DOCUMENT RESUME

ED 034 260 CG 004 692

AUTHOR	Pike, Lewis W.; Mahoney, Margaret H.
TITLE	Cross-Ethnic Cross Validation of Aptitude Batteries.
PUB DATE	[69]
NOTE	9p.
EDRS PRICE	EDRS Price MF-$0.25 HC-$0.55
DESCRIPTORS	*Aptitude Tests, Caucasians, Correlation, Culture Free Tests, *Ethnic Groups, Minority Groups, Negroes, Performance Criteria, Predictive Ability (Testing), Testing, *Test Validity

ABSTRACT

How well an aptitude test battery predicts rated job performance for Negroes and whites, and how well a battery selected for one group predicts performance for the other, is examined. Supervisory ratings were used as the criterion of job performance. Tests selected to predict performance in the job of Medical Laboratory technicians were validated separately by ethnic groups. Multiple correlation coefficients between the test battery and each of nine rating scales were computed and the resulting test batteries were cross validated, across ethnic group. Validity coefficents were generally higher for Negroes than for whites, and there were consistently higher validities for Negroes on paper and pencil tests assumed to be "culture bound" but higher validities for whites on tests assumed to be "culture free." On all nine rating scales, multiple correlations were greater for the Negro sample than for the white. The cross ethnic, cross validation indicated that a battery selected for a white sample would be generally valid for Negroes but the converse was less true. It was concluded that in some instances paper and pencil tests are as valid for Negroes as for whites even when weighted on a predominately white population. (Author/RM)

Cross-Ethnic Cross Validation of Aptitude Batteries

Lewis W. Pike

and

Margaret H. Mahoney

This paper examines how well a test battery predicts rated job performance separately for different ethnic groups, and how well a test battery selected for one ethnic group predicts performance for the other.

In employment testing there is increasing concern that predictors validated on a predominantly white group of job applicants may not be valid for Negroes or other minority groups. The concern applies particularly to pencil and paper tests.

In the general concern for fair employment testing, a procedure that has become widely recommended is that of validating tests for the job intended separately, by ethnic group. This procedure has been suggested by Guion (1966), and Krug (1966), among others.

Separate validation by ethnic group of tests selected to predict performance as Medical Laboratory technicians was a central part of the present study. The first paper this afternoon discussed the most direct index of test bias, the difference in intercepts of regression lines determined separately for the two ethnic groups. The present paper will focus more directly on the question of differential validity itself. The following questions will be considered: For the occupation in question, how do the validities of the several predictor tests compare, for the two ethnic groups? Further, how well does the battery of tests work for Negroes and for whites, using multiple correlation values based on optimal weightings for each group? And finally, how well will the battery weighted optimally for a white population work for a Negro group

(the most likely event, and the common concern) and vice versa, as indicated by cross-validating each set of weights, then applying them to persons of the opposite race?

Method

Both peer ratings and supervisory ratings were obtained as criteria, but only the supervisory ratings will be used in the present analysis. As you realize by now, from Dr. Flaugher's and Dr. Rock's presentations, we have serious reservations about accepting supervisors' averaged ratings as a criterion of job performance. Both the level of rating and the bases of rating will vary for both whites and Negroes, depending upon whether one or both raters is of the same ethnic group. Nevertheless, for this paper we are taking the ratings as given. When data are restricted to those within each rater-ratee ethnic combination, and including only those technicians rated by supervisors of both races, there is a serious restriction in sample size, and an increase in complexity of data reporting.

Results

Before considering the validity measures, it may be worth reviewing the general pattern of ratings and scores for the two ethnic groups. As you may recall, means were higher for whites on all but the first rating scale, but in every instance the difference was slight.

Mean scores for whites were higher than those for Negroes on all nine predictor tests, and in every instance the difference was significant at the .01 level.

