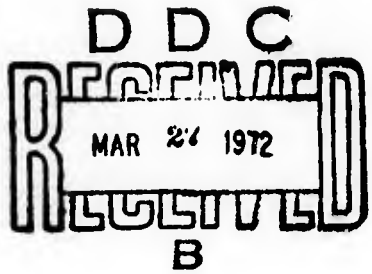
THE CENTER FOR THE STUDY OF ORGANIZATIONAL PERFORMANCE AND HUMAN EFFECTIVENESS

University of Minnesota
Minneapolis, Minnesota

Office of Naval Research Contract
ONR N00014-68-A-0141-0003

Approved for public release; distribution unlimited

R21



THE EQUIVALENCE OF SEMANTIC AND FIGURAL TEST PRESENTATION OF THE SAME ITEMS

Howard E. A. Tinsley and Rene' V. Dawis

Technical Report No. 3004

Security Classification

DOCUMENT CONTROL DATA - R & D

Security classification of title, body of abstract and indexing annotation must be entered when the overall report is classified

1. ORIGINATING ACTIVITY (Corporate author)	2a. REPORT SECURITY CLASSIFICATION
University of Minnesota Department of Psychology	UNCLASSIFIED
	2b. GROUP

3. REPORT TITLE

The Equivalence of Semantic and Figural Test Presentation of the Same Items

4. DESCRIPTIVE NOTES (Type of report and inclusive dates)

Technical Report

5. AUTHOR(S) (First name, middle initial, last name)

Rene' V. Dawis and Howard E. A. Tinsley

6. REPORT DATE	7a. TOTAL NO. OF PAGES	7b. NO. OF REFS
January, 1972	17	35

8a. CONTRACT OR GRANT NO.	9a. ORIGINATOR'S REPORT NUMBER(S)
N00014-68-A-0141-0003	3004
b. PROJECT NO. NR 151-323	
c.	9b. OTHER REPORT NO(S) (Any other numbers that may be assigned this report)
d.	

10. DISTRIBUTION STATEMENT

Approved for public release; distribution unlimited

11. SUPPLEMENTARY NOTES	12. SPONSORING MILITARY ACTIVITY
	Personnel & Training Research Programs Office of Naval Research Department of the Navy ~~Arlington, Virginia~~ 22217

13. ABSTRACT

A 30-item multiple-choice word analogy test and a corresponding 30-item picture analogy test (in which the pictures corresponded to the words in the work analogy test) were administered to 289 Civil Service employees. The equivalence of semantic (word) and figural (picture) test presentation of the same items was determined by comparing the responses of the same subject to the same item. Proportion of correspondent responses (both correct or both wrong) ranged from .69 to .91 with a median of .84. Correlation between scores on the two test forms was .86. Over 84% of the subjects gave correspondent responses with greater than chance frequency. Score distributions were practically identical. It was concluded that semantic and figural parallel test forms can be constructed.

14 KEY WORDS	LINK A		LINK B		LINK C	
	ROLE	WT	ROLE	WT	ROLE	WT
Test equivalence						
Analogy test						
Semantic test						
Figural test						
Culture-fairness of tests						
Test presentation mode equivalence						

The Equivalence of Semantic and Figural Test Presentation of the Same Items

Howard E. A. Tinsley
and
Rene' V. Dawis

In 1962, several books were published which, along with a subsequent awakening to the realities of racism in America, were helpful in focusing public attention on the cultural fairness of psychological tests (Black, 1962; Gross, 1962; Hoffman, 1962). A great deal of discussion followed regarding the feasibility and desirability of developing "culture-fair" tests (Anastasi, 1967; Ash, 1967; Doppelt and Bennett, 1967; Krug, 1964; and Wesman, 1968). The research evidence bearing on this question is mixed. A number of studies on the prediction of educational achievement have indicated that validity coefficients for black and other culturally disadvantaged students are as high or higher than validity coefficients for white or culturally advantaged students (Cleary, 1966; Hewer, 1965; Hills, 1964; Roberts, 1964; and Stanley and Porter, 1967); similar findings have been observed in at least one study of the prediction of vocational criteria (Gordon, 1953). On the other hand, a series of studies with employees of the Port of New York Authority (Lopez, 1966) revealed different relationships between predictors and job performance criteria for black and white toll collectors and maintenance men. Moreover, an extensive investigation by the Research Center for Industrial Behavior at New York University (Kirkpatrick, Ewen, Barrett, and Katzell, 1968) indicated that many tests performed equally well in different ethnic groups but that in some cases different tests worked best for different groups. Inclusion of an index of cultural disadvantage as a moderator variable improved test validity for some jobs. Included in this study were some 1200 persons; white, black and Puerto Rican clerical workers,

nursing students, and participants in job training programs for maintenance work and heavy equipment operation.

