Validity Generalization

A Critical Review

Edited by KEVIN R. MURPHY

Validity Generalization
A Critical Review

Validity Generalization
A Critical Review

Edited by

Kevin R.Murphy
Pennsylvania State University

LEA
2003
LAWRENCE ERLBAUM ASSOCIATES, PUBLISHERS
Mahwah, New Jersey London

This edition published in the Taylor & Francis e-Library, 2011.

To purchase your own copy of this or any of Taylor & Francis or Routledge's collection of thousands of eBooks please go to www.eBookstore.tandf.co.uk.

Lawrence Erlbaum Associates, Inc., Publishers
10 Industrial Avenue
Mahwah, New Jersey 07430

Cover design by Kathryn Houghtaling-Lacey

Library of Congress Cataloging-in-Publication Data

Validity generalization: a critical review/edited by Kevin R.Murphy.
p. cm. —(Applied psychology series)
Includes bibliographical references and index.
ISBN 0-8058-4114-8 (alk. paper)
1. Psychology—Research—Methodology. I. Murphy, Kevin R., 1952–
II. Series in applied psychology

BF76.5 .V35 2002
150′.7′2–dc21 2002024474
CIP

ISBN 1-4106-0687-2 Master e-book ISBN

Dedicated to the memory
of Jack Hunter (1939–2002)
who gave so much to the field of
Industrial and Organizational Psychology
—Frank Schmidt, University of Iowa

SERIES IN APPLIED PSYCHOLOGY

Edwin A.Fleishman, George Mason University,
Jeanette N.Cleveland, Pennsylvania State University
Series Editors

Gregory Bedny and David Meister
The Russian Theory of Activity: Current Applications to Design and Learning

Michael T.Brannick, Eduardo Salas, and Carolyn Prince
Team Performance Assessment and Measurement: Theory, Research, and Applications

Jeanette N.Cleveland, Margaret Stockdale, and Kevin R.Murphy
Women and Men in Organizations: Sex and Gender Issues at Work

Aaron Cohen
Multiple Commitments in the Workplace: An Integrative Approach

Russell Cropanzano
Justice in the Workplace: Approaching Fairness in Human Resource Management, Volume 1

Russell Cropanzano
Justice in the Workplace: From Theory to Practice, Volume 2

James E.Driskell and Eduardo Salas
Stress and Human Performance

Sidney A.Fine and Steven F.Cronshaw
Functional Job Analysis: A Foundation for Human Resources Management

Sidney A.Fine and Maury Getkate
Benchmark Tasks for Job Analysis: A Guide for Functional Job Analysis (FJA) Scales

J.Kevin Ford, Steve W.J.Kozlowski, Kurt Kraiger, Eduardo Salas, and Mark S.Teachout
Improving Training Effectiveness in Work Organizations

Jerald Greenberg
Organizational Behavior: The State of the Science (new edition coming 2003)

Uwe E.Kleinbeck, Hans-Henning Quast, Henk Thierry, and Hartmut Häcker
Work Motivation

Martin I.Kurke and Ellen M.Scrivner
Police Psychology into the 21st Century

Manuel London
Job Feedback: Giving, Seeking, and Using Feedback for Performance Improvement (new edition coming 2003)

Manuel London
How People Evaluate Others in Organizations

Manuel London
Leadership Development: Paths to Self-Insight and Professional Growth

Robert F.Morrison and Jerome Adams
Contemporary Career Development Issues

Michael D.Mumford, Garnett Stokes, and William A.Owens
Patterns of Life History: The Ecology of Human Individuality

Kevin R.Murphy and Frank E.Saal
Psychology in Organizations: Integrating Science and Practice

Kevin R.Murphy
Validity Generalization: A Critical Review

Ned Rosen
Teamwork and the Bottom Line: Groups Make a Difference

Heinz Schuler, James L.Farr, and Mike Smith
Personnel Selection and Assessment: Individual and Organizational Perspectives

John W.Senders and Neville P.Moray
Human Error: Cause, Prediction, and Reduction

Contents

Series Foreword

Series Editors

Jeanette N.Cleveland
Pennsylvania State University

Edwin A.Fleishman
George Mason University

There is a compelling need for innovative approaches to the solution of many pressing problems involving human relationships in today's society. Such approaches are more likely to be successful when they are based on sound research and applications. This Series in Applied Psychology offers publications that emphasize state-of-the-art research and its application to important issues of human behavior in a variety of societal settings. The objective is to bridge both academic and applied interests.

Twenty-five years have passed since the publication of Schmidt and Hunter's (1977) ground-breaking article laying out a model for meta-analysis and its special case of validity generalization (VG). This article and the work that followed changed the face personnel psychology and has had a profound impact on many other disciplines. There have been many papers, chapters, and books describing, extending, and applying the methods pioneered by Schmidt and Hunter in the last 25 years, but there has never been a comprehensive review describing how this method works, what it has accomplished, and where it is likely to be heading. The present volume fills this void.

Kevin Murphy has assembled a distinguished set of authors, many of whom have made important contributions to the metaanalysis and validity generalization literature, and asked each of them for constructive, critical

evaluations of the VG method, its applications, and its future development. The resulting chapters present a wide-ranging review of VG research and applications, as well as a number of new developments, unique perspectives, and insight-ful suggestions for the future of VG.

This book provides historical overviews of how this method was developed and how it has been applied. Of particular interest are the discussions of the impact of VG on the science and practice of personnel psychology, as well as the evolving relationship between VG research and professional standards for testing, assessment and personnel selection. Several chapters show, for example, how VG research has revolutionized research and applications linking personality measures and personnel selection. These are just a few examples of the unique contributions of this volume.

The book is critical in the best sense of the word. Each chapter presents fresh insights and balanced evaluations of the strengths and weaknesses of the VG method. It is clear from these chapters that VG is here to stay, and that it is likely to continue making important contributions. It is also clear that there are important problems yet to be solved; this book provides well-articulated frame-works (e.g., Bayesian statistics, generalizability theory) for attacking many of these issues. This timely and well-balanced set of chapters will be particularly valuable for both researchers and doctoral students in Industrial and Organizational Psychology. Further, human resources managers will find a number of chapters as a must-read as they design and finetune their testing and selection systems.

The present volume is welcomed for its coverage of new quantitative methods, developed in the context of the thriving areas of metaanalytic and VG research, which are clearly relevant and applicable to these important scientific issues. Psychologists, students, and others dealing with many areas of human performance will find much of value in this important and timely contribution.

Preface

It has been 25 years since the publication of Schmidt and Hunter's (1977) article "Development of a General Solution to the Problem of Validity Generalization" (*Journal of Applied Psychology,* 62, 643– 661). That article, and the subsequent stream of research, debate, and discussion about the meaning of validity generalization (VG), changed the face of personnel psychology. Prior to 1977, it was assumed that a new validity study would be needed virtually every time a test or selection procedure was tried out in some new setting. After 1977, it was often argued that a new local validity study was not only not needed, but that it might even add to the confusion rather than shedding new light on the validity of the test.

Developments in validity generalization led to wholesale changes in psychologists' assumptions about what conclusions could or could not be drawn from examining the cumulative literature. Prior to 1977, researchers often despaired of making sense of substantial bodies of research, largely because of the apparent instability of results from study to study. A test or intervention that seemed to work well in one organization would appear to fail in other similar settings, and given the extensive variability in study outcomes, it seemed that few good conclusions could be gained from looking at the research literature. Psychologists who wanted to know how a test or intervention would work in some particular setting would simply have to try it out there and see. VG research suggested that the fundamental relationships among tests and criteria and among the constructs they represent were simpler and more regular than they

appeared, and that the combined effects of sampling error, measurement error, and other statistical artifacts had blinded applied psychologists to the true worth of selection tests. More important, these statistical artifacts had blinded psychologists to the consistency in relationships between tests and between constructs.

The methods and conclusions of VG researchers have often been the subject of intense controversy. This controversy has in turn stimulated a number of developments and refinements of the concepts and calculations that underlie VG research. There are several good books describing VG methods, and numerous review articles that incorporate them, but there is no single source that provides an overview of the method, the controversies, the current status of VG, and the probable future of validity generalization research. The purpose of this volume is to provide that overview. Twenty-one authors have contributed chapters that outline the history and the contributions of the VG model, applications of this model to diverse domains, challenges to validity generalization, and alternative methods for attacking the key problems faced when using existing research to draw conclusions about the relationships among measures and among constructs.

In putting this volume together, I asked authors to take a critical view of VG, with the goal of describing what we do well with VG and identifying areas where more progress is needed. They more than met the challenge. The chapters in this volume document the many accomplishments and contributions of VG research, but they also provide concrete suggestions for further improving the process of validity generalization. More important, they provide a road map of where this method is likely to go in the future. It is clear from reading these chapters that VG is alive and well, but that daunting challenges remain to be faced in developing methods for extracting the most reliable information from the cumulative research literature. The authors of these chapters have helped lay the foundation for attacking these challenges and for building on the 25 years of progress that followed Schmidt and Hunter's original VG article.

During the production of this volume, I received the sad news that Jack Hunter had died. His contributions to psychology will live on, but we mourn his passing.

