DOCUMENT RESUME

ED 255 554 TM 850 195

AUTHOR Schmidt, Frank L.
TITLE From Validity Generalization to Meta-Analysis: The Development and Application of a New Research Integration Procedure.
PUB DATE 2 Apr 85
NOTE 16p.; Paper presented at the Annual Meeting of the American Educational Research Association (69th, Chicago, IL, March 31-April 4, 1985).
PUB TYPE Speeches/Conference Papers (150) -- Reports - Research/Technical (143)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS Effect Size; Error of Measurement; Estimation (Mathematics); *Generalizability Theory; *Meta Analysis; Occupational Tests; Test Validity
IDENTIFIERS Glass (GV); Office of Personnel Management

ABSTRACT

This paper describes how work by the United States Office of Personnel Management on the generalizability of employment test validities led to the development of a widely applicable meta-analysis method. The method focuses strongly on estimating the true variance of study correlations and effect size. This validity generalization procedure has been applied to over 500 research literatures in employment selection--each one representing a predictor-job performance combination. This procedure goes beyond the Glass methods of meta-analysis in providing methods for correcting both variances and means of correlations or effect sizes for the distorting effects of the artifacts of sampling error, measurement error, and range restriction. These meta-analysis methods have been successfully applied to non-employment selection topics. Twenty-seven references are listed. (BS)

FROM VALIDITY GENERALIZATION TO META-ANALYSIS: THE DEVELOPMENT AND APPLICATION OF A NEW RESEARCH INTEGRATION PROCEDURE

Frank L. Schmidt
U.S. Office of Personnel Management
and
George Washington University

March, 1985

Presented as part of the symposium "Validity Generalization as Meta-Analysis." M. J. Burke, Chair. AERA Annual Meeting, Chicago, April 2, 1985

From Validity Generalization to Meta-Analysis: The Development and Application of a New Research Integration Procedure

My purpose is to describe for you how our work on the question of the generalizability of employment test validities led to the development of a method of meta-analysis that is (a) widely applicable to research literatures and (b) different in important respects from the meta-analysis methods of Glass and his associates.

General History

We began our work on validity generalization in 1975, and in 1976, Jack Hunter and I were given the Cattell Research Design award for the method and its initial application to a data set. We published a description of our methods the following year. Glass published the first article on meta-analysis in 1976, and during 1977 Lee J. Cronbach, in correspondence with us, drew our attention to Glass' work. He suggested that our methods, like those of Glass, were potentially general and could be applied to research literatures in many different areas of the social and behavioral sciences. Our book on general meta-analysis methods was published in 1982 (Hunter, Schmidt, and Jackson, 1982), one year after the Glass, McGaw, and Smith (1981) book. Both Glass' methods and our methods are now being widely used in a variety of different areas of the behavioral and social sciences.

We initially developed our meta-analysis methods not as general research integration methods, but as a way of attacking a critically important problem in personnel psychology: the problem of "situational specificity" of employment test validities. For over 50 years, most personnel psychologists

had believed that employment test validities were specific to situations and settings, and that therefore every test had to be revalidated anew in every setting in which it was considered for use. This belief was based on the empirical fact that considerable variability was present from study to study in observed validity coefficients even when the jobs and tests studied appeared to be similar or identical. The explanation developed for this variability was that the factor structure of job performance was different from job to job and that the human observer or job analyst too poor an information receiver and processor to detect these subtle but important differences. The conclusion was that validity studies must be conducted--typically at considerable expense--in every setting. That is, the conclusion was that validity evidence could not be generalized across settings. Lawshe (1948) stated:

> A given test may be excellent in connection with one job and virtually useless in connection with another job. Furthermore, job classifications that seem similar from plant to plant sometimes differ significantly; so it becomes essential to test the test in practically every new situation (p. 13).

And in the words of Albright, Glennon, and Smith (1963):

> If years of personnel research have proven anything, it is that jobs that seem the same from one place to another often differ in subtle but important ways. Not surprisingly, it follows that what constitutes job success is also likely to vary from place to place. (p. 18).