Anastasi (1964) has pointed out that in designing "culturally fair" tests, it is important to distinguish between those cultural factors that affect both the test and criterion behavior, and those that influence only the test behavior. The former are necessary to insure the validity of the test. It is the "test-specific cultural factors" which "culturally bias" a test. Research with the mentally retarded indicates that verbal ability may constitute one such biasing factor in tests designed to predict vocational criteria. A number of investigators have reported that the diagnosis of a person as mentally retarded on the basis of his performance on a verbal test is vocationally meaningless (Bobroff, 1956; Collman and Newlyn, 1957; Kauppi, 1963; Kauppi and Weiss, 1967; Muench, 1944; and Seashore, Wesman and Doppelt, 1950). In general, the evidence indicated that the vocational adjustments of the mental retards was far too heterogeneous and showed far too much overlap with that of non-retarded workers for the diagnosis of a person as mentally retarded to have valid implications for vocational success. Kauppi and Weiss (1967, p. 348) concluded that "Knowing that a client is mentally retarded tells the counselor only that he is probably below average on verbal tasks. The label says little about other abilities, interests, needs or potential."

Other researchers have also indicated the desirability of eliminating the verbal ability "bias" found in tests. The United States Employment Service (Jurgensen, 1966) has experimented with non-reading forms of several GATB tests. Rimland (1967) has suggested the use of the Porteus Maze test, a non-verbal test of general mental ability. Freeberg (1970) has experimented with verbal tests and with tests in which pictorial information is accompanied by verbal information, with primary emphasis given to making

the tests consist of more culturally relevant stimulus materials. And Krug (1964) has suggested the use of biographical information and situation tests.

Guilford's (1956, 1959) Structure of Intellect provides the best theoretical framework within which to pursue this discussion. His model represents the human intellect as a cube having "content," "process," and "product" dimensions. Each cell in the cube represents a unique factor (or set of factors) of intellectual ability. Thus, a semantic test (one which uses words to ask questions, thereby requiring verbal ability) supposedly measures a factorially different ability than a figural test (one which uses pictures to ask questions). Because they measure different factors, the two types of tests may be differentially related to many criteria. Guilford (1959) has suggested that the abilities involved in using figural (picture) information are most closely related to success as a mechanic, machine operator, artist, or musician, and are related to success in certain aspects of engineering, while the abilities involved in using semantic (verbal) information are most closely related to success in educational settings where the learning of verbally presented facts and ideas is essential.

There is a great deal of evidence, then, to suggest that a test is culturally "biased" only because the test measures some culturally related factor which does not influence the criterion behavior, e.g., a verbal ability component may "bias" tests used to predict criteria not influenced by verbal ability. The work of Guilford indicates that semantic (verbal) and figural (picture) tests measure different factors. It is possible, therefore, that figural tests may operate as unbiased predictors in those instances where semantic tests are culturally biased. The choice between semantic and figural tests, however, represents a kind of "all-or-none"

choice. Because the semantic and figural tests used in past research have often been developed independently without any attempt to make them equivalent, they may differ in a number of respects, only some of which are related to verbal ability. It is important to identify those factors which contribute to differences in test scores on the two types of tests. This will allow the elimination from a test of those factors which have a biasing effect and the retention of those factors which contribute to the predictive validity of the test.

Heim (1954) has suggested two variables related to the structure of test items which he believes have an effect on the difficulty of the item. First, Heim has presented evidence that the type of question format (multiple choice or inventive answer) has a bearing on the difficulty of the item; Guilford (1959) has demonstrated that such questions measure factorially different abilities. Again, Heim (1954) has suggested that differences in the internal structure of an item might influence item difficulty. Ace and Dawis (in press) have demonstrated this to be true under certain conditions.