Correlations

The correlations between predictor tests and supervisors' ratings of Negro and white subjects are given in Table 9. These are corrected for attenuation in the predictors as well as in the criterion scales. The present data are intended to show the validities potentially available in the predictors used, for predicting performance as medical technicians. The predictor tests were arbitrarily kept brief to allow for the collection of a variety of predictor, criterion, and background measures.

Upon examining the pairs of validity coefficients in Table 9 column by column, it may be seen that validities for Negroes were higher than those for whites in all instances for the first, second, fourth, and sixth tests. The reverse was true on the fifth and eighth tests, where 17 of the 18 correlations were greater for whites. Thus, the general expectation that pencil and paper tests are less valid for Negroes was certainly not borne out in the present instance.

The oft-voiced concern that school-oriented tests are less valid for Negroes than for whites also failed to hold for the population studied. Of the four tests having consistently higher validities for Negroes than for whites, two are computational--the Subtraction-Multiplication and the Necessary Arithmetic tests. Another is a test of vocabulary, and the last, Number Comparison, is a standard test of clerical ability. Tests that showed higher validities for whites, on the other hand, are the Fine Finger Dexterity test and the Picture-Number test. The latter is a test of short-term memory, and would seem a likely candidate for a "culture-fair" test.

It would be interesting to know how typical or atypical the above results may be. As you will recall, the subjects of the study were incumbent medical technicians, rather than job applicants. On the other hand, there

was not the usual problem of restriction of range due to testing, since the technicians studied had not been selected for their jobs on the basis of tests.

Multiple correlation coefficients

For each ethnic group, multiple correlations were computed for the best weighted combination of the nine experimental tests. These correlations are given in the first and third columns of Table 10, for whites and Negroes, respectively. In comparing the two sets of multiple correlations, note that for every rating scale, Negro weights applied to the Negro sample yielded a higher multiple correlation than did the white weights applied to whites. Note further that the lowest multiple R for Negroes, .29 on the Overall rating, was exceeded by only two of the multiple R's for whites, .38 on Learning Ability and .36 on Flexibility. The conclusion is strengthened, then, that a battery of objective pencil and paper tests is indeed relevant for blacks as well as whites, in predicting rated job performance.

The comparatively high multiple correlations for Negroes could have come from the relatively culture-free tests, of course, such as Picture-Number (testing short-term memory) or Finger Dexterity. Such was not the case, however. For nearly every scale, Subtraction-Multiplication and Necessary Arithmetic test scores were assigned the largest weights in the multiple correlations for Negroes. Picture-Number also appeared in several scales, but with a negative weight. For the white sample, Necessary Arithmetic again figured prominently, having the largest weight for five of the nine scales. Unlike the Negro multiple correlations, however, those for whites included sizable positive weightings on Finger Dexterity and Picture-Number scores.

Cross-validation coefficients

How well will a test battery selected for a white sample make generally valid predictions about Negroes, as well? For the data just presented, this question can be answered by applying the weights determined on the white sample, to obtain multiple correlations for the Negro sample. The cross-ethnic cross-validation coefficients resulting from doing this are given in the second column of Table 10. Similarly, the results of applying weights derived from Negro data to the white sample are given in the fourth column of the table.

When the weights determined on the white sample were applied to the Negro sample, five rating scales actually had higher multiples than they did for the white sample. This of course reflects the fact that the tests contributing to those multiples had higher validities for the Negroes than for the whites. Multiples for three of the four remaining scales dropped only slightly. Thus, it appears that a battery selected for a white sample will make generally valid predictions among Negroes, as well. The converse was less true, as is apparent upon examination of the last two columns in Table 10. On most scales, there was considerable shrinkage in the multiple correlation when weights derived for the Negro sample were applied to the whites.

Summary

Tests selected to predict rated performance in the job of Medical Technician were validated separately by ethnic group. Multiple correlation coefficients between the test battery and each of nine rating scales were next computed, using for each the optimal weights for Negroes and for whites. The resulting test batteries were then cross-validated, across ethnic group. That is, optimal weights derived for one ethnic group were then applied to the other.

