McKenzie, Robert C., US Office of Personnel Management, Washington, DC. (Thurs. P.M.)

The Problem of Range Restriction in Test Validation

This paper examines some of the legal arguments advanced over the years with respect to the interpretation of the validity coefficient. The issue as to whether or not the validity coefficient is of a magnitude large enough to have "practical significance" is given considerable attention. The Equal Employment Opportunity Commission has argued that the validity coefficient is usually too low to have practical significance. Educational testing Service (ETS), however, has contended that the validity coefficient does have practical significance but it should be corrected for restriction in range. The Department of Justice claimed that range restriction correction formulas cannot be used because the assumptions underlying them cannot be met. The Department of Justice in its defense cited the Division 14 Principles for Validation and Use of Personnel Selection Procedures, which asserts the desirability of having validation samples be as similar as possible to the applicant pool. Thorndike's range restriction correction formulas and their underlying assumptions are carefully reviewed in this paper. These formulas are applied to data on actual selection instruments to obtain the estimated true validity coefficient that would be obtained if the valdity coefficient was based on the total applicant population. Also, criterion referenced tests are discussed and suggested as viable alternatives to norm referenced tests, along with factors contributing to criterion biases.

The Problem of Range Restriction in Test Validation

Robert C. McKenzie
U.S. Office of Personnel Management
Washington, D.C. 20415

The opinions expressed in this paper are those of the author and do not reflect official policy of the Office of Personnel Management.

Finding ways to improve selection systems have been at the forefront of personnel psychology for many years. Very sophisticated selection systems have been developed over the years but most have lacked good validity. The most recent descision of the Justice Department with respect to the Professional Administrative Career Exam (P.A.C.E.) has called for more tests to be developed which have higher validity and less adverse impact on minority group members.

Boldt (1977) has pointed out the practical significance that the validity coefficient plays in the legal process through the EEOC guidelines. The EEOC guidelines asserted that test validity must not only be statistically significant but of a magnitude of which to suggest that the benefits obtained from using a particular selection device is worth the trouble of using it. Boldt has also pointed out that the educational testing service has asserted that the correlation coefficient was the appropiate statistics but should be corrected for restriction in range.

The Department of Justice claimed that correlations between selection devices and criteria should not be corrected for range restriction because the assumption of homoscedasticity, linearity, and normality cannot be met. This question, however, raises an important question as to which validity coefficient is the best, the uncorrected or corrected. The purpose of this paper is to shed some light on some of the problems associated with validity coefficients and provide an overview of three main cases of correcting for range restriction in testing.

Generally when a test is administered, it is administered to applicants who walk into a testing center and take a test. However, in some situations, the tester may have control over the applicant pool, but this situation very rarely occurs. Since the total number of people who take the test makeup the applicant pool, the normative information which is obtained from the sample should be based on the total applicant pool. Very seldom, however, are validity coefficient provided for the entire applicant group to whom the test is administered. Consequently, validity coefficients can only be obtained for those people who are selected on the job. When this occurs, the range of test scores becomes restricted, and the sample ceases to become a representative sample of the general population of applicants and thus cannot be generalized to the total applicant population.

Range restriction, as applied to test scores, is a general term which means that the test scores for a particular group are concentrated in only a portion of the possible range of scores (Kaufman, 1972). Groups that are restricted in range have smaller standard deviations than groups that are not restricted. Also, when test scores are restricted in range, correlation between a test and a criterion will be lower than the scores for the unrestricted groups.

When a high standard of selectivity is employed, the effect of range restriction on the resulting validity coefficients becomes even more profound. Thorndike's (1949) frequently cited wartime study is a good

illustration of this relationship. In this study, a battery of tests to predict success in pilot training was given to a large group of men as a part of an Army Air Force Aviation Psychology Program. In using the strict selection standards that were in effect toward the end of World War II, only 13% of these men would have qualified to enter pilot training on the basis of test scores. However, all of the men, regardless of their test scores, were allowed to enter training for experimental purposes. Table 1 below illustrates the correlation coefficients between test scores and the criterion for the total and qualified groups entering training.