Of more interest to the present authors are three item characteristics suggested by Spearman (1927), who observed that the complexity, abstractness, and novelty of test items seemed to be the factors important in determining their difficulty. The present authors believe that if such variables as test instructions, time limits, administrator comments, item format, and the internal structure of the item are held constant, four factors may still operate to produce differences in the scores obtained on semantic (verbal) and figural (picture) tests. First, such tests may differ in the level of abstraction they require. Many concepts are easy to express verbally but are extremely difficult to express in a picture.

Examples include emotions (love, hate, affection), time other than present (past, future) and degree (better, best). Unless semantic and figural tests are equated for the level of abstraction, it is likely that the semantic test will require a higher level of abstract thinking. This may be detrimental if the criterion behavior is not related to ability in abstract thinking. Secondly, semantic and figural tests may differ in their "novelty". It seems likely that many respondents will find one type of test stimulus more familiar than the other. To do well on a semantic test requires a familiarity with the words used (a good vocabulary) while achievement on a figural test requires familiarity with the appearance of objects. Few people, for example, would recognize the word "ibex," yet most would recognize a picture of a wild mountain goat. Conversely, few people could identify a clutch or brake drum from a picture but many have those words in their vocabulary. A third way in which semantic and figural tests may differ is in their complexity. Campbell (1961) has reported that the effects of complexity (defined as the number of item properties to be taken into consideration in arriving at the correct answer) on the difficulty of symbol classification is due primarily to the nature of the classifying property. Classification by shape led to the least item difficulty; classification by size led to the most difficulty. Finally, semantic and figural tests usually represent different samples of test behavior. (This item characteristic is referred to hereafter as the item content, but should not be confused with Guilford's notion of "content" which refers to the type of stimulus material--words, pictures, numbers, or symbols--used to present the item.) Even when two semantic tests have been designed to be parallel measures of the same ability, they often do not yield identical ability estimates. Most semantic and figural tests have not been designed to be parallel, so differences in ability estimates are to be expected.

The present authors hypothesize, then, that the differences observed by Guilford (1959) in semantic and figural tests are due to differences in the abstractness, novelty, complexity, and content of the test items. Two tests which have been equated for these factors should yield roughly equivalent scores even though one uses semantic items while the other uses figural items. The remainder of this paper is concerned with an investigation of this question.

METHOD

Instrumentation: The analogy question format was selected for this research because of its wide use in tests and because analogy tests seem to represent Spearman's "g" more closely than other tests (Helmstadter, 1964, p. 99). A list of relationships which could be expressed in analogy format was compiled and used as a guide in constructing a pool of 100 picture analogies. A set of 30 picture analogies which included most of the pictures used in the 100 picture analogies was administered to a group of 46 college students. In addition to completing the analogy, the subjects were asked to identify the object in each picture. Most pictures were identified by greater than 90% of the subjects. Those pictures which were correctly identified by fewer than 80% of the students were discarded and new pictures were taken to represent the concept.[1] The total item pool of 100 analogies was then administered to 301 college students and the 30 picture analogies having the highest point-biserial correlation with total score were selected for further study. Next, thirty word analogies were constructed by expressing each picture analogy in word form, thus pairing every picture analogy with a word analogy of identical content. Because the items were so exactly paired in

---

[1] The authors wish to express their gratitude to Mr. Merle Ace, University of British Columbia, who supervised the construction of the 100 picture analogies and the analysis of the recognizability of the objects in the pictures.

terms of content, it is assumed that both items in each of the 30 item pairs were also equivalent in abstractness and complexity. The 60 analogy items were then combined in an instrument with the 30 word analogies first. For each type of analogy, the order of presentation was randomized.

Both tests were designed to minimize the novelty of the items. The pictures in the picture analogies were of commonplace items although the relationship expressed by the analogy was often complex. The object in each picture was correctly identified by 80% of college students. All but 15 of the words in the word analogies appear on the Dale and Chall (1948) list of 3000 words familiar to 80% of fourth graders. Because the novelty of the picture items was judged from the responses of college students while the novelty of the word analogies was judged from the responses of fourth grade students, the items may be imperfectly equated for their novelty with the picture analogies containing the more novel stimuli.

