DOCUMENT RESUME

ED 069 692 TM 002 143

AUTHOR	Passmore, David L.
TITLE	A Study of the Usefulness of Weighting Test Item Responses.
INSTITUTION	Minnesota Research Coordinating Unit in Occupational Education, Minneapolis.
SPONS AGENCY	Office of Education (DHEW), Washington, D.C.
PUB DATE	May 72
NOTE	15p.; Part of the fellowship program, "Preparing Researchers in Vocational Education"
EDRS PRICE	MF-$0.65 HC-$3.29
DESCRIPTORS	Attitude Tests; Bibliographies; *Discriminant Analysis; *Item Analysis; *Job Satisfaction; *Predictive Measurement; Questionnaires; *Research Methodology; Statistical Analysis; Tables (Data); Technical Reports; Test Interpretation; Vocational Adjustment; Work Attitudes
IDENTIFIERS	Minnesota Satisfaction Questionnaire; MSQ

ABSTRACT

The purpose of this study was to investigate the practicality of multiple discriminant function analysis for deriving item response weights. Item responses on a job satisfaction questionnaire administered to 219 professional workers and 242 semi-skilled customer workers were analyzed. Discriminant functional analysis was conducted on the total sample. Respondents were then randomly assigned to one of two subsamples. Two different discriminant function analyses were then undertaken to maximize group differences in each of the samples. Inconsistency of the results indicates that weights so derived are not generalizable to an independent sample from the same population. Though an increment in predictive efficiency of 11% might be realized with the differential weighting system, the veracity of the increment is doubtful due to the failure of the differential weighting system to cross-validate. Investigating other weighting techniques, such as latent trait measurement models, empirical techniques for weighting each multiple-choice alternative of a test item, and confidence weighting, is suggested. (DJ)

A STUDY OF THE USEFULNESS OF WEIGHTING TEST ITEM RESPONSES

by

David L. Passmore
USOE Research Fellow
Minnesota Research Coordinating Unit for Vocational Education
University of Minnesota
Minneapolis, Minnesota

May 1972

A STUDY OF THE USEFULNESS OF WEIGHTING TEST ITEM RESPONSES

David L. Passmore[1]

It may be important for vocational teacher educators to prepare tests which estimate the technical competencies of skilled workers who desire to enter vocational teacher preparation programs (Cf. Griess, 1967). Obviously, such competency tests need the sensitivity to separate technically competent from incompetent skilled workers. Similarly, an attempt may be made to differentiate between the interest patterns of various occupational criterion groups so that vocational counselors may determine the similarity of their advisees' interest patterns to those of the criterion groups. The Strong Vocational Interest Blank (Strong, 1943) was the result of such an effort. As another example, evaluators may be interested in devising criterion measures which discriminate between the products of various vocational education programs in terms of their ultimate work adjustment (Smith, Passmore, Moss, and Copa, 1971).

After criterion measures have been developed, there are a number of ways to enhance the detection of group differences. One way would be to multiply each examinee's responses to items on the test instrument by some set of numerical weights. There would be a different weight for each item. These weights would be derived to maximally separate the criterion groups in question. If the weighted responses to the test items would be summed, the difference between the mean weighted total scores of the criterion groups should be greater than the difference between their mean unweighted scores.

Unfortunately, measurement specialists have had little success in solving these measurement problems by the application of differential weighting techniques

(Stanley & Wang, 1970a). However, multiple discriminant function analysis (Fisher, 1936), a statistical technique used for maximally separating any number of criterion groups by differentially weighting a set of predictors, has received little attention but may be appropriate for yielding useful response weights in these circumstances. Federico (1971) applied this technique successfully using item responses from a questionnaire designed to estimate the training satisfaction of U.S. Air Force personnel. If his results are replicable, then perhaps multiple discriminant function analysis may be an appropriate tool for test constructors in occupational education.

The purpose of this study was to investigate the practicality of multiple discriminant function analysis for deriving item response weights. The following questions were explored:

(a) Do the results of discriminant function analysis of item response data provide a generalizable scheme for weighting scored item responses?

(b) Does the application of a differential weighting scheme provided by discriminant function analysis help to differentiate between groups of examinees better than an unweighted scoring of items?

Method

Subjects. Item responses on a job satisfaction questionnaire administered to the following two groups were analyzed:

(a) Group 1 - Professional workers (N=219) performing in the same occupation for five years or less. Most of these workers were college-educated males.

(b) Group 2 - Semi-skilled customer service workers (N=242) with five years or less experience in the same organization. The majority of these workers were females with a high school education.