The fact that our point of departure was the problem of situational specificity explains why our methods of meta-analysis are focused strongly on estimation of the true (i.e., nonartifactual) variance of study correlations and effect sizes. We hypothesized that most or all of the variance in test

validity coefficients across studies and settings was due to artifactual sources such as sampling error, and not to real differences between jobs. This focus on the variance of effect sizes and correlations is the primary difference between our methods and those of Glass. In validity generalization, merely showing that the mean is substantial is not sufficient to demonstrate generalizability. One must be able to show that the standard deviation of true validities is small enough to permit generalization of the conclusion that the test has positive validity in the great majority of situations. Figures 1A and 1C illustrate this point.

None of this means that we were unconcerned with accurate estimation of the mean. Accurate estimation of mean true validities is critical because the mean affects both generalizability (by affecting the lower credibility value) and expected practical utility. Practical utility is a direct multiplicative function of the expected operational validity, other things equal. Therefore, we introduced methods for correcting the mean observed validity for attentuation due to mean levels of range restriction and mean levels of measurement error in the measures of job performance. These corrections also differentiate our methods from those of Glass and his associates.

In 1978, simulation studies by Callender and Osburn drew our attention to an error in our 1977 formula for estimating the standard deviation of true validities. In 1979, we published a corrected estimation formula, along with applications to new data sets. The following year Callender and Osburn (1980) published a new formula for estimating this value--their multiplicative formula. They used simulation methods to show that this formula, our corrected original formula, and a second formula we had

developed (the interactive formula) all provided estimates that were quite accurate. In 1983, Raju and Burke published two additional formulas for estimating the $\underline{SD}_{\rho}$ validities corrected for artifacts. These formulas, based on Taylor series approximations, were also shown by computer simulation studies to be quite accurate. Thus, there are now at least five accurate equations available for estimating $\underline{SD}_{\rho}$.

The differences between these equations center on the estimation of between-study variance due to test and criterion reliability differences and due to range restriction differences. In typical validity studies, the sample size is small (e.g., 50-100), and as a result most of the observed variance in $\underline{r}$'s is due to simple sampling error (typically, from 60% to 100%). Usually less than 10% is due to differences between studies in measurement error and in range restriction. In fact, we have repeatedly found that correcting for sampling error alone leads to the same conclusions about validity generalizability as correcting for all four artifactual sources of variance (e.g., see Schmidt, Gast-Rosenberg, & Hunter, 1981; Pearlman et al., 1980).

In certain data sets--particularly those from the military--mean sample sizes are much larger (e.g., 300-400), and therefore total (or observed) variance is much smaller. In these cases, the percentage of observed variance due to measurement error and range restriction differences is larger. However, the amount of such variance remains very small. The percentage is higher only because the total (observed) variance is quite small.

Other Artifacts

Artifacts other than sampling error and differences between studies

in measurement error and in range restriction can cause variance in study outcomes. Computational, typographical, transcriptional, and computer program errors may be important sources of artifactual variance in many validity coefficient sets. No method or equation for estimating variance due to such sources has yet been devised. Therefore it is to be expected that even when all variance is in fact artifactual, meta-analysis will indicate that less than 100% of observed variance is due to artifacts. That is, the methods now available for estimating artifactual variance error on the conservative side.

Applications and Impact

To date, the validity generalization procedure has been applied to over 500 research literatures in employment selection, each one representing a predictor-job performance combination. These predictors have included nontest procedures such as evaluations of education and experience and interviews, as well as ability and aptitude tests. In many cases, artifacts accounted for all variance across studies; the average amount of variance accounted for by artifacts has been approximately 80%. As an example, consider the relation between quantitative ability and overall job performance in clerical jobs (Pearlman et al., 1980). This substudy was based on 453 correlations computed on a total of 39,584 people. Seventy-seven percent of the variance in observed validities was traceable to artifacts, leaving a neglible variance of .019. The mean effect size was 47. Thus, integration of this massive amount of data leads to the general and generalizable principle that the correlation between quantitative ability and clerical performance is .47, with very little, if any, true variation around this

value. Findings like this show the old belief that validities are situationally specific to be false and show that cumulative, generalizable knowledge is possible.