Subjects: The tests were administered to 289 Minneapolis Civil Service employees as part of a battery of tests. Twenty subjects were dropped for failure to respond to all of the items. The remaining 269 subjects were predominately white (96%) females (97%) who ranged in age from 18 to 64. The median age was 33; the modal ages were 20 and 21; 42% of the sample was 26 years of age or younger. All but 4 subjects had a high school education, 17% had some college, and 3.4% had a college degree. The median family income was $9000 per year; the modal income was $10,000 per year.

Analysis: This research was concerned with the question of whether semantic and figural analogy items of equivalent abstractness, novelty, complexity, and content would yield equivalent results. Analyses were performed at the item and the test level. Because each question had five alternatives, the expected chance probability of a subject's correctly answering both items

in the pair was .04 (.2 x .2), the expected chance probability of his incorrectly answering both items in the pair was .64 (.8 x .8), and the expected chance probability of correspondent responses (both responses correct or both responses incorrect) was .68. Accordingly, a 1-tailed z test was performed for each of the 30 pairs of items to determine whether the proportion of correspondent responses was significantly greater than .68. This represented an extremely stringent test of the hypothesis, however, as measurement at the single item level is seldom precise. At the test level, the data were analyzed as two 30-item analogy tests. A 2-tailed t-test was performed to determine whether the mean total scores on the semantic and figural forms were equivalent, an F test was performed to determine whether the variances of the total scores on the two forms were equivalent, and the product-moment correlation was computed between the total scores on the two forms.

The above analyses indicated the extent to which the semantic and figural items yielded statistically equivalent or correspondent results for the total sample. Also of interest was the degree to which the two types of items yielded equivalent measurement for each individual. A 2-tailed z test was performed for each of the 269 subjects to determine whether the proportion of correspondent responses to the item pairs departed significantly from the expected chance rate of .68. This, again, is a somewhat stringent test. Even the most rigorously developed of parallel forms will not yield identical scores for all subjects. It is justifiable, therefore, to ask whether the observed differences in scores can be explained in terms of the error of measurement. To answer this question, standardized difference scores were computed for each subject. First, the standard error of measurement of the picture form was computed from the item analysis data. Then the difference between the total scores for each person on the word and picture analogies

was expressed as a proportion of this standard error of measurement. Wright (1967), in commenting on this procedure, points out that if the variation in scores is of the same magnitude as that expected from the error of measurement of the test, then the distribution of standardized difference scores should have a mean of zero and a standard deviation of 1.0.

RESULTS

For each of the 30 pairs of analogy items, the 1-tailed z test was employed to determine whether the proportion of subjects making correspondent responses significantly exceeded the proportion expected by chance. The proportion of correspondent responses was significantly greater than chance at the .005 level of confidence for 27 items; the 3 remaining items failed to achieve significance at the .05 level of confidence (see Table 1).

Insert Table 1 about here

The data were also analyzed to determine whether the 30 semantic and 30 figural analogies could be regarded as parallel forms having equal means and equal variances. The mean total score was 18.0 for the semantic analogies and 18.2 for the figural analogies; the variances were 24.4 and 28.3 respectively. Neither the two sample t-test for the difference between means ($t = .45$, $df = 268$) nor the F-test for homogeneity of variance ($F = 1.16$, $df = 268, 268$) was significant at the .05 level of confidence. The correlation between scores on the semantic and figural forms was .86.

For each subject, a 2-tailed z test was performed to determine whether the proportion of correspondent responses made by that person was significantly different from the proportion that could be expected by chance. In order for the proportion of correspondent responses to exceed significantly

Table 1

Number and Proportion of Correspondent Responses for Thirty Semantic Analogy-Figural Analogy Pairs (N=269)

Item	N	Proportion of correspondent responses
1	217	.80
2	186	.69*
3	206	.76
4	218	.81
5	205	.76
6	230	.85
7	205	.76
8	226	.84
9	214	.79
10	218	.81
11	230	.85
12	228	.84
13	208	.77
14	220	.82
15	195	.72*
16	240	.90
17	246	.91
18	209	.77
19	236	.87
20	246	.91
21	204	.76
22	230	.85
23	228	.84
24	227	.84
25	229	.85
26	228	.84
27	189	.70*
28	228	.84
29	235	.87
30	226	.84

* Not significant at the .05 level. All other pairs are significantly correspondent at the .005 level.

the proportion expected by chance (.63), the subject needed to make 26 (86.7%) correspondent responses; 125 (46.5%) of the 269 subjects fell in this category. Only 6 (2.2%) of the examinees made significantly fewer than chance (15 or less) correspondent responses; the remaining 138 subjects fell in the chance range. In this latter group, 102 (73.9%) made correspondent responses to more than 63% of the item pairs. In all, then, 227 (84.4%) subjects made correspondent responses with greater than chance frequency while 49 (15.6%) made correspondent responses with less than chance frequency.