—Kevin R.Murphy

1

The Logic of Validity Generalization

Kevin R.Murphy
Pennsylvania State University

A few minutes in any college library is enough to illustrate one of the important characteristics of research in the behavioral and social sciences (i.e., that the number of books, papers, chapters, and reports published in these areas is simply enormous). For example, the various journals of the American Psychological Association publish tens of thousands of pages of peer-reviewed studies each year. The sheer volume of published work often makes the task of summarizing, integrating, and making sense of this research daunting. For example, a recent keyword search of the *PsychInfo* database using the term *attitude change* returned 1,800 citations. A search using the term *psychotherapy* returned more than 54,000 citations. In industrial and organizational psychology, a similar phenomenon has been noted, especially in the area of selection test validity. There have been thousands of studies examining the validity and utility of tests, interview methods, work samples, systems for scoring biodata, assessment centers, etcetera (e.g., a *PsychInfo* using the term *personnel selection* yielded more than 2,300 citations), and the task of interpreting this body of research is a challenging one.

For much of the history of personnel psychology, the task of interpreting this literature fell on the authors of textbooks and narrative reviews (notably Ghiselli, 1966, 1970). Throughout the 1960s and 1970s, reviews of literature on the validity and utility of tests and other selection methods

highlighted two recurrent problems: relatively low levels of validity for tests that seemed highly relevant to the job, and substantial inconsistencies in validity estimates from studies that seemed to involve similar tests, jobs, and settings. This pattern of findings led personnel psychologists to conclude that it would be difficult to predict or determine what sorts of tests might or might not be valid as predictors of performance in a particular job, that the validity of particular tests varied extensively across settings, organizations, etcetera, even when the essential nature of the job was held constant, and that the only way to determine whether a test was likely to be valid in a particular setting was to do a local validity study. The application of meta-analytic methods, and in particular, the validity generalization model to these same validation studies has led some very different conclusions about the meaning of this research. Applications of meta-analysis, and particularly validity generalization analyses, to studies of the validity of tests, interviews, assessment centers, and the like has led to the conclusions that (1) professionally developed ability tests, structured interviews, work samples, assessment centers, and other structured assessment techniques are likely to provide valid predictions of future performance across a wide range of jobs, settings, etcetera, (2) the level of validity for a particular test can vary as a function of characteristics of the job (e.g., complexity) or the organizations, but validities are often reasonably consistent across settings: and (3) it is possible to identify abilities and broad dimensions of personality that are related to performance in virtually all jobs (for reviews of research supporting these points, see Hartigan & Wigdor, 1989; Hunter & Hunter, 1984; McHenry, Hough, Toquam, Hanson, & Ashworth, 1990; Nathan & Alexander, 1988; Ree & Earles, 1994; Reilly & Chao, 1982; Schmidt & Hunter, 1999; Schmidt, Hunter, & Outerbridge, 1986; Schmitt, Gooding, Noe, & Kirsch, 1984; Wigdor & Garner, 1982. For illustrative applications of VG methods, see Callender & Osburn, 1981; Schmidt, Hunter, Pearlman, & Shane, 1979). Schmidt and Hunter (1999) reviewed 85 years of research on the validity and utility of selection methods and concluded that cognitive ability tests, work samples, measures of conscientiousness and integrity, structured interviews, job knowledge tests, biographical data measures and assessment centers all showed consistent evidence of validity as predictors of job performance.

The purpose of this chapter is to discuss the methods used by researchers to study the cumulative literature in areas such as test validity, and in particular, to lay out the logic behind the methods used in research on validity generalization (VG). Research on validity generalization is based

on an integration of meta-analysis and psychotric theory, and in order to understand the methods and results of VG research, it is important to examine the method and its logic in some detail.

METHODS OF META-ANALYSIS

The problem of making sense of the outcomes of hundreds or thousands of studies is in many ways similar to the problem of making sense of the data collected in any particular study. For example, if you conduct a study in which 200 subjects each complete some task of measure, the first step in making sense of the data you have collected is often to compute a variety of statistics that both describe that you found (e.g., means, standard deviations) and lend support to inferences you might make about what those data mean (e.g., confidence intervals, significance tests). One of the key insights of methodologists in the 1970s and 1980s was that the same could also be applied to the problem of making sense of a body of research. That is, if you wanted to make sense of the results of 125 different validation studies, each of which reported the correlation between some test and some measure of performance, one thing you would probably do would be to compute the mean and the standard deviation of the validities across studies. Many of the current methods of meta analysis take a more sophisticated approach to the problem than simply computing the average across all studies (e.g., they might weight for sample size), but the starting point for virtually all methods of meta analysis is essentially to compute some descriptive statistics that summarize key facets of the research literature you hope to summarize and understand. Differences in approaches to meta-analysis start to emerge as we move from descriptive statistics (i.e., what happened) to inferential ones (i.e., what does this mean).

The term *meta-analysis* refers to a wide array of statistical methods that are applied to the outcomes of multiple studies to describe in some sensible fashion what these studies have typically found, and draw inferences about what those findings might mean. Validity generalization represents a specialized application of meta-analysis that attempts to integrate both psychometric and statistical principles to draw inferences about the meaning of the cumulative body of research in a particular area (this method is sometimes also referred to as *psychometric meta-analysis*). In particular, validity generalization analyses attempt to draw inferences about the meaning of a set of studies, each of which has attempted to draw conclusions about fundamental relationships among the constructs being

studied on the basis of imperfect measures, finite samples, and studies that vary on a number of dimensions (e.g., the level of reliability of the measures used).

There are a number of methods of quantitatively summarizing the outcomes of multiple studies, any or all of which might be referred to as meta-analysis. For example, Rosenthal (1984) developed methods of combining the *p* values (i.e., probability that experimental results represent chance alone) from several independent studies to obtain an estimate the likelihood that the particular intervention, treatment, etcetera has some effect. Glass, McGaw, and Smith (1981) developed methods of combining effect size estimates (e.g. the difference between the experimental and control group means, expressed in standard deviation units) from multiple studies to give an overall picture of how much impact treatments or interventions have on key dependent variables. Schmidt and Hunter (1977) developed methods of combining validity coefficients (i.e., correlations between test scores and criterion measures) from multiple studies to estimate the overall validity of tests and other selection methods. Several variations on the basic VG model proposed by Schmidt and Hunter have been reviewed by Burke (1984) and Hedges (1988). Hedges and Olkin (1985) elaborated a general statistical model for meta-analysis that includes as a special case a variety of procedures similar to those developed by Schmidt and Hunter. Brannick (2001) discussed applications of Bayesian models in meta-analysis (see also Raudenbush & Bryk, 1985). Finally, Thomas (1990) developed a mixture model that attempts to describe systematic differences in validity among specific subgroups of validity studies.

The methods developed by Schmidt and Hunter have been widely applied, particularly within the field of personnel selection. For example, Schmidt (1992) noted that "meta-analysis has been applied to over 500 research literatures in employment selection, each one representing a predictor-job performance pair" (p. 1177). The most frequent application of these methods has been in research on the relationship between scores on cognitive ability tests and measures of overall job performance; representative examples of this type of validity generalization analysis include Pearlman, Schmidt, and Hunter (1980), Schmidt, Gast-Rosenberg, and Hunter (1980) and Schmidt, Hunter, and Caplan (1981). However, applications of metaanalysis and validity generalization analysis have not been restricted to traditional test validity research. Hunter and Hirsh (1987) reviewed meta-analyses spanning a wide range of areas in applied psychology (e.g., absenteeism, job satisfaction). Other recent applications

of meta-analytic methods have included assessments of the relationship between personality traits and job performance (Barrick & Mount, 1990), assessments of race effects in performance ratings (Kraiger & Ford, 1985) and assessments of the validity of assessment center ratings (Gaugler, Rosenthal, Thornton, & Bentson, 1987). Finally, Hom, Carnikas-Walker, Prussia, and Griffeth (1992) combined meta-analysis with structural modeling to assess the appropriateness of several competing theories of turnover in organizations.

VALIDITY GENERALIZATION: THE BASIC RATIONALE

The basic model developed by Schmidt and Hunter (1977) has gone through several developments and elaborations (Burke, 1984; James, Demaree, Mulaik, & Ladd, 1992; Raju & Burke, 1983; Schmidt et al., 1993), and the accuracy and usefulness of the model has been widely debated (e.g., Hartigan & Wigdor, 1989; James, Demaree, & Mulaik, 1986; Kemery, Mossholder, & Roth, 1987; Thomas, 1990). Although there is still considerable discussion and controversy over specific aspects of or conclusions drawn from validity generalization analyses, the core set of ideas in this method are simple and straightforward.

As noted earlier, the problem the validity generalization model was designed to address is that of making sense of research literature in which many, if not most of the relevant studies are of dubious quality. For example, many studies of the validity and utility of selection tests feature small sample sizes or unreliable criteria. Because sampling error leads to random variations in study outcomes and measurement error artificially lowers (i.e., attenuates) validities, it is reasonable to expect that validity coefficients from different studies will seem to vary randomly from study to study and will generally seem small. However, the effects of sampling error and unreliability are both relatively easy to estimate, and once the effects of these statistical artifacts are taken into account, you are likely to conclude that the actual validity of the test or assessment procedure studied is probably both larger and more consistent than a simple examination of the observed validity coefficients would suggest.

For example, suppose that there are 100 studies of the validity of structured interviews as predictors of job performance, and in each study the reliability of the performance measure is .70 and N (i.e., the sample size) is 40. If the average of the observed validity coefficients is .45, the formula for the correction for attenuation suggests that the best estimate

of the validity of the these interviews is probably closer to .54 (i.e., .45 divided by the square root of .70) than .45. Thus, a simple correction for measurement error suggests that the interviews are probably more valid than they seem on the basis of a quick examination of the validity studies themselves.