These two groups were chosen because of certain characteristics which might be indicative of differences in job satisfaction. Mann (1953) found differences

in levels of job satisfaction between male and female workers and, also, between workers with various degrees of education. Wilensky (1964, pp. 138-140) hypothesized that there would be differences in the patterns of the job satisfaction of professional and semi-skilled workers.

It is noted that these subjects were not selected from an educational situation but were selected merely for convenience sake to test the practicality of discriminant function analysis as a measurement tool for test constructors in occupational education.

Instrumentation. Each examinee's job satisfaction was estimated by means of the Minnesota Satisfaction Questionnaire (MSQ) which was developed by the staff of the Work Adjustment Project at the University of Minnesota (Weiss, Dawis, England, & Lofquist, 1967). The content of each of the 21 Likert-type items on the MSQ deals with some aspect of job satisfaction. The internal consistency reliability of the MSQ, estimated by an analysis of variance technique (Hoyt, 1941), for this sample of respondents was .90 and the standard error of measurement based on raw scores was 1.06.

Data analyses. Reference is made to Figure 1. Discriminant function analysis was conducted [see (1)] on the item responses of the Total Sample. Pre-

---

Insert Figure 1 About Here

---

dictor variables of interest in this study were the 21 MSQ items. Discriminant coefficients were calculated for each item so that criterion groups were maximally separated when these weights were applied to the examinees' MSQ responses. This analysis was expedited through the use of a computer program developed by Veldman (1967). Respondents in the Total Sample were then randomly assigned to two subsamples of nearly equal size which were designated Sample A and Sample B.

Two different discriminant function analyses, (2) and (3), were undertaken to maximize group differences in each of these subsamples.

In order to determine the generalizability of the functions derived at (2) and (3), a double cross-validation pattern (Katzell, 1951, pp. 20-21; Mosier, 1951, p. 11) was designed. Cross-validation of the weights derived in (2) was accomplished [see (4)] by applying B's weights to Sample A's responses. To cross-validate the weights derived in (3), the discriminant function derived for Sample B was used [see (5)] to weight the item responses of Sample A. Both cross-validations were handled by means of a computer program called Program CROSVAL (Passmore & Irvin, 1972). This same program also generated an unweighted total score for all examinees.

Results and Discussion

Discriminant functions. Results of discriminant function analyses conducted on the MSQ item responses from the Total Sample, Sample A, and Sample B are shown in Table 1.

Insert Table 1 About Here

Do the results of discriminant function analysis of item response data provide a generalizable scheme for weighting scored item responses? This problem was attended to by means of the previously described double cross-validation design. In Sample A, for example, the rank ordering of respondents, obtained when A's weights [derived at (2) in Figure 1] were applied to A's responses, was correlated with the rank ordering of the same respondents secured when B's weights [derived at (3) in Figure 1] were applied to A's responses. A high correlation would present evidence for the generalizability of the weighting procedure since either set of weights would produce the same ranking of examinees

(Ghiselli, 1965, p. 101; Gulliksen, 1950, pp. 314-315). Table 2 shows the results of such rank-order correlations for Samples A and B.

Insert Table 2 About Here

Moderate relationships were observed denoting a somewhat similar ranking of examinees but both correlations were negative. This indicated a distrubing trend: in one case low discriminant scores were associated with low job satisfaction but in the other case low discriminant scores were associated with high job satisfaction. The inconsistency of these results suggested that weights derived by discriminant function analyses on this set of MSQ item responses might not be generalizable to an independent sample of examinees selected from the same population. Giese (1965) conducted a similar study on another set of test items and found that the discriminant weights derived also failed to cross-validate.

A plausible explanation for these outcomes might be that test items, individually, are often not very valid or reliable even though the entire instrument may have high validity and reliability (Katzell, 1951). This individual test item instability may be due to sampling fluctuations or to the fact that, many times, test items are merely distant verbal corollaries of criterion behavior. Given this instability, it is not surprising that discriminant weights fail to cross-validate. If the sample size was large enough, say 800 to 1000 subjects (Katzell, 1951), and if test items were closer to work samples of their criteria (Cronbach, 1966, p. 55), then it might be possible to eradicate a large portion of this instability. Differential weighting techniques might be in a less tenuous position in these circumstances. However, few civilian testing situations can control these sources of error in a practical manner.