TABLE 1

	Correlation with Criterion	
Predictor	Total Group (N = 1036)	Qualified Group (N = 136)
Pilot Stanine (Composite Score)	.64	.18
Mechanical Principles Test	.44	.03
Complex Coordination	.40	-.03
Instrumental Comprehension Test	.45	.27
Finger Dextrity Test	.18	.00
General Information Test	.46	.20
Arithmetic Reasoning Test	.27	.18

From Thorndike, 1949, page 171

These results show the effect that can occur on validity coefficients when restricted and unrestricted groups are compared. According to table 1, the composite aptitude score has the highest correlation (.64) when the total group is taken into consideration, but a substantially lower correlation (.18) when the group is restricted. Judging by the

qualified group alone, the Complex Coordination Test and the Mechanical Principles Test were among the worst predictors. However, for the total group, both the Mechanical Principles and Complex Coordination Tests were among the best predictors.

In order to make practical use of validity statistics for a restricted group, it is necessary to have statistical correction procedures to estimate what validity coefficients would have been obtained if it had been possible to obtain test and criterion data from a representative sample of those to whom the selection device was applied. Thorndike (1949) discussed three types of correction procedures for range restriction, Case I, Case II, and Case III.

Case I occurs when there is some degree of truncation on the criterion (Schmidt and Hunter, 1977; Gullisksen, 1950). In most practical situations, a test or multivariables such as exceptionally good recommendations and exceptionally good academic records are used for selection. However, there may also be situations wherein the criterion itself is used to select employees. For example, a manufacturing company may wish to develop a test to predict job performance and may accept all applicants regardless of their test scores, but weed out the bottom 50% whose performance falls below a given minimum standard. The Case I model is rarely used in validity studies because very seldom is selection made on the basis of the criterion. Selection is almost always made on some type of test. Nevertheless, this case can be put into practice by selecting on the criterion at a point which corresponds to the cutting score where selectees have been proven to be successful. Figure 1 is an illustration of selection occurring on the criterion.

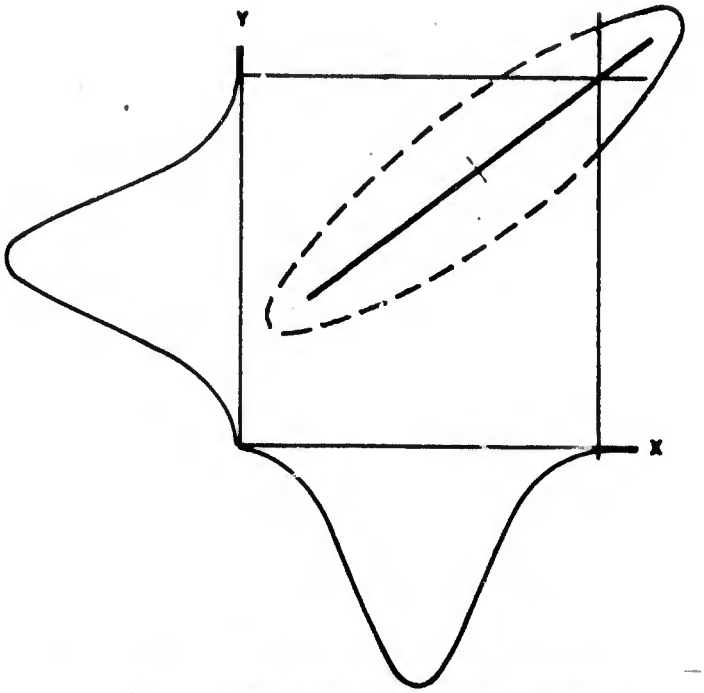

Figure 1. is an illustration of selection occurring on the criterion (Y).

Let Y represent the criterion, and X the predictor. Suppose that the selection ratio is set at .05, then if the assumptions of linearity and homoescedacity hold, setting a cutting score at the top 5% on the criterion will correspond to the top 5% if the test were used to select employees. The dotted and solid line in figure 1 shows the ellipse for the total unrestricted group, and the solid line represents the ellipse for the restricted group. The higher the selection ratio the more severe the restriction in range. Essentially, This is a case of regression of X on Y, instead of the more typical case of regression of Y on X. While this approach tends to be somewhat costly, it does offer an alternative to written tests.

Case II

Case II situation occurs when there is direct truncation on the predictor variable. This model leads to underestimates of the effect of range restriction, if selection is partly on the basis of the criterion as well as the predictor. However, the Case II formula provides slight overestimates of the corrected coefficients when truncation on the test is not perfect. For example, when some applicants below the cutoff score are selected, or selection is made on the basis of the sums of scores on two or more tests rather than a single test (Schmidt and Hunter, 1977).

Table 2 was developed by using the Case II formula (appendix A), and facilitates the determination of R_{12}, the estimated correlation between predictor and criterion in an unrestricted sample.

