Test-Wiseness
Estimating Desirability	1.08	.16	-3.00	3.00
Estimating Communality	.68	.27	-3.00	3.00
Ability to Identify Items	15.86	1.20	.00	19.39
Role-Playing--General Activity	-10.55	5.72	28.48	-21.93
Role-Playing--Sociability	-13.35	5.71	30.18	-18.74
Role-Playing--Emotional Stability	-18.56	4.31	25.77	-24.37
Role-Playing--Personal Relations	-17.33	4.13	21.47	-24.50

Table 2

Intercorrelations of Test-Wiseness Measures and Their Correlations with Standard GZTS Scales

	Test-Wiseness Measures							Standard GZTS Scales			
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	G.A.	S	E.S.	P.R.
(1) Estimating Desirability	(.63)	.23*	.07	-.02	-.04	.06	.10	.08	.02	.21*	.11
(2) Estimating Communality		(.86)	.09	.14	.03	.08	.14	.06	-.07	.27**	.20*
(3) Ability to Identify Items			(.62)	.02	.04	-.02	-.10	.16	.04	-.29**	-.04
(4) Role-Playing--General Activity				(.86)	.49**	.42**	.06	.27**	.09	-.05	-.11
(5) Role-Playing--Sociability					(.90)	.39**	.15	-.06	-.01	.01	-.05
(6) Role-Playing--Emotional Stability						(.84)	.40**	.18	.14	.36**	.16
(7) Role-Playing--Personal Relations							(.79)	.08	-.09	.27**	.47**

Note.--The internal-consistency reliability coefficients for the test-wiseness measures appear in parentheses.

*Significant at the .05 level; **significant at the .01 level.

Table 3

Correlations of Test-Wiseness Measures and Standard GZTS

Scales with Response Style Measures

Response Style Measure	Est. Des.	Est. Comm.	Abil. Ident. Items	Role-Playing G.A.	Role-Playing S	Role-Playing E.S.	Role-Playing P.R.	Standard GZTS G.A.	Standard GZTS S	Standard GZTS E.S.	Standard GZTS P.R.
Social Desirability:											
Edwards' SD	.26**	.26**	-.22*	.00	-.03	.29**	.26**	.22*	.46**	.75**	.43**
Stricker's SD	.14	.24*	-.20*	-.15	.00	.20*	.22*	.05	.38**	.66**	.55**
CPI Wb	.23*	.20*	-.23*	-.02	.05	.25*	.07	.15	.43**	.62**	.41**
MMPI K	.12	.23*	-.16	-.08	-.02	.15	.09	.03	.42**	.58**	.44**
MMPI F	-.27**	-.31**	.20*	.13	.01	-.18	-.20*	-.02	-.25*	-.44**	-.38**
MMPI F-K	-.22*	-.31**	.21*	.12	.02	-.19	-.16	-.03	-.42**	-.62**	-.49**
Defensiveness:											
Wiggins' Sd	.02	-.11	-.04	-.01	-.23*	.12	.05	.27**	.35**	.19	.06
Marlowe-Crowne SD	.06	.06	-.12	-.12	-.12	.18	.01	.07	.31**	.42**	.29**
MMPI L	-.04	-.04	-.20*	.01	-.12	.17	.03	-.01	.25*	.24*	.26**

*Significant at the .05 level; **significant at the .01 level.