Because each validity coefficient comes from a fairly small sample, it is natural to expect some variability in study results; this variability can be estimated using a well-known formula for the sampling error of correlation coefficients (Hunter & Schmidt, 1990). For example, suppose that the standard deviation of the validity coefficients coming from these 100 studies was .18. On the basis of sampling error alone, you would expect a standard deviation of .12, given an *N* of 40 and a mean observed validity of .45 (see Hunter & Schmidt, 1990, for a detailed discussion of the formulas used to make such estimates). One conclusion that is likely to be reached in a validity generalization study is that much of the observed variation is test validities is likely to be due to the effects of sampling error rather than to the effects of real variation in test validity (here, 66% of the observed variability in validities might be due to sampling error).

Although the results of various approaches meta-analytic do not always agree (Johnson, Mullen, & Salas, 1995), these methods lead to similar general conclusions about the validity of selection tests, interviews, assessment centers, etcetera. In particular, it seems highly likely that test validities are generally both larger and more consistent across situations than the results of many individual validity studies would suggest (Hartigan & Wigdor, 1989; Schmidt, 1992; see, however, Murphy, 1993). Indeed, given the nature of much of the available validation research (i.e., small *N*, unreliable measures, range restriction), this general finding is virtually a foregone conclusion, although it directly contradicts one of the most widely held set of assumptions in personnel psychology (i.e., that validities are generally small and inherently unstable). Similarly, applications of the VG model to quantitative reviews of research on the validity of personality inventories as predictors of performance (e.g., Barrick & Mount, 1991; Hough, Eaton, Dunnette, Kamp, & McCloy, 1990; Tett, Jackson, & Rothstein, 1991) has overturned long-held assumptions about the relevance of such tests for personnel selection. Personnel researchers now generally accept the conclusion that scores on personality inventories are related to performance in a wide range of jobs.

The VG model suggests that there are a variety of statistical artifacts that artificially depress the mean and inflate the variability of validity

coefficients, and further that the effects of these artifacts can be easily estimated and corrected for. It is useful to discuss two broad classes of corrections separately, corrections to the mean and corrections to the variability in the distribution of validity coefficients that would be found in a descriptive meta-analysis.

Corrections to the Mean

There are several reasons why validity coefficients might be small. The most obvious possibility is that validities are small because the test in question is not a good predictor of performance. However, there are several statistical artifacts that would lead you to find relatively small correlations between test scores and measures of job performance, even if the test is in fact a very sensitive indicator of someone's job-related abilities. Two specific statistical artifacts that are known to artificially depress validities have received extensive attention in literature dealing with validity generalization, the limited reliability of measures of job performance and the frequent presence of range restriction in test scores, performance measures, or both.

There is a substantial literature dealing with the reliability of performance ratings (Viswesvaran, Ones, & Schmidt, 1996; Schmidt & Hunter, 1996) and other measures of job performance (Murphy & Cleveland, 1995). This literature suggests that these measures are often unreliable, which can seriously attenuate (i.e., depress) validity coefficients. For example, Viswesvaran et al.'s (1996) review showed that the average inter-rater reliability estimate for supervisory ratings of overall job performance was .52. To correct the correlation between a test score *(X)* and a measure of performance *(Y)* for the effects of measurement error in *Y,* you divide the observed correlation by the square root of the reliability of the performance measure. If you use inter-rater correlations as an estimate of reliability, corrected correlations will be, on average, be 38.7% larger than uncorrected correlations (i.e., if you divide the observed correlation by the square root of .52, the correction will lead to a 38.7% increase in the size of *r*). Murphy and DeShon (2001) questioned the use of inter-rater correlations as estimates of the reliability of ratings, but the general principle that low reliability will lead to what appears to be low levels of validity is beyond debate.

Performance ratings are normally collected in settings where range restriction is ubiquitous, especially when ratings are used to make

administrative decisions about ratees (e.g., salary, promotion; Murphy & Cleveland, 1995). For example, Bretz, Milkovich, and Read (1992) concluded that "the norm in U.S. industry is to rate employees at the top end of the scale" (p. 333). Ratees who consistently receive either very high ratings or very low ones are likely to be moved out of the job (e.g., promotions for high-rated employees, transfers or dismissals for low-rated employees), artificially truncating the distribution of ratings. Evidence of leniency and range restriction in performance ratings is so pervasive that several commentators (e.g., Ilgen, Barnes-Farrell, & McKellin, 1993; Jawahar & Williams, 1997) have urged caution in using ratings as criteria in validation studies. Range restriction can also artificially depress validity coefficients.

There has been considerable discussion in the literature about the best ways to correct for attenuation and range restriction (Hartigan & Wigdor, 1989; Hunter & Schmidt, 1990; Schmidt et al., 1993; Viswesvaran et al., 1996), and there are difficult issues with both corrections that have never been satisfactorily resolved (Cronbach, Gleser, Nanda, & Rajaratnam, 1972; Lumsden, 1976). However, the idea that both range restriction and the limited reliability of the measures used in validity studies depress validity coefficients, and that we can at least estimate and partially correct for this effect, is well accepted.

Corrections to the Variance

Meta-analyses of validity coefficients have often shown that the validity for the same type of test or measure varies considerably across jobs, organizations, settings, etcetera. This variability in validity coefficients is one of the chief reasons for the long-held assumption that it was necessary to validate tests in each setting where they were used. The validity generalization model suggests that some, and perhaps all of the variability in validity coefficients, might be explained in terms of a few simple statistical artifacts and that once the effects of these artifacts are removed, you are likely to conclude that the validity of tests is substantially similar across settings. Many potential explanations for variability in test validity have been put forth (e.g., the reliability of performance measures is higher in some organizations than in others, which can lead to apparent differences in validity), but much of the literature dealing with validity generalization has focused on the simplest and probably the most important explanation for differences in validity coefficients across studies (i.e., simple sampling

error). Many validity studies, particularly studies from the early 1970s and before, used small samples, and it is well known that statistical results of all sorts, including validity coefficients, are highly unstable when samples are small. Corrections for sampling error and for other artifacts that artificially inflate the variability in test validities often suggest that much of the apparent instability of validities is a reflection of weaknesses of validity studies (small samples, variation in the degree of unreliability and range restriction) rather than a reflection of true differences in the validity of tests across settings. For example, in McDaniel, Whetzel, Schmidt, and Maurer's (1994) analysis of the validity of situational interviews the standard deviation of the validities they reviewed was .14. After applying statistical corrections based on the VG model, this value shrunk to .05.

The cumulative effect of corrections that raise the mean and shrink the variance of the distribution of validities can be substantial. Returning to the example used above, McDaniel et al. (1994) reported that the mean of 16 validity coefficients for situational interviews was .27, and the standard deviation was .14. After correcting for statistical artifacts, the estimated population mean validity rose to .50, and with a standard deviation of .05. These researchers concluded that corrected validity of situational interviews was .43 or larger at least 90% of the time.

VG: THE INFERENTIAL MODEL

Validity generalization research involves developing and applying a particular type of inferential model, with the hope that this model can be used to understand the meaning of the cumulative body of research in a particular area. Traditionally, the model presented by VG researchers has resembled a variation on classic true score models used in psychometric research, in which there is an attempt to decompose the variance of test scores into its component parts. The classic true score theory of reliability states that the variance in observed scores can be broken down into that due to true scores and that due to measurement error, or:

$$\sigma^2_{observed} = \sigma^2_{true} + \sigma^2_{error} \quad (1)$$

which implies that any method that allows you to estimate the variance in scores due to measurement error also allows you to determine how much is due to true scores, or

$$\sigma^2_{true} = \sigma^2_{observed} - \sigma^2_{error} \quad \textbf{(2)}$$

In validity generalization studies, the *scores* to be analyzed are validity coefficients calculated in the different validity studies. In VG, the equivalent to Equation 1 is given by

$$\sigma^2_{observed} = \sigma^2_{true} + \sigma^2_{artifacts} \quad \textbf{(3)}$$

That is, the observed variance in validities from study can be broken down into true differences in validity from setting to setting and variance due to a number of statistical artifacts, such as sampling error, differences across settings in the amount or range restriction, the reliability of tests, etcetera. If this variance can be estimated, it is possible to draw inferences about the true variability in test validity on the basis of an equivalent to Equation 2, or

$$\sigma^2_{observed} = \sigma^2_{true} + \sigma^2_{artifacts} \quad \textbf{(4)}$$

As in traditional reliability theory, Equations 3 and 4 are based in part on the assumption that artifacts such as range restriction, test reliability, etcetera are independent of the true validity of tests. This assumption is probably not justified (James et al., 1992), but there is not much evidence that these artifacts are systematically correlated with test validity, and it is probably a reasonable approximation of the true state of affairs. Many of the modifications and refinements of the basic VG equations over the years (e.g., Schmidt et al., 1993) have been developed to capture more fully and accurately the variance due to statistical artifacts, which in turn should lead to better estimates of the true variability in test validity across jobs, settings, etcetera.

An alternative method of tackling the same problem (i.e., to understand the observed distribution of test validities) is to try and develop a process model. That is, the task of the VG researcher can be described as one of developing a that could explain key features of the observed distribution of test validities, then see if the model fits well enough to accept it as at least a reasonable approximation. Here, the *data* to be modeled are a set of correlations between (for example) scores on some test and measures of job performance, all obtained from relatively small and presumably independent samples. The mean of this distribution is likely

to be relatively small and the variance of this distribution is likely to be relatively large (see, e.g., Ghiselli, 1966). The problem is to generate a model of the process that might lead to this observed distribution of test-performance correlations.