Today many organizations--including the Federal government, the U.S. Employment Service, and some large corporations--use validity generalization findings as the basis of their selection testing programs. Validity generalization has been included in the recently adopted 1985 APA-AERA-NCME Standards for Educational and Psychological Tests.

Comparison of Meta-Analysis Methods

Gene Glass originated the term "meta-analysis" and advanced the first formal methods for meta-analysis. These methods are composed of the following steps:

1. Effect sizes are expressed in SD units or in correlatior form.
2. The mean effect size across studies is computed. This represents the expected magnitude of a treatment condition or the expected size of a correlation. For example, Smith and Glass (1977) found psychotherapy has an average effect size of .68 standard deviation units of the control group. White (1976) found that the mean correlation between SES and academic achievement is .25.
3. Properties or characteristics on which studies differ are coded and then correlated with effect sizes in an effort to find the causes of differences between studies in reported effect sizes. Studies may differ on age or sex of subjects, methodology used, and other variables. Although numerous variables are typically coded, the general finding has been that few are correlated with

study outcomes. Because the sample size for this type of analysis is the number of studies--not the number of people--there are often severe problems of capitalization on chance and low statistical power. For example, there may be 70 studies and 50 study characteristics.

Glass and his associates have applied his methods of meta-analysis to a variety of heretofore confused research literatures. In almost every case, the research literature has been clarified and general principles have been established. As one example, Glass and Smith (1979) have applied meta-analysis to the vast, conflicting, and heretofore uninterpretable literature on the effects of class size on pupil achievement. Based on 725 studies, their results revealed a very definite monotonic relation between class size and achievement, with the achievement difference ranging up to .90 SD units for the smallest (N = 1) vs. the largest (N = 40) classes. Further, the effect sizes were larger for the better controlled studies.

Our meta-analysis procedures go beyond the Glass methods in providing methods for correcting both variances and means of correlations or effect sizes for the distorting effects of the artifacts of sampling error, measurement error, and range restriction. Steps in this procedure are:

1. Effect sizes are expressed as correlations or d-values and the (sample-size-weighted) average effect size is computed across studies. This mean effect size is then corrected for the attenuating effects of instrument unreliability and range restrictions. This latter is a step not included in Glassian meta-analysis.

2. One then determines whether the variance in effect sizes across studies is due solely to statistical and measurement artifacts. This step is also not included in Glassian meta-analysis. If one can reject the hypothesis that the observed variance of effect sizes is greater than the variance expected from artifacts, one concludes that the mean corrected effect size estimates the true effect size, and a general principle has been established. The mean corrected effect size then incorporates and summarizes the results of all previous studies.
3. If one <u>cannot</u> reject the hypothesis that the variance of effect sizes is greater than the expected from artifacts, one then determines whether any of the study characteristics are correlated with effect size. Here the focus should be on theoretically meaningful moderators. In areas outside of employment testing, there are often sound theoretical reasons for expecting moderators. This step we borrowed from Glass and his associates (while recognizing and warning against the severe problems of capitalization on change and low statistical power).
4. If the remaining variance is still too large to be accounted for by artifacts, it is adjusted for the effects of these artifacts, and this adjusted variance is used to set confidence or credibility intervals around the mean effect size. Again, this is a step not included in Glassian meta-analysis.

Our meta-analysis procedures have now been applied to numerous topics outside the area of validity generalization in employment selection. Some examples include:

1. Correlates of role conflict and role ambiguity (Fisher, Gittelson, 1984, and Jackson & Schuler, in press).

2. Effects of realistic job previews (Premack and Wanous, 1984; Cascio and McEvoy, 1984).
3. Evaluation of Fieldler's theory of leadership (Peters, et al., in press).
4. Accuracy of self-ratings of ability and skill (Mabe & West, 1982).
5. Relation of LSAT scores to performance in law schools (Linn & Dunbar, 1981).
6. Relation of job satisfaction to absenteeism (Terborg, et al., 1982).
7. Ability of financial analysts to predict stock growth (Coggin & Hunter, 1983).
8. Premorbid functioning and recidivism in Schizophrenia (Stoffelmeyr, Dillavou, & Hunter, 1983).