TABLE 2

Validity Coefficients for Unrestricted Group (R_{12}) Estimated From Values for Restricted Group (r_{12})

S_1 / s_1	$r_{12} \rightarrow$ .10	.15	.20	.25	.30	.35	.40	.45
1.25	.12	.19	.25	.31	.37	.42	.48	.53
1.50	.15	.22	.29	.36	.43	.49	.55	.60
1.75	.17	.26	.34	.41	.48	.55	.61	.66
2.00	.20	.29	.38	.46	.53	.60	.66	.71
2.50	.24	.35	.45	.54	.62	.68	.74	.78
3.00	.29	.41	.52	.61	.69	.75	.79	.83
4.00	.37	.52	.63	.72	.78	.83	.87	.90
5.00	.45	.60	.71	.79	.84	.88	.91	.93
10.00	.71	.83	90	.93	.95	.97	.97	.98

From Kaufman, 1972, page 6.

$\frac{S_1}{s_1}$ is the ratio of the standard deviation of the unrestricted group to the standard deviation of the restricted group on the predictor test; r_{12} is the actual obtained validity coefficient for the unrestricted group. The unrestricted coefficient is found by looking up the ratio of the unrestricted standard deviation to the restricted coefficient at the top of the table. For example, if the ratio of the standard deviation was found to be 1.25, and the restricted validity coeficient was .25, then table 2 would be entered as follows:

$\frac{S_1}{s_1} = 1.25$, $r_{12} = .25$, which results in an unrestricted validity coefficient of (R_{12}) .31.

The magnitude of a pearson product moment correlation coefficient is directly related to the standard deviation of the two variables being correlated. A reduction in either or both of the standard deviations will lower the correlation coefficient between the two variables.

Case III

The third case of range restriction occurs when there is truncation on a third variable. Gullisksen (1950) distinquishes between explicit selection and incidental selection. Explicit selection is defined as direct selection occurring on the basis of a given variable (test), and incidental selection occurs when there is indirect selection occurring on the basis of a given variable (criterion) or another test which is highly correlated with the explicit variable. For instance, suppose a researcher is interested in trying out a new test (Y) to see how well it predicts job performance, and the scores are available for the first test (X) which was

used for selection. Selection is then incidental with respect to test Y (new test) because of the high correlation between the new test and the initial test, thus selecting on the first test is basically the same as selecting on the second test. Restriction in range occurs on the new research test because there will be some applicants who passed the first test who will not pass the second test. The Case III situation is more practical than the Case I situation and is very frequent in concurrent validity studies.

Thorndike (1949) discussed another Case III in which one of the correlations used in the above formula maybe available for the total group rather than the restricted group. For example, a research test may have been given to a general unselected group and its correlation with the score on the selection test is based on this group (see appendix A for Case III formula).

These formulas are basically used for pearson-product moment correlations based on continuous variables; however, if the formula for biserial correlation is to be used in restricted groups, the distribution of traits underlying that dichotomy must be normally distributed in the restricted groups (Thorndike, 1949).

Campbell (1976) and Linn (1968) have pointed out that attempting to estimate reliabilities and validities in the appropriate population can pose a serious problem. For instance, they argue that the assumptions of linearity and homoscedasticity (equality of conditional variances) may easily be violated. For example, in a bivariate distribution, much less variability tends to be exhibited at the extremes as opposed to the middle, and so rather than being linear, the standard score of the regression line tends to be steeper at the ends than in the middle. The violation of the homoscedasticity assumption tends to inflate the corrections, while the departure from linearity tends to deflate it. Linn (1968) has shown that by correcting correlation coefficients as if the test (x)

is the explicit selection variable, results into the formula overcorrecting when the correlation between x (test 1) and z (criterion) is low and undercorrecting when the correlation is in the middle range. The undercorrection and overcorrection phenomena becomes more profound as the degree of actual selection on z (unavailable predictor) becomes extreme. This relationship is shown in figure 2 below which was abstracted from the Linn (1968) report.