Does the application of a differential weighting scheme provided by discriminant function analysis help to differentiate between examinees better than an unweighted scoring of items? Using calculations made on the responses of the Total Sample [see (1) in Figure 1], the correlation between the ranking of all examinees provided by differentially weighting test item responses and the ranking derived from using unweighted responses was .91. At first glance, it would appear that one technique had little utility over the other. Since unweighted scores are easier to calculate, the decision might be to use the unweighted scoring technique. However, Stanely and Wang (1970b, p. 678) caution against the sole use of correlation for such decisions. A high correlation may very well denote a similar ranking of subjects but it does not indicate the decrement in criterion group overlap that may be afforded by the differential weighting procedure. Overlap has been defined as the percentage of persons in one group whose scores may be matched by persons in a second group (Elster and Dunnette, 1971, p. 686). In the Total Sample, the overlap between groups 1 and 2, as measured by Tilton's (1937) index, was 30% when unweighted scoring techniques were used, but was 19% when discriminant weights from Function I were applied to MSQ responses. Therefore, an increment in predictive efficiency of 11% might be realized with the differential weighting system. However, the veracity of this increment is doubtful due to the failure of the differential weighting system to cross-validate.

Implications

For practical implementation. In addition to the previously mentioned problems with the instability of test items and the crudeness of our measurement procedures, a weighted sum of test items is not very useful as the number of items approaches 30 (Gulliksen, 1950, p. 326). Since more than 30 items are

often used in most practical testing situations to ensure the collection of reliable information about examinees, discriminant function analysis appears to be of little practical value to test constructors in occupational education for deriving item response weights. However, Nunnally (1967, p. 278) has not found item weighting techniques to be impractical with tests composed of ten or fewer items. Some manipulative performance tests may require few tasks for completion and perhaps discriminant function analysis may be useful in weighting these tasks. Also, such tests of psychomotor ability often enjoy the benefits of being actual work samples of their criteria thus reducing some of the sources of instability encountered in this study.

For further research. There are several weighting techniques which may benefit test constructors in occupational education but at this time they remain unexplored. Latent trait measurement models (Lord & Novick, 1968, pp. 395-479), which represent radical departures from classical test theory, may potentially provide a valid source of item weights. Also, empirical techniques for weighting each multiple-choice alternative of a test item have been profitably studied (e.g. Hendrickson, 1970). Closely related to empirical option weighting methods are techniques which allow the examinee to express the confidence he has in the correctness of his test item response. Echternacht, Boldt, & Sellman (1971) present an example of the use of confidence weighting as a diagnostic aid in technical training.

Summary and Conclusions

Although the logic behind the use of multiple discriminant function analysis for deriving item response weighting schemes is compelling, empirical evidence fails to demonstrate the generalizability or utility of such procedures. It has been found, however, that very short tests may profit from item weighting techniques. Perhaps some tests of performance ability which contain few tasks

may benefit from this type of weighting scheme. But for most paper and pencil tests of even moderate length, item weighting techniques seem to be of little use to the test constructor in occupational education. Yet, research relating to empirical weighting of each multiple-choice alternative as well as the study of confidence weighting techniques may be profitable in the future. The use of latent trait measurement models may also prove a fruitful avenue for research. However, the following cautionary remarks made over twenty years ago by Katzell (1951) may add a historical perspective to the problem:

> "...the solutions to these questions [item weighting] do not lie in the realm of elegant analytical techniques...We should not be too unhappy with the turn of events, for in item analyses we are, after all, dealing with rather unstable and coarse data to which the application of highly sensitive methods would not be unlike casting pearls before swine (p. 17)."

References

Cronbach, L. J. New light on test strategy from decision theory. In A. Anastasi, (Ed.), Testing problems in perspective. Washington, D.C.: American Council on Education, 1966.

Elster, R. S. & Dunnette, M. D. The robustness of Tilton's measure of overlap. Educational and Psychological Measurement, 1971, 31 (3), 685-698.

Federico, Pat-Anthony. Identifying item validity indices utilizing a multivariate model. AFHRL-TR-71-16, Lowry AFB, Colorado: Technical Training Division, Air Force Human Resources Laboratory, April 1971.

Fisher, R.A. The use of multiple measurements in taxonomic problems. Annals of Eugenics, 1936, 8, 376-386.

Ghiselli, E. E. Theory of psychological measurement. New York: McGraw-Hill, 1964.

Giese, D. L. Comparison among item weighting techniques applied to the general college comprehensive examination. Unpublished doctoral dissertation, University of Minnesota, 1965.