The standard error of measurement for the figural analogy test (as computed from the item analysis data) was 2.32. The mean and standard deviation of the distribution of standardized difference scores were -.03 and 1.18 respectively.

DISCUSSION

The analyses at both the item and the test level support the conclusion that the semantic and figural analogies used in this study were measuring the same trait. At the item level, correspondent responses occurred at a significantly greater than chance frequency for 27 of the 30 analogies. An analysis of the three discrepant item pairs suggests that the quality of the pictures may account for their failure to support this conclusion. The analogies in question read:

Peanut: _____ :: Lettuce : Cabbage

1. Plowed field 2. Butter 3. Potato 4. Raddish 5. Beans

Carrot: _____ :: Orange : Innertube

1. Block 2. Alligator 3. Canoe 4. Fire 5. Telephone

_____ : Hinge :: Arm : Elbow

1. Handle 2. Door knob 3. Door frame 4. Desk leg 5. Door

In the picture form of the first analogy, the peanut, the raddish, and the beans are particularly difficulty to identify. In the picture form of the second analogy, the innertube looks like a ring bologna or a sausage and the block looks like a bar of soap. Incidentally, these five pictures did not appear in the 30 picture analogies employed for the evaluation of picture clarity. In the third analogy, the correct answer is door. It seems likely that the distinction between door and door frame is not as clear in the picture form as it is in the word form of the analogy. It was concluded, therefore, that the proportion of correspondent responses failed significantly to exceed the proportion expected by chance because of the clarity of some of the pictures used in these analogies.

At the test level, the distributions of scores on the semantic and figural tests were practically the same. The figural test scores were slightly more variable than the semantic test scores but the difference was not significant. The product-moment correlation between scores on the two forms (.86) was high considering the experimental nature of the two forms. All of the evidence indicates that the two tests are measuring the same trait.

The distribution of standardized difference scores also supports this conclusion. The data indicate that most of the differences observed in scores on the two forms can be attributed to errors of measurement. The small amount of difference score variance remaining after the variance attributable to errors of measurement has been removed may well be due to differences in the novelty of the stimuli or to the use of uninterpretable pictures.

The above conclusions are based upon data from the entire sample. An analysis of the data for an individual at a time leads to essentially the

same conclusion. Over 84% of the subjects gave more correspondent responses than would have been predicted on the basis of chance responding. Only 2.2% gave significantly fewer correspondent responses than expected on the basis of chance.

These results, then, indicate that the distinction between semantic and figural tests needs to be examined more closely. Semantic and figural "parallel forms" can be constructed. This implies that the differences which have been observed in performance on such tests are not necessarily the result of differences in the stimulus material (pictures and words), but can be the result of other characteristics which usually covary with stimulus differences. The present authors suggest that the abstractness, novelty, complexity, and content of the items may be the most meaningful dimensions on which these items vary. Research on "culture-fair" tests may be more profitably spent in investigating these dimensions rather than in comparing semantic (word) and figural (picture) tests.

REFERENCES

Ace, M. E. and Dawis, R. V. The effects of two properties of item structure on item difficulty in verbal analogies. In press, Educational and Psychological Measurement.

Anastasi, A. Culture-fair testing. Educational Horizons, 1964, 43, 26-30.

Anastasi, A. Psychology, psychologists, and psychological testing. American Psychologist, 1967, 22, 297-306.

Ash, P. Selection techniques and the law: Discrimination in hiring and placement. Personnel, 1967, 6, 8-17.

Black, H. They shall not pass. New York: Morrow, 1962.

Bobroff, A. Economic adjustment of 121 adults, formerly students in classes for mental retardates. American Journal of Mental Deficiency, 1955-1956, 60, 225-235.

Campbell, A. C. Some determinants of the difficulty of non-verbal classification items. Educational and psychological measurement, 1961, 21, 899-913.