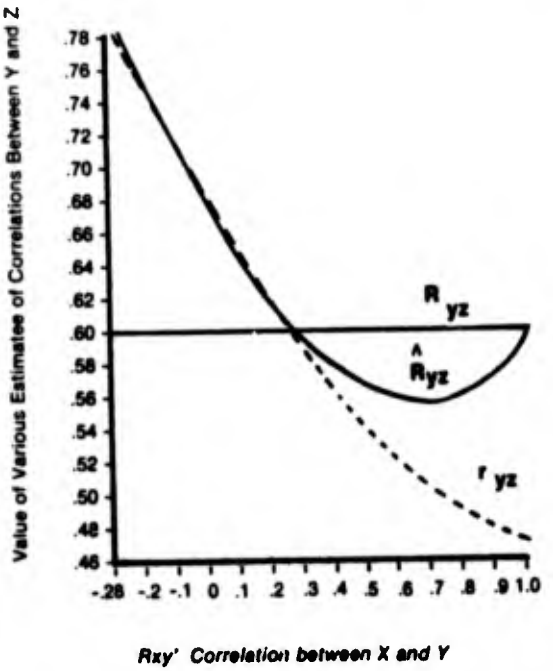

Figure 2. An illustration of the effects of correction of range as a function of the correlation between the explicit and implicit predictors.

Let x be the explicit selection variable (test used for selection), y the implicit selection variable (a second test), and z the criterion but also subject to implicit selection. Then the correlation between y and z in the selected group (r_{yz}) is a function of the correlation between x and y in the unrestricted group (R_{xy}). The broken line in the figure 2 shows the value of r_{yz} in the restricted population as a function of R_{xy} for values of R_{xy}. If y is treated as an explicit selection variable and r_{yz} is corrected for homogeniety of variances of y in the restricted population,

the corrected values of $\hat{R}_{yz}$ will still be an underestimate of R_{yz} as long as x and y are reasonably correlated.

A number of studies have shown that the assumptions of linearity, normality and homoscedasticity are very rarely violated. Sevier (1957), using sample sizes ranging from 105 to 250, has shown that out of 24 tests, only one violated the assumption of linearity, and one out of eight violated the assumption of homoescedasticity. Ghiselli and Kahneman (1962) examined 60 aptitude variables in a sample size of 200 and showed that 40 percent of the variables departed significantly from the linear-homoescedastic model. However, ninety percent of these variables held up when cross validated. Tupes (1964) re-analyzed the Ghiselli and Kahneman studies and found that only 10 percent of these relationships departed from linearity at the .05 level.

Schmidt (1978) has pointed out that the assumptions for range restriction correction formulas in real data will be violated only infrequently. However, even if they are violated (e.g., lower conditional criterion variances in the restricted group), there is every reason to believe that the amount of induced bias will be small relative to massive bias induced by failure to correct. This point is illustrated by re-examining the figure abstracted from the Linn (1968) article[1]. Campbell (1976), however, reviewed the Linn article and came to the same conclusion as did Linn. He then further emphasized that these kinds of consideration make using correction

---

1. As noted by Schmidt (1978), the Linn article is very confusing because figure 2 has been called figure 1, and vice versa. However, by considering the example on which figure 1 is based, a very serious violation of the assumption is made. In this example, restriction has been on a third variable (X), and so the appropiate model is Thorndike's (1949) Case III formula. However variable X is unknown and thus the correct correction formula cannot be used. The incorrect Case II model is used instead, which assumes direct restriction on the predictor.

formulas rather risky business. In almost every real situation an estimate of the population parameter will most likely be biased in some respect, except, in clear and straightforward situations.

Criterion-Referenced Tests

Perhaps a better way to insure that selection systems have validity is to make sure that the test is an adequate sample of the job content. Criterion-referenced tests although less sophisticated than norm-referenced tests provide better samples of the job domain. These tests are designed to measure performance relative to a criterion standard. That is, an employee's test performance is compared to an established criterion performance standard. Normative data as to how well one employee is performing relative to another is not necessary. The job is usually simulated with 90% accuracy, so that the performance on the test is basically the same as the actual performance on the criterion. Criterion-referenced test are basically the same as performance tests.

Job performance tests such as work sample tests are especially applicable as use for criterion-referenced measures (Buck, 1975). These tests like criterion-referenced tests, must sample the job domain accurately and also must measure the employee's ability to perform critical job tasks. As Buck (1975) has pointed out, criterion referenced tests are often confused with criterion related tests. A criterion related test implies that there is some kind of statistical relationship between a test and some measure of job performance (i.e., supervisory ratings, self ratings, peer ratings, etc.). While on the other hand, criterion-referenced tests refer to the minimal acceptable level that an applicant must meet in order to achieve a mastery level on the job. Essentially, criterion-referenced tests are used to determine who is qualified, they do not measure variability in performance. Popham and Husek (1971) contend

that criterion-referenced tests are only suitable when there is no constraints on how many people are to be selected for a particular job. However, when there are constraints, norm-referenced tests are usually more appropiate for selection. Like norm-referenced tests, criterion-referenced tests can also be used when the applicant pool is exceptionally large by setting the cutting score above the minimally qualifying standard. For example, if the minimally qualifying standard cut score is set at 70%, it can be set at 90% which would insure the adaptation of a treatment that would select superior employees. The advantages of this aproach of testing over norm-referenced tests is the accuracy with which the job domain is sampled and the fact that less complex tests can be developed which in turn would give better measures of validity.