Griess, J. Feasibility of providing trade competency exams for teachers on a national basis. Albany, New York: New York State Education Department, 1967, ED 021794.

Gulliksen, H. Theory of mental tests. New York: Wiley and Sons, 1950.

Hendrickson, G. F. An assessment of the effect of differentially weighting options within items of a multiple-choice objective test using a Guttman weighting scheme. Unpublished doctoral dissertation, The John Hopkins University, 1970.

Hoyt, C. J. Test reliability estimated by means of analysis of variance. Psychometrika, 1941, 6 (3), 153-160.

Katzell, R. A. Cross-validation of item analyses. Educational and Psychological Measurement, 1951, 11 (1), 16-22.

Lord, F. M. & Norvick, M. R. Statistical theories of mental test scores. Reading, Massachusetts: Addison-Wesley, 1968.

Mann, F. C. A study of work satisfaction as a function of the discrepancy between inferred aspirations and achievement. (Doctoral dissertation, University of Michigan) Ann Arbor, Michigan: University Microfilms, 1953, No. 5701.

Mosier, C. L. Problems and designs of cross-validation. Educational and Psychological Measurement, 1951, 11 (1), 5-11.

Echternacht, G.J., Boldt, R.F., & Sellman, W.S. A user's handbook for confidence testing as a diagnostic aid in technical training. AFHRL-TR-71-34, Lowry AFB, Colorado: Technical Training Division, Air Force Human Resources Laboratory, July 1971.

Nunnally, J. C. Psychometric theory. New York: McGraw-Hill, 1967.

Passmore, D. L. & Irvin, D. B. Program CROSVAL: A computerized scheme for cross-validation. Unpublished manuscript, University of Minnesota, 1972.

Smith, B. B., Passmore, D. L., Moss, J. & Copa, G. Project IMPROVE: A state-operated system for evaluation. Paper presented at the American Vocational Association Convention, December, 1971.

Stanley, J. C. & Wang, M. D. Weighting test items and test-item options, an overview of the analytical and empirical literature. Educational and Psychological Measurement, 1970a, 30, 21-35.

Strong, E. K. Vocational interests of men and women. Stanford, California: Stanford University Press, 1943.

Tatsuoka, M. M. Discriminant analysis. Champaign, Illinois: Institute for Personality and Ability Testing, 1970.

Tilton, J. W. The measurement of overlapping. Journal of Educational Psychology, 1937, 28, 656-662.

Veldman, D. FORTRAN programming for the behavioral sciences. New York: Holt, Rinehart, and Winston, 1967.

Wiess, D. J., Dawis, R. V., England, G. W. & Lofquist, L. H. Manual for the Minnesota Satisfaction Questionnaire. Minnesota studies in vocational rehabilitation: xxii, Minnespolis: Work Adjustment Project, Industrial Relations Center, University of Minnesota 1967.

Wilensky, H. L. Varieties of work experience. In H. Borow (Ed.), Man in a world of work. Boston: Houghton-Mifflin, 1964.

Footnotes

[1]Research Fellow in USOE fellowship program, "Preparing Researchers in Vocational Education", University of Minnesota. The assistance of the following University of Minnesota personnel is acknowledged: F. Marion Asche, USOE Research Fellow; Darwin Hendel, Work Adjustment Project; Dr. Cyril J. Hoyt, Educational Psychology.

Table 1

Results of Discriminant Function Analyses of MSQ Item Responses[a]

MSQ Item Number	Discriminant weights		
	Function I: Total Sample	Function II: Sample A	Function III: Sample B
1	.0268	-.0158	.0698
2	.2354	-.2209	.1372
3	.2697	-.2659	.1180
4	.1913	-.1190	.2086
5	-.0606	-.0904	-.1155
6	.0823	-.5426	.5347
7	.0652	.0133	-.1150
8	-.3326	.1961	-.2628
9	.1553	-.0184	.1763
10	-.1077	.0734	-.1083
11	.2290	-.1649	.1960
12	-.3975	.3232	-.1875
13	-.1718	.2694	-.0537
14	.1235	.0766	-.2006
15	-.0851	.0363	-.1737
16	.3810	-.1487	.4531
17	.2967	-.4072	.1490
18	-.0075	.0064	-.0483
19	-.2572	.1548	-.1445
20	-.3321	.3058	-.2563
21	-.0364	-.0475	-.1766

[a]Wilk's Lambda, an index of the degree of separation by the discriminant function (explained in Tatsuoka, 1970, pp. 22-24), was statistically significant for functions I, II, and III (.503, .309, .430 respectively at $p < .0001$ for each).