In some of these non-employment selection applications, the results have been similar to those for employment tests: most or all of the observed between-study variance in effect sizes or correlations has been found to be due to statistical artifacts (principally sampling error). However, in other cases, considerable variance has remained after correcting for the effects of artifacts, indicating the appropriateness of moderator analyses. And in many of these cases, the subsequent moderator analysis has provided evidence for theoretically predicted and meaningful moderators.

References

Albright, L. E., Glennon, J. R., & Smith, W. J. (1963). The use of psychological tests in industry. Cleveland, OH: Howard Allen.

Callender, J. C., & Osburn, H. G. (1980). Development and test of a new model for validity generalization. Journal of Applied Psychology, 65, 543-558.

Cascio, W. F., & McEvoy, G. M. Extension of utility analysis to turnover reduction strategies. Presented in the Symposium "Overcoming the Futilities of Utility Applications. (Steven Wroten, Chair) American Psychological Association Convention, Toronto, Canada, August 1984.

Coggin, T. D., & Hunter, J. E. (1983). Problems in measuring the quality of investment information: The perils of the information coefficient. Financial Analysts Journal, May/June, 3-10.

Cooper, H. (1984). The integrative research review: A social science approach. Beverly Hills: Sage.

Glass, G. V. (1976). Primary, secondary, and meta-analysis of research. Educational Researcher, 5, 3-8.

Glass, G. V, McGaw, B., & Smith, M. L. (1981). Meta-analysis in social research. Beverly Hills, CA: Sage Publications.

Glass, G. V, & Smith, M. L. Meta-analysis of research on the relationship of class-size and achievement. Evaluation and Policy Analysis, 1979, 1, 2-15.

Hunter, J. E., Schmidt, F. L., & Jackson, G. B. (1982). Meta-analysis: ..ing research findings across studies. Beverly Hills, CA: Sage Publications.

Fisher, C. D., & Gittelson, R. (1983). A meta-analysis of the correlates of role conflict and ambiguity. Journal of Applied Psychology, 68, 320-333.

Jackson, S. E., & Schuler, R. S. (in press). A meta-analysis and conceptual critique of research on role ambiguity and role conflict in work settings. Organizational Behavior and Human Performance.

Lawshe, C. H. (1948). Principles of personnel testing. New York: Mcgraw-Hill.

Linn, R. L., Harnisch, D., & Dunbar, S. B. (1981). Validity generalization and situational specificity: An analysis of the prediction of first year grades in law school. Applied Psychological Measurement, 5, 281-289.

Mabe, P. A., III, & West, S. G. Validity of self-evaluations of ability: A review and meta-analysis. Journal of Applied Psychology, 1982, 67, 280-296.

Pearlman, K., Schmidt, F. L., & Hunter, J. E. (1980). Validity generalization results for tests used to predict training success and job proficiency in clerical occupations. Journal of Applied Psychology, 65, 373-406.

Peters, L. H., Harthe, D., & Pohlman, J. (1984). Fieldler's contingency theory of leadership: An application of the meta-analysis procedures of Schmidt and Hunter. Psychological Bulletin, in press.

Premack, S. & Wanous, J. P. (1984). Meta-analysis of realistic job preview experiments. Presented as part of the Symposium, "Use of meta-analysis in industrial/organizational psychology." Hannah R. Hirsh, Chair. The 92nd Annual APA Convention, Toronto, Canada, August 27.

Raju, N. S., & Burke, M. J. (1983). Two new procedures for studying validity generalization. Journal of Applied Psychology, 68, 382-395.

Rosenthal, R. (1984). Meta-analysis procedures for social research. Beverly Hills, CA: Sage.

Schmidt, F. L., & Hunter, J. E. (1977). Development of a general solution to the problem of validity generalization. Journal of Applied Psychology, 62, 529-540.