Validity generalization analyses implicitly involve a process model of the sort previously described. For example, suppose you started with the assumption that scores on tests really did tell you a good deal about the job performance of applicants, and furthermore that the relationships between the test and performance were in fact reasonably consistent in their actual performance across settings. This implies that if you had a set of validity studies that combined essentially perfect measures, large and well-defined samples, minimal range restriction, etcetera you would find consistently large correlations between tests and performance. Suppose, however, that most researchers use unreliable measures of performance. This would tend to lower the correlations. Suppose, also, that there was substantial levels range restriction in many studies. This would also lower the correlations, yielding a distribution with a small mean (but also a small variance). Suppose further that the amount of range restriction, criterion unreliability, etcetera varied from study to study. This would lead to some variation in the observed distribution of correlations. Suppose further that validity studies often employed small samples. This would lead to further variation in validity coefficients. The net effect of all of these statistical artifacts would be to produce exactly what reviewers in the 1960s and 1970s bemoaned, validity coefficients that were often both small and variable. Yet none of the processes described above reflect real deficiencies in tests. Rather, they reflect deficiencies in validation research (i.e., small samples, unreliable measures). Furthermore, it might be hard to identify plausible alternative models that would explain this state of affair, although several possibilities are considered in the following.

Suppose, for example, you made the same assumption as was traditionally made by personnel psychologists in the 1960s and 1970s (i.e., that tests that "worked" in one setting might turn out to be virtually useless in other similar settings). If the validity of tests did in fact vary from setting to setting, the process model described earlier would lead one to expect an even more pessimistic set of conclusions than those reached by researchers at the time. That is, if there were real and meaningful differences in the validity of tests across most of the settings in which they were applied, the observed distribution of correlations should show even more variability than they typically do. It is certain that some of the variability in results across settings

is due to the often lamentable quality of the validity studies themselves, and if this artifactual variance in validity was piled on top of true and meaningful differences in validity across settings, it seems certain that the outcomes of validity studies would be much more varied that they have turned out to be. The VG model provides a coherent explanation of the observed distribution of validities, but a model that starts with the assumption that validity truly varies from job to job probably cannot fit the data, given the known effects of sampling error, limited reliability, etcetera.

Are Alternative Models Plausible?

One strength of the VG model is that it provides a coherent explanation for the observed distribution of validities. That is, if you start with a test that in fact works pretty well and pretty consistently, then carry out a large number of validity studies, each combining small *N* with unreliable performance measures, you will get pretty much what is actually known to occur (i.e., validity coefficients that are generally small and highly variable). Thus, the model "works," in the sense that it explains how the observed distribution of validities could arise from a population in which the actual validity is relatively large and relatively consistent. Any alternative model will have to take into account the known effects of low reliability, small sample sizes, differences in reliability across settings, and so forth, when explaining why validity coefficients are typically low and variable. This means, for example, that a model that takes the observed distribution of validities at face value (i.e., that concludes that validity truly is low and variable) is unlikely to be correct; if there was substantial real variation in validity across settings, the VG model suggests that the observed variance in validities should be much larger than it actually is.

There are, of course, other processes that might lead to the same outcome as the VG model. The most obvious possibility is a selfcensorship model. Suppose, for example, that you are studying the validity of a cognitive ability test, and you know that other studies in your area typically report uncorrected validity coefficients in the ballpark of .30. You obtain a validity coefficient of−.20 (or, for that matter, a coefficient of .70). It is likely that researchers who find validity coefficients that are very different from those normally reported with reanalyze their data, check their assumptions, evaluate their measures closely, etcetera, whereas researchers who obtain values that are closer to the norm will take them at face value. This process would lead to a tendency to doubt, revise, or discard validity estimates

that were too far out of the range of researchers' expectations and to keep those that fell within the distribution of results researchers were more comfortable with, distorting the distribution of *reported* validity coefficients (meta-analysts typically have no way of knowing whether the final result reported in a validity study is the result of data checking, transformations, or other sorts of data manipulation) in such a way that the observed distribution was similar to the distribution one would expect if the VG model was true.

A variation on this theme might be to argue that external censorship would produce the same result (i.e., a tendency to report results hat conform to establish patterns and to withhold publication of results that deviate sharply). As a journal editor, I receive frequent reminders of reviewers' tendencies to question findings that deviate sharply from current trends in the literature and to take at face value results that conform to those trends, and it is reasonable to argue that this tendency will favor the publication of studies that report similar results to those of existing studies and work against the publication of studies with discordant results. The result will be a body of published literature in which the observed variation in outcomes is less than the variation in outcomes of all studies.

It is not the argument that a self-censorship or an external censorship model is correct, or that these sort of processes plays any meaningful role in validity research. Rather, the point is made that demonstrating that the VG model could be used to explain the observed distribution of validities is a different matter than demonstrating that the VG model does explain this distribution. The VG model is certainly plausible, if certain assumptions are met, but the data that have been analyzed by VG researchers does not demonstrate that this model is correct. To make such a demonstration, one must either show that no plausible alternative exists or show that the model corresponds so closely to the phenomenon being studied that no alternatives would both explain the observed distribution of validities and correspond to the process by which validity coefficients were obtained in the first place.

There have been few attempts to generate, much less test alternative models for explaining the observed distribution of validity coefficients. Because few serious contenders have even been put forth, it is difficult to know how a proponent of the VG model could truly demonstrate that no alternatives exist. A more fruitful direction to pursue might be to compare some of the key features of the model with the known characteristics of validity studies. The closer the match, the lower the likelihood that

meaningful alternative models could be constructed that would explain the observed distributions of validity coefficients and lead to substantially different conclusions about the true value and validity of tests.

Key Assumptions of the VG Model

The assumptions that underlie meta-analyses in general and validity generalization analyses in particular are rarely articulated (for an exception see Hedges & Olkin, 1985), and it is likely that some researchers will dispute any conclusions that other researchers reach about the assumptions that underlie these techniques. The approach taken here is to examine the similarities and differences between statistical analyses in meta-analysis and comparable analyses in primary research. The statistical model for meta-analysis is usually a simple extension of the same model that would be used in analyzing data from a single study, and an explicit comparison of the problems of statistical inference when the data are individual observations versus the results of multiple studies can help to illustrate the challenges faced my meta-analysts in creating a useful model to explain their data.

For the purpose of illustrating the similarities and differences in the assumptions at different levels of analysis, consider two studies, one that involves collecting and analyzing data from a pool of subjects (primary analysis) and the other involving a meta-analysis of published and unpublished studies in a particular area. In the first study, 100 subjects are asked to carry out some experimental task, and their responses are recorded and analyzed. In the second study, the results form 100 existing studies are recorded and analyzed. In both cases, the goal of these statistical analyses is to both describe what happened in this particular study or set of studies and to draw inferences about what these findings mean in some broader sense.

The most obvious set of assumptions have to do with sampling. In a primary study, we treat subjects as a random sample from some particular population. In a meta-analysis, studies are treated as repeated samples from a common population. Regardless of the level of analysis, the population is rarely well-defined, and the process by which subjects are sampled from that population is rarely a random one. On this basis alone, some skepticism about the accuracy and meaningfulness of inferential statistics is often warranted, and there are as many reasons to worry about the results of primary analyses as about those of meta-analyses. A second

sampling assumption, the independence of observations, is likely to pose more serious problems in interpreting meta-analyses (Hedges & Olkin, 1985) than in interpreting primary research.

In traditional between-subjects research designs, it is reasonable to assume that the responses of subject *i* are independent of those of subject *k.* There are, of course, many designs (e.g., within-subjects or mixed designs, group data collections) in which this assumption is violated, and there is a rich methodological literature dealing with the problems caused by nonindependence and potential solutions to these problems. In meta-analysis, the assumption of independence is obviously inappropriate. For example, if there are 100 studies of the relationship between self-esteem and job performance, it is unreasonable to assume the researchers who design and carry out the 101st study will do so with complete disregard to the methods, results, and conclusions of previous studies (indeed, one of the factors motivating the development of meta-analysis and VG was researchers' inability to properly consider the implications of existing research when evaluating their own studies). It is often not clear how nonindependence will influence the design, analysis, or reporting of study results, but it is virtually certain that the results of studies in a particular area will be influenced by the body of available research. At present, there are no good methods for modeling, much less for correcting the effects of this nonindependence on the observed distribution of validity coefficients.

A second set of assumption involves the data themselves. It is usually assumed that data from multiple sources, either multiple subjects or multiple subjects, can be aggregated without fundamentally changing the nature of the variables being analyzed. This assumption is likely to be a serious problem in a meta-analysis. Except in special cases where a meta-analysis focuses on some specific test or measure (e.g., the validity of Minnesota Multiphasic Personality Inventory, 2nd edition [MMPI-2] scales for predicting social adjustment), it is common for the studies included in an analysis to feature many different operational definitions of the independent and dependent variables, For example, a study of the validity of spatial ability tests as predictors of job performance would certainly include a wide range of measures of both ability and performance. This raises the question of whether these can or should be combined, and if so, what the combined results mean.