Table 2

Correlation Between the Ranking of Examinees
by Two Different Weighting Procedures

	A's weights on A's responses	B's weights on B's responses
A's weights on B's responses	—	-.63 **
B's weights on A's responses	-.74 *	—

* $r \neq 0$, $p < .005$, $df = 228$
** $r \neq 0$, $p < .005$, $df = 229$

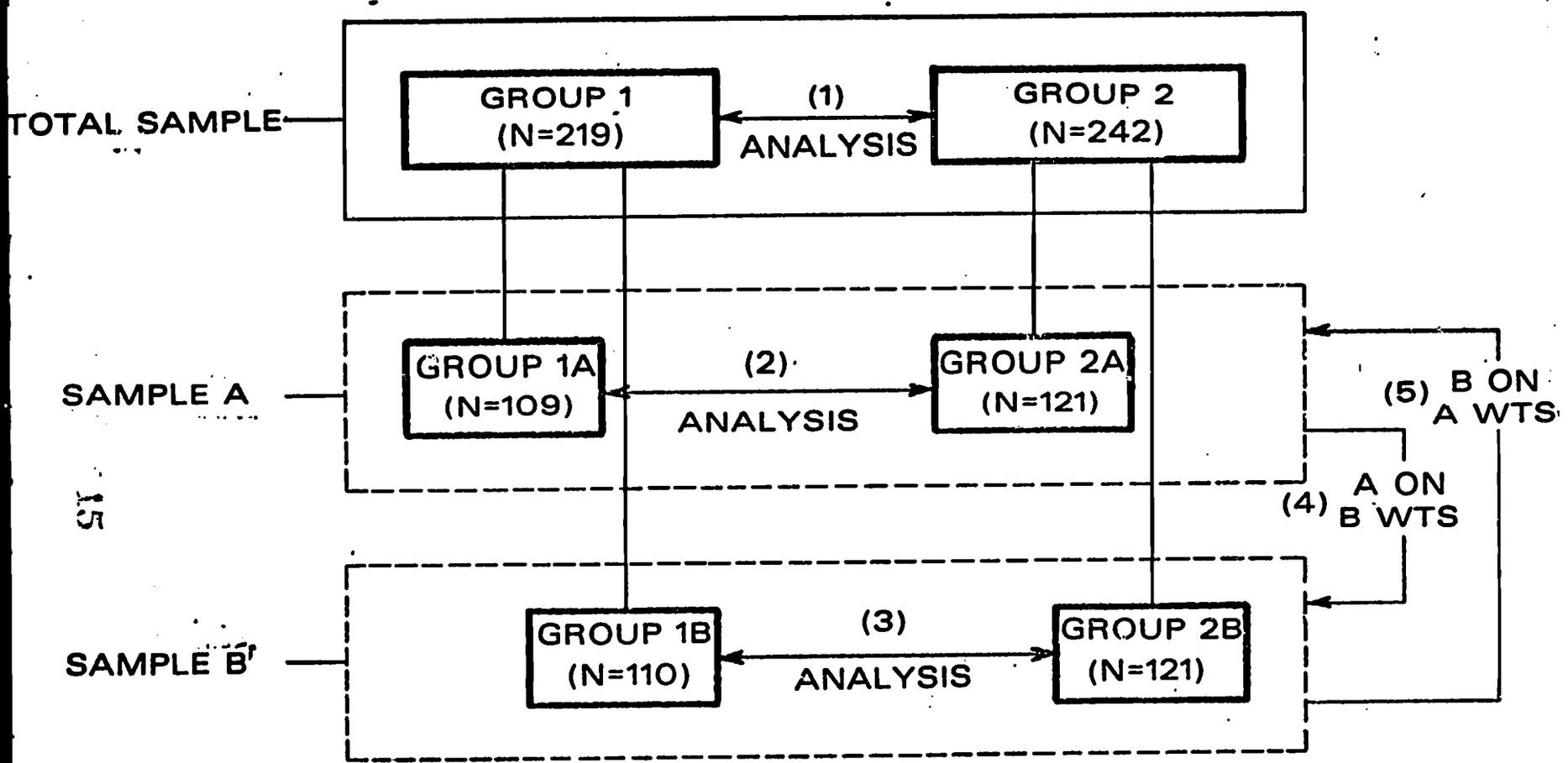

FIG. 1 DESIGN USED TO EXAMINE DIFFERENTIAL WEIGHTING OF TEST ITEM RESPONSE