Schmidt, F. L., Hunter, J. E., Pearlman, K., & Shane, G. S. (1979). Further tests of the Schmidt-Hunter Bayesian validity generalization procedure. Personnel Psychology, 32, 257-281.

Schmidt, F. L., Hunter, J. E., & Pearlman K. (1981). Task differences and validity of aptitude tests in selection: A red herring. Journal of Applied Psychology, 66, 166-185.

Schmidt, F. L., & Hunter, J. E. (1981). Employment testing: Old theories and new research findings. American Psychologist, 36, 1128-1137.

Smith, M. L., & Glass G. V. Meta-analyses of psychotherapy outcome studies. American Psychologist. 1977, 32, 752-760.

Stoffelmeyr, B. E., Dillavou, D., & Hunter, J. E. (1983). Premorbid functioning and recidivism in schizophrenia: A cumulative analysis. Journal of Consulting and Clinical Psychology, 1983, 51, 338-352.

Terborg, J. R., & Lee, T. W. Extension of the Schmidt-Hunter validity generalization procedure to the prediction of absenteeism behavior from knowledge of job satisfaction and organizational commitment. Journal of Applied Psychology. 1982, 67, 280-296.

White, K. R. (1982). The relation between socio-economic status and academic achievement. Psychological Bulletin. 91, 61-81.

Table 1

Chronology of Selected Events in the Development of Meta-Analysis

1976 - Glass publishes first article on meta-analysis.

- Our first meta-analysis study receives Cattell Research Award.

1977 - Our first meta-analysis study is published.

- Lee J. Cronbach suggests our methods be generalized to all research literatures.

- Smith and Glass publish massive meta-analysis of effects of psychotherapy.

1978 - Callender and Osburn point out error in our first equation of SD_ρ.

1979 - We publish corrected formula for SD_ρ and meta-analyses of new data sets.

1980 - Callender and Osburn publish multiplicative equation for SD_ρ.

- We publish large meta-analysis studies of test validity for clerical job and computer programmers.

1981 - Glass, McGaw, and Smith publish first book on meta-analysis.

- Our group and Callender and Osburn publish separate meta-analysis of test validity in the Petroleum industry.

1982 - Our book on meta-analysis is published.

1983 - Raju and Burke publish two new equations for SD_ρ and ρ.

1984 - Rosenthal publishes book on meta-analysis.

- Cooper publishes book on meta-analysis.

Figure 1

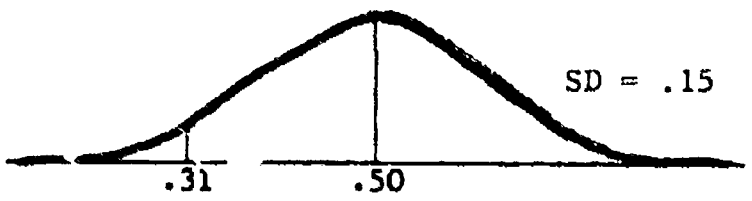

Figure 1A

Typical Validity Generalization Result for Cognitive Abilities in Predicting Job Performance. The Mean is .50 and the 10th Percentile (90% Credibility Val) is .31.

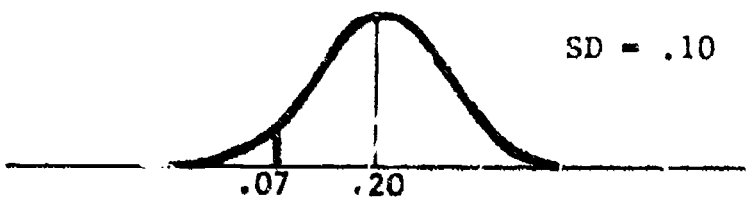

Figure 1B

Example of Validity Generalization Result with Relatively Low Mean and Low Variance. Despite Lower Mean, Validity is Still Generalizable; Representative of Results Obtained for Sales Clerks.

Figures 1C

Validity Generalization Result for Performance Tests Used with Clerical Workers. Substantial Mean but Large Variance; 90% Credibility Value is Negative.