Suppose, for example, that the meta-analysis described earlier includes 12 different (and nonequivalent) measures of spatial ability, five different measures of performance, and samples that differ in important ways from

study to study. Even if the validities turn out to be reasonably consistent (i.e., not much more different than one would expect on the basis of sampling error and other artifacts), it is still unclear what population parameter is being estimated. It is conventional to describe this sort of study as an examination of the validity of spatial ability tests as a predictor of job performance, and given the results described here, one might conclude that this relationship in invariant across the settings studied. However, it is difficult to describe precisely what this relationship really means. The population that is defined by the sorts of studies included in this validity generalization analysis is probably best described as studies involving a range of things that are referred to as spatial tests as predictors of another range of things referred to as performance measures, and even if the empirical results turn out to be reasonably consistent from study to study, the conventional interpretation (i.e., spatial ability predicts job performance) is not justified in terms of any specific statistical model.

In principle, there would be nothing to stop you from doing a grand meta-analysis of the relationship between all *X* variables and all Y variables studied in the social and behavioral sciences. Indeed, given the sample sizes typical in this research (e.g. Sedlmeier & Gigerenzer, 1989) it is likely that a large proportion of the variability in the r_{xy} values (for randomly chosen *Xs* and *Ys*) reported in all behavioral science research will probably be due to sampling error. The problem with this sort of meta-analysis is that its results would be impossible to interpret. The same general problem is, however, faced in all meta-analyses. Whenever studies that are nominally similar (e.g. *quantitative* and *job performance* in the title) are grouped together, there will be some ambiguity in describing exactly what the *population validity* estimated on the basis of that set of studies truly means. As a rule of thumb, it will probably be easier to interpret the results of meta-analyses that combine validity studies of relatively standardized tests or assessments (e.g. commercially published cognitive ability tests) than those that combine studies of tests or assessment procedures that are highly diverse (e.g. selection interviews, work sample tests, simulations).

Testing Meta-Analytic Models

There is an extensive literature dealing with the accuracy, bias, and efficiency of various meta-analytic procedures, and an examination of the methods used to test these models both confirms the importance of the assumptions laid out earlier and points out critical weaknesses of these

tests. The method of choice for evaluating VG models, or meta-analytic procedures in general is to conduct Monte Carlo studies (for a sample of illustrative applications, see Brannick, 2001; Callender & Osburn, 1981; Kemery et al., 1987; Law, Schmidt, & Hunter, 1994; Osburn, Callender, Greener, & Ashworth, 1983; Oswald & Johnson, 1998). A typical Monte Carlo study mirrors quite precisely the assumptions laid out above. That is, it is normal to generate data from a known set of population parameters, to create a large number of independent replications, to calculate validity coefficients in each of the samples generated, then to see whether various meta-analytic models do a better or worse job of accounting for characteristics of the observed distribution of validities. These methods can be extremely useful, and they often lead to important insights into the relationships among the various approaches to meta-analysis and validity generalization. If the assumptions that guided these models were at least reasonable approximations of the actual process by which studies are conducted, there would be little controversy about the accuracy of VG estimates or about the advantages and disadvantages of different models for meta-analysis and VG.

The central weakness of most tests of the accuracy of VG estimates is the gap between the assumptions needed to develop and test these models and the actual process by which validity results are produced and generated. The same, of course, can be said of most statistical models used in primary research. Assumptions of normality, random sampling, independence, and so forth are routinely violated in most studies in the behavioral and social sciences. Concerns about the possible effects of such violations have led to the development of a wide range of nonparametric techniques, and also to a large literature dealing with the robustness of statistical tests under various sorts of violations of statistical assumptions (Hunter & May, 1993; McDonald, 1999; Wilcox, 1992; Zimmerman & Zumbo, 1993; Zwick & Marascuilo, 1984). Some studies of the sensitivity of meta-analytic procedures to violation these assumptions have emerged (e.g., Oswald & Johnson, 1998), but even these are often limited by their dependence of the Monte Carlo framework. That is, most assessments of the accuracy and reasonableness of conclusions from VG analyses have been carried out in environments where assumptions of random sampling, independent observations, etcetera are literally met. The problem of determining the usefulness of these models under more realistic conditions is just starting be addressed (e.g., Steel & Kammeyer-Mueller, 2002), and it is likely that

this effort will require a very different set of methods than the standard Monte Carlo simulations that sure now used to evaluate VG models.

Some Conclusions About VG Models and Their Alternatives

In examining the literature dealing with alternative interpretations of VG results, three conclusions are discussed here. First, no serious contender has been put forth that accounts for the observed distribution of validities as well as the likely effects of statistical artifacts such as limited reliability and sampling error. Right now, the only explanation that has been put forth and has survived serious scrutiny is the explanation offered by VG researchers (i.e., that statistical artifacts mask the level and the consistency of validity coefficients).

Second, the fact that the VG model can explain the observed distribution of validities does not imply that it does explain this distribution. Key assumptions of the VG model are obviously wrong, and the discrepancies between the model and the reality it tries to explain are far from trivial. In particular, the results of validity studies cannot be thought of as independent observations sampled from some well-defined population. Rather, the results of any particular study are likely to be affected by whatever was known (or thought to be known) about the phenomenon at the time the study was performed, and these dependencies are likely to complicate the interpretation of meta-analytic findings.

Third, one potential route for developing and testing alternate models might be to focus on violations of key assumptions. The censorship models put forward earlier as a potential explanations for the observed variance in validities are an example. Mixture models (e.g., Thomas, 1990) designed to examine potential changes in validity as the operational definitions of X and Y change provide another example of using differences between the assumptions of the VG model and the reality it is designed to explain to generate alternative explanations.

VALIDITY GENERALIZATION VERSUS SITUATIONAL SPECIFICITY OF VALIDITIES

Murphy (1994) noted that researchers and practitioners sometimes confuse the claim that validity generalizes with the claim that validity is essentially constant across situations. This confusion has arisen largely because of changes, over time, in the way personnel researchers have conceptualized and discussed validity.

In the early years of validity generalization research (e.g., late 1970s to mid 1980s), researchers often talked about validity as if it were a dichotomous variable, that is, tests are either valid or not valid. This way of thinking closely mirrors the treatment of validity in the legal system, in which tests that led to adverse impact were held to be illegal unless they were shown to be valid. If such a showing was made, it did not matter much whether the test was just above the minimum threshold for defining validity or if it was a highly sensitive predictor of performance. Early research on validity generalization focused largely on the question of whether test validities exceeded some minimum level in most validity studies. Later research has focused more strongly on the consistency of validity across situations, in particular on the hypothesis that the level of validity achieved by a test might be situationally specific.

Distinguishing Between Validity Generalization and Situational Specificity

In the VG literature, the existence of substantial variability in the level of validity across situations (after correcting for statistical artifacts) is referred to as Situational specificity. If the correlation between test scores and job performance truly depends on the job, organization, or the situation, validity is said to be situationally specific. Validity generalization, on the other hand, refers to the classification of tests or other assessment devices as *valid* or *not valid.* If a test demonstrates at least a minimal level of validity in a sufficiently wide range of situations, validity is said to generalize. If a test cannot be consistently classified as *valid*, validity generalization fails.

The processes involved in testing the validity generalization and Situational specificity hypothesis overlap in many ways. In both cases, you start by calculating the mean and variance of the observed distribution of validities. Next, you correct for unreliability, range restriction, and other statistical artifacts that might affect the mean of the validity distribution, and correct for sampling error, variation across studies in range restriction and unreliability, and other statistical artifacts that might affect the variance of the distribution of validities (see Hunter & Schmidt, 1990, for formulas and sample calculations). At this point, the two processes diverge.

Tests of the Situational specificity hypothesis involve a comparison between the observed variance in validities and the variability expected solely on the basis of sampling error and other artifacts. If the variability

expected on the basis of statistical artifacts is as large, or nearly as large as the observed variability in validities, the Situational specificity hypothesis is rejected. Schmidt, Hunter, and their colleagues suggested a "75% rule," in which the Situational specificity hypothesis is rejected if the variability expected on the basis of statistical artifacts is at least 75% as large as the observed variance in validities (e.g., Schmidt et al., 1979). Other authors (e.g., Hedges & Olkin, 1985) use statistical tests of the homogeneity of correlations coefficients to evaluate this hypothesis. In many meta-analyses, the observed variance in validity coefficients is equal to or less than the variance that would be predicted on the basis of statistical artifacts alone, and this is often taken as evidence that true validities do not vary. Several aspects of the situational specificity hypothesis, including the decision rules used to evaluate the consistency of validities, are discussed in sections that follow.

The procedure for determining validity generalization is quite different from those used to evaluate situational specificity. After applying corrections for unreliability, sampling error, etcetera, the test of validity generalization involves comparing the bottom of the corrected validity distribution (e.g. the value at the 10th percentile of the corrected distribution) to some standard that represents a minimal level of validity (e.g. a validity coefficient of .00, or .10). For example, if the value at the 10th percentile of a corrected validity distribution was greater than .10, proponents of validity generalization would conclude that you could be 90% confident that the test would be at least minimally valid in essentially all new applications.

Gaugler, Rosenthal, Thornton, and Bentson (1987) conducted a meta-analysis of assessment center validities; results from this study can be used to illustrate the procedures used to evaluate validity generalization vs. situation specificity. Their review included 44 correlations (from 29 separate studies) between assessment center ratings and measures of job performance. The mean and the standard deviation of these validity coefficients were .25 and .15, respectively. After correcting for sampling error, unreliability, and other statistical artifacts, Gaugler et al. (1987) reported that: (a) the best estimate of assessment center validity was given by a corrected mean validity of .36, (b) the corrected validities varied substantially across studies (i.e., a corrected standard deviation of .14), and (c) 90% of the corrected validities were greater than .18. This set of results led them to conclude that the assessment center method was at least minimally valid in virtually all reported applications (i.e., assessment

center validity generalized) but that the level of validity was not consistent across studies, suggesting that characteristics of the jobs, organizations, assessment exercises, and so forth could substantially affect the validity of assessment center ratings.