Two conclusions may be reached from examining the differential validities. One, the belief that pencil and paper tests are generally less valid for Negroes than for whites was not supported by the present study. Validity coefficients were generally somewhat higher for the Negro group than for the whites. Two, there were consistently higher validities for Negroes than for whites on tests which might be considered culture bound, including Subtraction-Multiplication, Necessary Arithmetic, and Vocabulary, but higher validities for whites on tests one might assume to be "culture-free"--including Finger Dexterity and Picture-Number.

Evidence that the pencil and paper tests were as valid for the Negro subjects as for the whites, and that presumably culture-bound tests played as important a role compared to "culture-free" tests for Negroes as they did for whites, was even more pronounced when multiple correlations were examined. On all nine rating scales, multiple correlations computed for the Negro sample were greater than those computed for whites. Further, the more culture-bound tests such as Subtraction-Multiplication and Vocabulary were generally weighted more heavily for the Negro sample than for the white.

When applying multiple-regression principles to derive optimal weights for a test battery, cross-validation is important to check on applicability of the weights in new samples, even for the same population. Shrinkage is ordinarily expected. Even more important is the need to cross validate across ethnic groups, given the concern that there may be important differences, such that weights giving relatively high validity for one ethnic group may be entirely inappropriate for another.

Cross-ethnic cross validation of the weights derived from the white sample indicated that a test battery selected on this basis would be generally valid for Negroes, as well. The converse was less true. There was generally large attrition in multiple correlation when weights derived for the Negro sample were applied to whites.

The findings of this study are not, of course, an indication that problems due to differential validities between whites and Negroes do not exist. The present data are for a particular set of tests, used to predict success on a particular job, and using a particular criterion. The present findings do, however, indicate that in at least some instances, paper and pencil tests are as valid for Negroes as for whites, even when weighted on a predominantly white population.

Table 9

Correlations between predictor tests and supervisors' ratings on selected criterion scales, corrected for attenuation in criteria and predictors

Rating scale	Predictor test								
	1. Subtr-Mult	2. Vocab	3. Hidden Figure	4. Nec Arith	5. Finger Dext	6. Number Compar	7. Gestalt Compl	8. Pict Number	9. Paper Folding
Flexibility	30	-00	22	38	31	20	32	29	34
	48	16	06	46	19	22	20	-05	21
Planning	18	01	04	21	19	06	18	18	13
	51	16	05	34	14	24	10	-12	02
Interest	16	08	06	21	15	09	08	17	14
	40	14	05	27	05	10	04	-11	00
Learning Ability	30	09	17	40	32	21	25	27	38
	55	32	03	59	29	40	29	10	46
Job Knowledge	11	17	-01	16	12	-01	04	08	16
	41	27	10	49	11	24	19	-03	14
Technique	14	08	08	21	21	10	18	26	20
	37	21	06	35	10	23	11	-09	11
Low Need for Supervision	06	06	04	12	08	-01	04	12	08
	36	14	04	39	07	14	06	-00	14
Communication	08	22	11	17	07	01	04	07	13
	32	31	-03	35	08	20	18	-08	18
Overall	15	07	06	20	14	05	13	19	14
	40	13	03	26	13	24	07	-03	13

Note.-- In each pair of correlations, the upper and lower values are for the white and Negro sample, respectively.

Table 10

Multiple correlation coefficients and cross-ethnic cross validation coefficients for predicting supervisors' ratings from aptitude test scores

Rating scale	White weights		Negro weights	
	White sample (N = 297)	Negro sample (N = 166)	Negro sample (N = 166)	White sample (N = 297)
1. Flexibility	36	34	41	24
2. Organization	19	18	36	11
3. Interest	15	17	32	07
4. Learning Ability	38	40	42	32
5. Job Knowledge	17	21	42	13
6. Technique	23	-01	35	07
7. Low Need for Supervision	11	04	33	05
8. Communication	17	21	34	15
9. Overall	16	17	29	13