Reference Notes

1. Schmidt, F.L. Personal communication, January 1978

References

Boldt, R.F. Robustness of range restriction in court, Unpublished report, 1978.

Buck, L.S. Use of criterion-referenced tests in personnel selection: A Summary Status Report. Technical Memorandum, 1975 6

Campbell, J.P. Psychometric theory. Handbook of Industrial and Organizational Psychology, 1976.

Gullisksen, H. Theory of Mental Tests. New York: McGraw Hill, 1956.

Popham, W.J., & Husek, T.A. Implications of criterion-referenced measurement. In W.J. Popham (ED.), Criterion-referenced measurement (An introduction). New Jersey: Educational Publication, Inc., 1971.

Kaufman, A. S. Restriction of Range: Questions and Answers. Test Services Bulletin, The Psychological Corporation, 1972.

Levin, J. The occurence of an increase in correlation by restriction of range. Psychometrika, 1972 37, 93-97.

Linn, Robert L. Range restriction Problems in the use of selected groups for test validation. Psychological Bulletin, 1968, 69, 69-73.

Schmidt, F.L. & Hunter, John E. The development of a general solution of the problem of validity generalization. Journal of Applied Psychology, 36, 1977.

Thorndike, Robert L. Personnel Selection. Test & Measurement Techniques, New York, 1949.

Ghiselli, E.E. & Kahneman, D. Validity and non-linear heteroscedastic models. Personnel Psychology, 15 , 1962,

Sevier, F.C. Testing the assumptions of underlying multiple regression. Journal of Experimental Education, 1957, 25, 323-330.

Tupes, E.C. A note on "Validity and Non-linear heteroscedastic models". Personnel Psychology, 17, 1964.

APPENDIX A

Thorndike's Case I formula:

$$R_{xy} = \sqrt{1 - \frac{s_y^2}{S_y^2}(1 - r_{xy}^2)}$$

The Case I formula is primarily concerned with the correlation between x and y. Where R_{xy} is the correlation between a test (x) and a criterion (y) in the unrestricted group, Sy is the standard deviation of the criterion in the unrestricted group, sy is the standard deviation of the criterion in the restricted group.

Thorndike's Case II formula:

$$R_{xy} = \frac{r_{xy}\frac{S_x}{s_x}}{\sqrt{1 - r_{xy}^2 + r_{xy}^2\frac{S_x^2}{s_x^2}}}$$

A special case of Case III when the validity coefficient is known for the unrestricted group:

$$R_{xy} = \frac{r_{xy}\sqrt{1 + R_{xz}^2\left(\frac{s_z^2}{s_z^2} - 1\right)} + R_{xz}\, r_{yz}\left(\frac{s_z}{s_z} - \frac{s_z}{s_z}\right)}{\sqrt{1 + r_{yz}^2\left(\frac{s_z^2}{s_z^2} - 1\right)}}$$

The terms are define in the Case I formula.

Thorndike's Case III formula:

$$R_{xy} = \frac{r_{xy} + r_{xz} + r_{yz}\left(\frac{S_z^2}{s_z^2} - 1\right)}{\sqrt{\left[1 + r_{xz}^2\left(\frac{S_z^2}{s_z^2} - 1\right)\right]\left[1 + r_{xz}^2\left(\frac{S_z^2}{s_z^2} - 1\right)\right]}}$$

These symbols are analogous to the Case II formula, with the exception of the third variable being correlated. R_{xy} is the correlation between the explicit variable (test) and the implicit variable (a new research test) for the unrestricted group, r_{xy} is the correlation between the explicit variable and the implicit variable for the restricted group, r_{xz} is the correlation between test X and the criterion Z for the restricted group, r_{yz} is the correlation between implicit variable (new research test) and the implicit variable (criterion Z for the restricted group, and S_z^2 and s_z^2 are the unrestricted and restricted group variances for the criterion Z.