In principle, there is no necessary relationship between tests of situational specificity and tests of validity generalization. The most common finding, at least in the area of ability testing, has been that validities are both: (a) generalizable, in the sense that such tests appear to be at least minimally valid predictors in virtually all settings, and (b) consistent, in the sense that the level of validity is reasonably comparable across settings (Hunter & Hirsch, 1987; Hunter & Hunter, 1984; Schmidt, 1992). However, it is also possible to conclude that validities are generalizable, but not consistent. That is, tests might show some validity in virtually all settings, but might be substantially more useful in some jobs, organizations, etcetera than in others.

On the whole, it is easier to demonstrate validity generalization than to demonstrate consistent levels of validity across situations. Mean validities are reasonably high for most structured selection procedures (see Hunter & Hunter, 1984; Reilly & Chao, 1982; Wiesner & Cronshaw, 1988), which means that the lower bound of the validity distribution is almost always greater than 0, .10, or whatever other standard is used to define minimal validity for this class of tests and assessment procedures. Demonstrations of situational specificity, on the other hand, are typically more difficult and controversial.

What Inferences Can Be Drawn From Tests of Situational Specificity?

Earlier it was noted that whereas many methods of meta-analysis provide what are basically descriptive statistics, VG analyses focus on inferential statistics (i.e., estimates of unknown population parameters). The descriptive-inferential distinction highlights one of the most difficult problems in using the results of VG analyses to draw inferences about situational specificity (i.e., the problem of deciding the conditions under which inferences can be drawn from the sample of studies included in a meta-analysis to the specific application in mind).

The logic of using situational specificity tests to make projections about the validity of a particular test in a particular situation is straightforward. If validities have not varied (except for variation due to sampling error

and other statistical artifacts) across a large number of studies included in a VG analysis, it is reasonable to conclude that they will also not change when we apply the test in a new and different situation. This description suggests that there are four key questions that need to be answered in determining whether inferences about the level of validity you can expect from a particular test can be on the basis of VG analyses: (1) did the VG analysis provide convincing evidence to refute the hypothesis of situational specificity?; (2) is the sample of validity coefficients included in the analysis sufficiently large and diverse to provide a reasonable picture of the population?; (3) is the test you are trying to validate a member of the same population of measures as that included in the VG analysis?; and (4) is the situation in which you wish to apply this test drawn from the population of situations sampled in the VG analysis?

First, it is important to ask whether a VG analysis provides credible evidence about situational specificity. Analyses that are based on relatively weak studies (e.g., studied with small *N* and unreliable criteria) may not allow you to convincingly sort out variance due to statistical artifacts from variance due to meaningful changes in validity across jobs, organizations, or settings. For example, many early validity generalization analyses featured average sample sizes of approximately 60 to 75 (see Table 1 in McDaniel et al., 1986), whereas more recent studies often have sample sizes ranging from approximately 600 to 750, depending on the criterion (Schmitt et al., 1984). Meaningful inferences about situational specificity depend first and foremost on the quality of the database that supports those inferences, and even studies that include a very large number of studies (as has been the case in some meta-analyses of the abilityperformance relationship) may not provide a firm basis for making inferences about situational specificity if most of the underlying studies are poorly designed.

Second, it is important to ask whether the sample of studies included in a meta-analysis spans the population of potential applications of the test. VG analyses that are based on a small number of validities, or that are based on studies taken from a very restricted range of potential applications may not provide a useful basis for making inferences about the consistency of test validity. For example, McDaniel et al. (1994) drew inferences about the validity of situational interviews on the basis of 16 validity coefficients, and about psychological interviews on the basis of 14 coefficients. They were appropriately cautious in interpreting these findings, and potential consumers of meta-analysis must also be cautious about overinterpreting consistency in a small set of validity studies. They

must be even more cautious about drawing broad inferences when the sample of validity studies spans only a small part of the range of situations in which a test might be used. For example, validity studies are more likely to be done in lower-level jobs (e.g., clerical jobs, semiskilled labor) than in managerial or professional jobs. When drawing inferences from a VG analysis, it is important to have detailed information about the range of jobs, situations, and so forth represented by the set of validity studies examined. Unfortunately, this sort of information is virtually never presented in the publications describing meta-analyses or VG analyses, and it is often necessary to go back to the original validity studies to determine what sorts of populations have actually been studied.

Third, it is important to determine whether the test you are hoping to use is a member of the same population of instruments that was examined in the body of literature summarized in a VG analysis. For example, there are probably hundreds of tests currently available that measure or claim to measure cognitive abilities (Murphy & Davidshofer, 1998). These tests do not all measure the same abilities (although they probably overlap substantially in their measurement of general cognitive ability, or *g*: Ree & Earles, 1994), and some tests are certainly better measures than others. Metaanalyses and VG studies rarely provide a detailed, explicit description of the population of tests, measures, etcetera they are designed to sample, and the process of determining whether the test you are hoping to use is really a member of the population of instruments sampled in a meta-analysis is sometimes little more than guesswork. In general, inferences that the test will work in the same way as tests sampled in the literature have worked are most likely to hold up if your tests is highly similar to the tests included in this meta-analysis.

Finally, it is important to consider whether the situation in which you hope to use a test is essentially similar to the situations sampled by the validity studies included in the VG analysis. For example, suppose that in most validity studies, range restriction is a relatively small concern (or is one that has been corrected for), and that validity coefficients reported in a meta-analysis are consistently in the .40's. In your organization, applicants go through extensive screening, and only a handful of candidates are allowed to go on for testing. Should you conclude that the correlation between test scores and performance is likely to be .40 in your organization? Probably not.

In sum, use of meta-analytic results suggesting that validity is essentially constant in a particular sample of studies to infer that it will

remain essentially constant in some new setting depends on the same sorts of assumptions and concerns that pertain to all inferential statistics. In particular, concerns over whether the test, situation, etcetera, that you have in mind is a member of the same population sampled in the meta-analysis are vital in determining what inferences can or cannot be made on the basis of meta analyses in particular and VG analysis in particular. Meta-analytic methods are tremendously useful in describing general trends in the research literature, and these trends often give selection practitioners a very good idea of what will or will not work. However, it is easy to overinterpret VG analyses and to make inferences about how well particular tests will work in particular settings that are not empirically justified. One of the great challenges in this sort of analysis is determining what inferences can be made (e.g., the inference that cognitive ability tests are at least minimally valid in most jobs seems a safe one) and what sort cannot be made on the basis of metaanalyses of validity studies.

STRONG INFERENCES FROM WEAK DATA

The methods and conclusions of meta-analysis in general and VG in particular have been the source of some controversy (e.g., Hartigan & Wigdor, 1989; Schmidt et al., 1985). To some extent, this controversy can be a reflection of general concerns about the logic of meta-analyses, particularly those that attempt to pull together very diverse sets of studies (e.g., studies employing a wide range of independent and dependent variables) to estimate a single population parameter (Eyesenck, 1978). This controversy also reflects the distaste of many reviewers for the answers provided by VG analyses (e.g., Hartigan & Wigdor, 1989, comment extensively on the social implications of VG research). However, the most lasting source of controversy is likely to be one that reflects the audacity of the endeavor itself. VG researchers frequently draw what appear to be very strong conclusions on the basis of what appear to be relatively weak data (e.g., see Schmidt's, 1992, widely cited paper telling us "what the data really mean"). That is, VG researchers often draw relatively strong conclusions about the relationships among constructs on the basis of the cumulative weight of what are often poorly conducted studies. For example, VG analyses have led some researchers to conclude that cognitive ability is related to job performance in virtually all jobs (Schmidt & Hunter, 1999). Schmidt, Hunter, and Pearlman (1981) went further concluding that the validity of selection tests for predicting performance was virtually

invariant across jobs. This conclusion was based largely on the basis of findings that the observed variance in validities in large-scale databases did not greatly exceed the variance expected on the basis of statistical artifacts. Subsequent research demonstrated that validity levels did in fact vary as a function of job level and complexity (Gutenberg, Arvey, Osburn, & Jenneret, 1983).

Can we really draw strong inferences from weak data? The VG model suggests that we often can. The VG model explains the pattern of results usually found when examining individual validity studies (validity coefficients that seem so weak and variable) in terms of the known effects of statistical artifacts. Once the effects of these artifacts are estimated and removed, there is often little room for other variables (e.g., differences in operationalizations of independent and dependent variables across studies) to have much effect on validities. The fact that the observed data fits this particular model well is an important argument in favor of accepting this explanation. On the other hand, the fact that a model fits the data pretty well is no assurance that it is the correct model (Birnbaum, 1973). There is a clear and immediate need for serious research on alternatives to the VG model. We have gone about as far as we can with demonstrations that it fits; the next step in drawing strong inferences must be an assessment of alternative explanations for these same data. Our ability to draw strong inferences will in the end depend on the explanation provided by the VG model is not only plausible, but that it is also right.