Clearly, T. A. Test bias: Validity of the Scholastic Aptitude Test for negro and white students in integrated colleges. Educational Testing Service Research Bulletin, RB-66-31, 1966.

Collman, R. D. and Newlyn, D. Employment success of mentally dull and intellectually normal ex-pupils in England. American Journal of Mental Deficiency, 1956-1957, 61, 434-490.

Dale, E. and Chall, J. S. A formula for predicting readability: instructions. Educational Research Bulletin, 1948, 27, 37-54.

Doppelt, J. E. and Bennett, G. K. Testing job applicants from disadvantaged groups. Test Service Bulletin, No. 57, The Psychological Corporation, May, 1967.

Freeberg, N. E. Assessment of disadvantaged adolescents: A different approach to research and evaluation measures. Journal of Educational Psychology, 1970, 61, 229-240.

Gordon, M. A. A study of the applicability of the same minimum qualifying scores for technical schools to white males, WAF, and negro males. San Antonio: Human Resources Research Center, Lackland Air Force Base. Technical Air Force Base. Technical Report 53-54, 1953.

Gross, M. L. The brain watchers. New York: Random House, 1962.

Guilford, J. P. The structure of intellect. Psychological Bulletin, 1956, 53, 267-293.

Guilford, J. P. Three faces of intellect. American Psychologist, 1959, 14, 469-479.

Heim, A. W. The appraisal of intelligence. London: Methuen and Company, Ltd., 1954.

Helmstadter, G. C. Principles of psychological measurement. New York: Appleton-Century-Crofts, 1964.

Hewer, V. H. Are tests fair to college students from homes with low socio-economic status? Personnel and Guidance Journal, 1965, 43, 764-769.

Hills, J. R. Prediction of college grades for all public colleges of a state. Journal of Educational Measurement, 1964, 1, 155-159.

Hoffman, B. The tyranny of testing. New York: Crowell-Collier, 1962.

Jurgensen, C. E. Advisory panel appraises suitability of USES testing. The Industrial Psychologist, 1966, 4, 41-44.

Kauppi, D. R. The application of general semantics to the classification of mentally retarded. Paper presented at the International Conference on General Semantics, Denver, 1968. (Minneapolis: University of Minnesota, Work Adjustment Project, Research Report Number 17, mimeographed.)

Kauppi, D. R. and Weiss, D. J. The utility of the classification "Mentally Retarded" in vocational psychology. Proceedings, 75th Annual Convention, APA, 1967, 347-348.

Kirkpatrick, J. J., Ewen, R. B., Barrett, R. S. and Katzell, R. A. Testing and fair employment: Fairness and validity of personnel tests for different ethnic groups. New York: New York University Press, 1968.

Krug, R. E. Some suggested approaches for test development and measurement. Presented at symposium: The Industrial Psychologist, Selection and Equal Employment Opportunity, American Psychological Association Convention, Los Angeles, California, 1964.

Lopez, F. M. Current problems in test performance of job applicants. Personnel Psychology, 1966, 19, 10-17.

Muench, G. A. A follow-up of mental deffectives after 18 years. Journal of Abnormal and Social Psychology, 1944, 39, 407-418.

Rimland, B. Proposal for research on selection for military effectiveness, with emphasis on the testing of marginal personnel, (draft) U. S. Naval Personnel Research Activity, San Diego, California, 1967.

Roberts, S. O. Ethnic background and test performance in selecting and training negroes for managerial positions. The Executive Study Conference. Princeton, New Jersey: Educational Testing Service, 1964.

Seashore, H. G., Wesman, A. G. and Doppelt, J. E. The standardization of the Wechsler Intelligence Scale for Children, Journal of Consulting Psychology, 1950, 14, 99-110.

Spearman, C. The abilities of man. London: Macmillan and Company, Ltd., 1927.

Stanley, J. C. and Porter, A. C. Correlation of Scholastic Aptitude Test scores with college grades for negroes versus whites. Journal of Educational Measurement, 1967, 4, 199-218.

Wesman, A. G. Intelligence testing. American Psychologist, 1963, 23, 267-274.

Wright, B. D. Sample-free test calibration and person measurement. Invitational conference on testing problems. Princeton, New Jersey: Educational Testing Service, 1967, 85-99.