One of the unique strengths of the VG method is its marriage of meta-analysis and psychometrics. Schmidt and Hunter (1977) convincingly argued that we cannot draw appropriate conclusions simply by looking at correlations obtained using unreliable measures in finite, range-restricted samples. There is no doubt that analyses of appropriately corrected correlations would be more informative than analyses of observed correlations, and the theory that underlies these corrections is generally straightforward. However, there are reasons to believe that the operational procedures used to estimate reliability and perhaps range restriction effects do not provide the sorts of estimates that psychometric theory demands (Murphy & DeShon, 2000a, 2000b). For example, numerous VG analyses in recent years have used inter-rater correlations (which typically run in the .50 to .55 range) to estimate the reliability of performance ratings, which implies that observed correlations very substantially underestimate the true correlations between tests and performance measures. Interrater reliabilities are common in many areas of psychology, and there may be

good reasons to use them in some settings (e.g., where raters are roughly parallel), but it is clear that they do not estimate the sort of reliability coefficient that underlies the basic psychometric theory and many of the corrections that are at the heart of VG analyses. The greatest single difficulty in consummating a marriage between meta-analysis and psychometric theory may be in developing procedures that can be used to correctly estimate the sources of variance in the sorts of criteria routinely used in VG analysis. Murphy and DeShon (2000b) noted that validity research has not yet caught up with developments and refinements in psychometric thinking that have occurred over the last 30 years, and that we many be farther from obtaining credible estimates of population validities than a casual reading of the VG literature would suggest.

REFERENCES

Barrick, M.R., & Mount, M.K. (1991). The big five personality dimensions and job performance. *Personnel Psychology, 44,* 1–26.

Birnbaum, M.H. (1973). The devil rides again: Correlation as an index of fit. *Psychological Bulletin, 79,* 239–242.

Brannick, M. (2001). Implications of empirical Bayes meta-analysis for test validation. *Journal of Applied Psychology, 86,* 468–480.

Bretz, R.D., Milkovich, G.T., & Read, W. (1992). The current state of performance research and practice: Concerns, directions, and implications. *Journal of Management, 18,* 321–352.

Burke, M.J. (1984), Validity generalization: A review and critique of the correlational model. *Personnel Psychology, 37,* 93–113.

Callender, J.C., & Osburn, H.G. (1981). Testing the constancy of validity with computer-generated sampling distributions of the multiplicative model variance estimate: Results for petroleum industry validation research. *Journal of Applied Psychology, 66,* 274–281.

Cronbach, L.J., Gleser, G.C., Nanda, H., & Rajaratnam, N. (1972). *The dependability of behavioral measurements: Theory of generalizabtlity for scores and profiles.* New York: Wiley.

Eyesenck, H.J. (1978). An exercise in mega-silliness. *American Psychologist, 33,* 517.

Gaugler, B., Rosenthal, D., Thornton, G.C.III, & Bentson, C. (1987). Meta-analysis of assessment center validity. *Journal of Applied Psychology, 72,* 493–511.

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

Ghiselli, E.E. (1966). *The validity of occupational aptitude tests.* New York: Wiley.

Ghiselli, E.E. (1970). The validity of aptitude tests in personnel selection. *Personnel Psychology, 26,* 461–477.

Gutenberg, R.L., Arvey, R.D., Osburn, H.G., & Jenneret, P.R. (1983). Moderating effects of decision-making/information processing job dimensions on test validities. *Journal of Applied Psychology, 68,* 602–608.

Hartigan, J.A., & Wigdor, A.K. (1989). *Fairness in employment testing: Validity*

generalization, minority issues, and the General Aptitude Test Battery. Washington, DC: National Academy Press.
Hedges, L.V. (1988). Meta-analysis of test validities. In H.Wainer & H.Braun (Eds.), *Test validity* (pp. 191–212). Hillsdale, NJ: Lawrence Erlbaum Associates.
Hedges, L.V., & Olkin, I. (1985). *Statistical methods for meta-analysis.* New York: Academic Press.
Hom, P.W., Carnikas-Walker, F., Prussia, G.E., & Griffeth, R.W. (1992). A metaanalytical structural equations analysis of a model of employee turnover. *Journal of Applied Psychology, 77,* 890–909.
Hough, L.M., Eaton, N.K., Dunnette, M.D., Kamp, J.D., & McCloy, R.A. (1990). Criterion-related validities of personality constructs and the effect of response distortion on those validities. *Journal of Applied Psychology, 75,* 581–595.
Hunter, J.E., & Hirsh, H.R. (1987). Applications of meta-analysis. In C.L.Cooper & I.T.Robertson (Eds.), *International review of industrial and organizational psychology* (pp. 321–357). Chichester, UK: Wiley.
Hunter, J.E., & Hunter, R.F. (1984). Validity and utility of alternative predictors of job performance. *Psychological Bulletin, 96,* 72–98.
Hunter, J.E., & Schmidt, F.L. (1990). *Methods of meta-analysis: Correcting error and bias in research findings.* Newbury Park, CA: Sage.
Hunter, M.A., & May, R.B. (1993). Some myths concerning parametric and nonparametric tests. *Canadian Psychology, 34,* 384–389.
Ilgen, D.R., Barnes-Farrell, J.L., & McKellin, D.B. (1993). Performance appraisal process research in the 1980s: What has it contributed to appraisals in use. *Organizational Behavior and Human Decision Processes, 54,* 321–368.
James, L.R., Demaree, R.G., & Mulaik, S.A. (1986). A note on validity generalization procedures. *Journal of Applied Psychology, 71,* 440–450.
James, L.R., Demaree, R.G., Mulaik, S.A., & Ladd, R.T. (1992). Validity generalization in the context of situational models. *Journal of Applied Psychology, 77,* 3–14.
Jawahar, I.M., & Williams, C.R. (1997). Where all the children are above average: The performance appraisal purpose effect. *Personnel Psychology, 50,* 905–926.
Johnson, B.T., Mullen, B., & Salas, E. (1995). Comparison of three major metaanalytic approaches. *Journal of Applied Psychology, 80,* 94–106.
Kemery, E.R., Mossholder, K.W., & Roth, L. (1987). The power of the Schmidt and Hunter additive model of validity generalization. *Journal of Applied Psychology, 72,* 30–37.
Kraiger, K., & Ford, J.K. (1985). A meta-analysis of race effects in performance ratings. *Journal of Applied Psychology, 70,* 56–65.
Law, K.S., Schmidt, F.L., & Hunter, J.E. (1994). A test of two refinements in procedures for meta-analysis. *Journal of Applied Psychology, 79,* 978–986.
Lumsden, J. (1976). Test theory. *Annual Review of Psychology, 27,* 251–280.
MacDonald, P. (1999). Power, Type I, and Type III error rates of parametric and nonparametric statistical tests. *Journal of Experimental Education 67,* 367–379.
McDaniel, M.A., Hirsh, H.R., Schmidt, F.L., Raju, N., & Hunter, J.E. (1986). Interpreting the results of meta-analytic research: A comment on Schmitt, Gooding, Noe, and Kirsch (1984). *Personnel Psychology, 39,* 141–148.
McDaniel, M.A., Whetzel, D.L., Schmidt, F.L., & Maurer, S.D. (1994). The validity of employment interviews: A comprehensive review and meta-analysis. *Journal of Applied Psychology, 79,* 599–616.
McHenry, J.J., Hough, L.M., Toquam, J.L., Hanson, M.A., & Ashworth, S. (1990). Project

A validity results: The relationship between predictor and criterion domains. *Personnel Psychology, 43,* 335–355.
Murphy, K.R. (1993). The situational specificity of validities: Correcting for statistical artifacts does not always reduce the trans-situational variability of correlation coefficients. *International Journal of Selection and Assessment, 1,* 158–162.
Murphy, K. (1994). Advances in meta-analysis and validity generalization. In N.Anderson & P.Herriot (Eds.), *International handbook of selection and appraisal: Second Edition* (pp. 323–342). Chichester, UK: Wiley.
Murphy, K.R., & Cleveland, J.N. (1995). *Understanding performance appraisal: Social, organizational and goal-based perspectives.* Thousand Oaks, CA: Sage.
Murphy, K.R., & Davidshofer, C.O. (1998). *Psychological testing: Principles and applications* (4th ed). Englewood Cliffs, NJ: Prentice Hall.
Murphy, K., & DeShon, R. (2000a). Inter-rater correlations do not estimate the reliability of job performance ratings. *Personnel Psychology, 53,* 873–900.
Murphy, K., & DeShon, R. (2000b). Progress in psychometrics: Can industrial and organizational psychology catch up? *Personnel Psychology, 53,* 913–924.
Nathan, B.R., & Alexander, R.A. (1988). A comparison of criteria for test validation: a meta-analytic investigation. *Personnel Psychology, 41* 517–535.
Osburn, H.G., Callender, J.C., Greener, J.M., & Ashworth, S. (1983). Statistical power of tests of the situational specificity hypothesis in validity generalization studies: A cautionary note. *Journal of Applied Psychology, 68,* 115–122.
Oswald, F.L., & Johnson, J.W. (1998). On the robustness, bias, and stability of statistics from meta-analysis of correlation coefficients: Some initial Monte Carlo findings. *Journal of Applied Psychology, 83,* 164–178.
Pearlman, K., Schmidt, F.L., & Hunter, J.E. (1980). Validity generalization results for tests used to predict job proficiency and training success in clerical occupations. *Journal of Applied Psychology, 65,* 373–406.
Raju, N.S., & Burke, M.J. (1983). Two new approaches for studying validity generalization. *Journal of Applied Psychology, 68,* 382–395.
Raudenbush, S.W., & Bryk, A.S. (1985). Empirical Bayes metaanalysis. *Journal of Educational Statistics, 10,* 75–98.
Ree, M.J., & Earles, J.A. (1994). The ubiquitous predictiveness of g. In M.G. Rumsey, C.B.Walker, & J.H.Harris (Eds.), *Personnel selection and classification* (pp. 127–136). Hillsdale, NJ: Lawrence Erlbaum Associates.
Reilly, R.R., & Chao, G.T. (1982). Validity and fairness of some alternate employee selection procedures. *Personnel Psychology, 35,* 1–67.
Rosenthal, R. (1984). *Meta-analysis procedures for social research.* Beverly Hills, CA: Sage.
Schmidt, F.L. (1992). What do data really mean? Research findings, meta-analysis, and cumulative knowledge in psychology. *American Psychologist, 47,* 1173–1181.
Schmidt, F.L., Gast-Rosenberg, I., & Hunter, J.E. (1980). Validity generalization results for computer programmers. *Journal of Applied Psychology, 65,* 643–661.
Schmidt, F.L., & Hunter, J.E. (1977). Development of a general solution to the problem of validity generalization. *Journal of Applied Psychology, 62,* 643–661,
Schmidt, F.L., & Hunter, J.E. (1996). Measurement error in psychological research: Lessons from 26 research scenarios. *Psychological Methods, 1,* 199–223.
Schmidt, F.L., & Hunter, J.E. (1999). The validity and utility of selection methods in

personnel psychology: Practical and theoretical implications of 85 years of research findings. *Psychological Bulletin, 124,* 262–274.

Schmidt, F.L., Hunter, J.E., & Caplan, J.R. (1981). Validity generalization results for two groups in the petroleum industry. *Journal of Applied Psychology, 66,* 261–273.

Schmidt, F.L., Hunter, J.E., & Outerbridge, A.N. (1986). Impact of job experience and ability on job knowledge, work sample, performance, and supervisory ratings of job performance. *Journal of Applied Psychology, 71,* 432–439.

Schmidt, F.L., Hunter, J.E., & Pearlman, K. (1981). Task differences as moderators of aptitude test validity In selection: A red herring. *Journal of Applied Psychology, 66,* 166–185.

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., Hirsch, H.R., Sackett, P.R., Schmitt, N., Tenopyr, M.L., Kehoe, J., & Zedeck, S. (1985). Forty questions about validity generalizations and meta-analysis with commentaries. *Personnel Psychology, 37,* 407–422.

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., Law, K., Hunter, J.E., Rothstein, H.R., Pearlman, K., & McDaniel,M. D. (1993). Refinements in validity generalization procedures: Implications for the situational specificity hypothesis. *Journal of Applied Psychology, 78,* 3–14.

Schmitt, N., Gooding, R.Z., Noe, R.D., & Kirsch, M. (1984). Metaanalyses of validity studies published between 1964 and 1982 and the investigation of study characteristics. *Personnel Psychology, 37,* 407–422.

Sedlmeier, P., & Gigerenzer, G. (1989). Do studies of statistical power have an effect on the power of studies? *Psychological Bulletin, 105,* 309–316.

Steel, P., & Kammeyer-Mueller, J. (2002). Comparing meta-analytic moderator search techniques under realistic conditions. *Journal of Applied Psychology, 87,* 96–111.

Tett, R.P., Jackson, D.N., & Rothstein, M. (1991). Personality measures as predictors of Job performance: A meta-analytic review. *Personnel Psychology, 44,* 703–742.

Thomas, H. (1990). A likelihood-based model for validity generalization. *Journal of Applied Psychology, 75,* 13–20.

Viswesvaran, C., Ones, D.S., & Schmidt, F.L. (1996). Comparative analysis of the reliability of job performance ratings. *Journal of Applied Psychology, 81,* 557–574.

Wiesner, W.H., & Cronshaw, S.F. (1988). A meta-analytic investigation of the impact of interview format and degree of structure on the validity of the interview. *Journal of Occupational Psychology, 61,* 275–290.

Wigdor, A.K., & Garner, W.R. (1982). *Ability testing: Uses, consequences, and controversies.* Washington, DC: National Academy Press.

Wilcox, R.R. (1992). Why can methods for comparing means have relatively low power, and what can you do to correct the problem? *Current Directions in Psychological Science, 1,* 101–105.

Zimmerman, D.W., & Zumbo, B.D. (1993). The relative power of parametric and nonparametric statistical methods. In G. Keren & C. Lewis (Eds.), *A Handbook for Data Analysts in the Behavioral Sciences: methodological Issues* (pp. 481–518). Hillsdale, NJ: Lawrence Erlbaum Associates.

Zwick, R., & Marascuilo, L.A. (1984). Selection of pairwise comparison procedures for parametric and nonparametric analysis of variance models. *Psychological Bulletin, 95,* 148–155.

2

History, Development, Evolution, and Impact of Validity Generalization and Meta-Analysis Methods, 1975–2001[1]

Frank Schmidt
University of Iowa

John Hunter
Michigan State University

EARLY HISTORY AND DEVELOPMENT

Background and Origins

In the Industrial/Organizational (I/O) PhD program at Purdue and other universities in the 1960s, the textbooks and most courses taught that selection procedure validities were situationally specific. The dominant theory held that the validity of the same test for what seemed to be the same job varied from employer to employer, region to region, across time periods, and so forth. It was believed and taught that the same test could have high validity in one location or organization and at the same time have zero validity in another for the same or very similar jobs. This belief was supported by the finding that observed validity coefficients for similar tests and jobs varied substantially across different validity studies and the finding that some of these validity coefficients were statistically significant

[1] For ease of expression this chapter is written in the first person singular of the first author. However, the test reflects the views of both authors.

and others were not. The explanation for this puzzling variability was that jobs that appeared to be the same differed in important but subtle ways in what was required to perform them. The conclusion, we were taught, was that the validity of selection procedures had to be estimated anew for each different situation or setting by a validity study conducted in that setting. It was impossible to generalize validity from one setting to others.

For most of my undergraduate career I was a biology major. Perhaps partly for this reason, I found it hard to believe that general and generalizable relationships did not exist in this area because it seemed to me that the goal of research in science was the discovery of generalizable relationships and the explanation of these relationships through theory. In a course on personnel selection taught by Hubert Brogden, I asked for his explanation for the variability of validity estimates. He replied that he thought most of the variability was due to sampling error stemming from use of small samples in civilian validity studies. Brogden's background was different from that of the other faculty: He formerly had been the research director of what is now the Army Research Institute (ARI). His own research had employed the large samples characteristic of military research, and he reported that validities behaved in a lawful and predictable manner. There were no puzzles: If a test was valid for a job in one study, it was valid in all such studies. Other faculty did not accept this view.

After moving to U.S. Office of Personnel Management (OPM) from the Michigan State I/O program in 1974, I began research on the currently important question of the technical feasibility of criterionrelated validity studies. As the employment agency for the federal government, we were under pressure from EEOC (Equal Employment Opportunity Commission) to conduct criterion-related validity studies whenever they were *technically feasible*, a term that had not been adequately defined. It occurred to me that low statistical power to detect validity was a way to define technical infeasibility. Hunter, Urry, and I (Schmidt, Hunter, & Urry, 1976) showed that the typical validity study in the literature had statistical power of only about .50 to detect validity given its presence (meaning that these studies were not technically feasible to begin with). These findings led me back to the question of validity generalizability that I had dropped years ago at Purdue. I realized that these findings explained one aspect of the data supporting the situational specificity theory: the fact that about half of all reported validities were statistically significant and half were not. I began to think about the possibility that most or all of the variability in validity estimates for a given test-job combination might be artifactual.

One day in 1975, sitting in my office looking at a distribution of observed validities reported by Ed Ghiselli (1969) in one of his books, the thought occurred to me that I could estimate the amount of variance due to sampling error variance in that distribution by averaging the amount of sampling error variance in the individual studies. I was shocked to find that this came out to be about 70% of the observed variance. I was also surprised to see that once I had subtracted out sampling variance, the remaining variance was small enough to produce a corrected (residual) standard deviation (SD) of validities small enough that almost all observed validities were positive. Then, when I corrected the mean validity for criterion unreliability and mean range restriction (using reasonable estimates of these based on figures in the literature), I saw that almost all the validities were not only positive but also substantial in magnitude.

Excited about this result, I called Jack Hunter at Michigan State. In his opinion, there were no errors in my calculations. He thought this general approach was likely to be fruitful and was also excited about the implications, later sending me an enthusiastic letter to that effect. We discussed ways to refine this procedure so it could be applied systematically to data in the literature. For example, we needed ways to estimate how much variability in validities was due to differences between studies in criterion reliability and range restriction—which I had not at that time taken into account. When we had done this and applied the procedure to several validity distributions from Ghiselli, the resulting paper won the 1976 Society for the Industrial/Organizational Psychology (SIOP) James McKeen Cattell Research Design Award. The following year an expanded version of this paper became our first published validity generalization paper (Schmidt & Hunter, 1977).

About this time we started corresponding with a number of psychologists interested in this topic. (Most of this correspondence is preserved and was available for review in preparing this chapter.) Edwin Ghiselli thought the idea was excellent; he very much regretted that, as a result of a dispute with the Berkeley psychology department, he had just recently destroyed all of his validity files, making it impossible for us to use his data. Bob Guion praised the concept and method, as did Anne Anastasi, who cited the research favorably in subsequent editions of her book *Psychological Testing*, greatly facilitating its acceptance. Marvin Dunnette responded positively and later became an important public defender of this research; he also later confirmed our findings in his own data sets. Lee J.Cronbach wrote saying the method should not be limited to personnel selection