CHAPTER 5

Validity Generalization as a Test Validation Approach

Michael A. McDaniel

Validity generalization is an application of meta-analysis to the correlations between an employment test and a criterion, typically job or workplace training performance. By employment test, I mean any method used to screen applicants for employment, including cognitive ability tests, personality tests, employment interviews, and reference checks. The purpose of this chapter is to review the use of validity generalization as a test validation strategy. I begin by reviewing the issues of situational specificity, differential validity, differential prediction, and racial and ethnic differences in test performance. In the 1970s, these were issues for which professional consensus had not been reached or professional consensus was later shown to be incorrect. These issues prominently influenced the *Uniform Guidelines on Employee Selection Procedures* (Equal Employment Opportunity Commission, Civil Service Commission, Department of Labor, & Department of Justice, 1978). Subsequent to adoption of the *Uniform Guidelines,* compelling research findings have altered professional consensus on these issues such that the *Uniform Guidelines* are at variance with current professional guidance, thus rendering the *Uniform Guidelines* technically flawed and in need of revision.

Situational Specificity

Dating from the 1920s, researchers observed that different studies using the same employment test showed different validity results (Schmidt & Hunter, 1998). This led to the hypothesis that there were yet-to-be-discovered characteristics of the validity setting (the situation) or the job that caused tests to be valid in one situation but less valid or not valid in another situation. This phenomenon was termed the "situational specificity" hypothesis (Schmidt & Hunter, 1998). It also appeared that the yet-to-be-discovered differences in situations and jobs that caused the varying validity results could not be identified through job analysis. Because the validity of a test appeared to vary across settings, it was argued that employers who used employment tests would need to conduct local studies to determine the validity of the tests in their settings with their jobs. By the 1970s, the situational specificity theory had come to be regarded as a well-established fact (Guion, 1975; Schmidt & Hunter, 2003), and the practice of local validation studies was common.

Beginning in 1977, Schmidt and Hunter began to develop and report data questioning the situational specificity hypothesis. They observed that sampling error was the major source of variation in validity coefficients across studies (Schmidt & Hunter, 1977). They also observed that measurement error reduces observed validity coefficients and that differences across studies in the degree of measurement error would also cause validity coefficients to vary across studies. They further observed that range restriction similarly reduced observed validity coefficients and that likewise differences across studies in the degree of range restriction result in variation in validity coefficients.

As a result and in support of this research, Schmidt and Hunter (1977) developed psychometric meta-analytic procedures (Hunter & Schmidt, 2004) for determining the extent to which the variation in validity across studies was due to sampling error and differences across studies in measurement error and range restriction. They continued to confirm in subsequent research that most of the variation across studies was due to these statistical artifacts (Pearlman, Schmidt, & Hunter, 1980; Schmidt, Gast-Rosenberg, & Hunter, 1980).

Thus, it was confirmed that the variation across studies that had been previously attributed to mysterious situational factors was largely due to these statistical artifacts, primarily simple random sampling error. This application of meta-analysis became known as *validity generalization* when applied to employment validity results. The groundbreaking work of Schmidt and Hunter (1977) and many subsequent meta-analyses of test validity data indicated that the hypothesis of situational specificity, which had been thought to be well-established, was disconfirmed. Unfortunately, the *Uniform Guidelines* had been adopted before this information regarding the situational specificity hypothesis became well known.

Differential Validity and Differential Prediction

By the 1970s, the concept of situational specificity and racial and ethnic differences in test scores also fueled the debate concerning differential validity and differential prediction (Boehm, 1972; Bray & Moses, 1972; Kirkpatrick, Ewen, Barrett, & Katzell, 1968). These hypotheses held that validity or prediction accuracy might vary by racial or ethnic subgroup. Since it appeared that some unknown situational factors caused the validity of tests to vary, it was possible that other unknown factors could cause a test to be valid for Whites but not for minorities. The common finding that Blacks, on average, scored lower than Whites on employment tests fueled the speculation that local validation studies were the best way to determine if a test was valid for all subgroups. However, by the mid-1980s it had become clear that differential validity is very rare (Schmidt, 1988; Schmidt & Hunter, 1981; Wigdor & Garner, 1982). Research on differential validity evolved into research on differential prediction, because even if a test were equally valid for all subgroups, the optimal regression lines to predict job performance might differ. Differential prediction might occur in either differing slopes or differing intercepts. However, research indicated that differing slopes occur at no higher than chance levels (Bartlett, Bobko, Mosier, & Hannan, 1978). Differing intercepts are somewhat less rare, but the error in prediction tends to favor minority groups; that is, when the prediction of job performance for minority groups and Whites is based on a common regression line, performance of the minority groups is often overpredicted when

differential prediction exists (Hartigan & Wigdor, 1989; Schmidt, Pearlman & Hunter, 1980). Thus, in the 1980s, it became clear that differential validity is a rarely detected phenomenon, and differential prediction, to the extent that it does exist, does not bias employment tests against minorities. Unfortunately, the *Uniform Guidelines* were written before this research was conducted and widely recognized.

Mean Racial Differences in Employment Test Scores

It is difficult to find an example of a cognitively loaded employment test on which Black or Hispanic minorities perform as well as Whites, on average. Over time, these differences have proven intractable despite various efforts to reduce them, and the magnitude of the differences has been relatively stable. To date there appear to be few, if any, interventions that are consistently effective in reducing these mean differences (Sackett, Schmitt, & Ellingson, 2001). A comprehensive review of Black-White mean differences on general cognitive ability tests indicated that White mean scores are about one standard deviation higher (Roth, BeVier, Bobko, Switzer, & Tyler, 2001). Previous research had indicated smaller differences, but Roth et al. showed how such differences were due to restricted range in the samples. Thus, mean racial differences are smaller in college samples because college samples only include those who have obtained a high school diploma and met other selection criteria of the college. Research places the Hispanic-White mean score difference as somewhat smaller but still substantial. Minority groups have also shown lower mean scores than Whites on less cognitively loaded measures, such as employment interviews (Roth, Van Iddekinge, & Huffcutt, 2002), assessment centers (Fiedler, 2001; Goldstein, Yusko, & Braverman, 1998; Goldstein, Yusko, & Nicolopoulos, 2001), and situational judgment tests (Nguyen, McDaniel, & Whetzel, 2005). To the extent (albeit limited) that personality tests show mean racial differences, they tend to disfavor Blacks (Goldberg, Sweeney, Merenda, & Hughes, 1998; Hough, Oswald, & Ployhart, 2001). Virtually all employment tests show mean racial differences.

Although there had been some early evidence that Black scores were substantially lower than White scores on measures of general cognitive ability, the size and the intractability of that difference became more apparent after the end of the 1970s. Unfortunately, the *Uniform Guidelines* were written before the extent and persistence of racial or ethnic score differences were fully appreciated. Under the *Uniform Guidelines,* when an employer's test use results in differential hiring rates based on this very common finding of mean racial or ethnic score differences, the employer is faced with requirements to demonstrate validity and business necessity. Specifically, employers need to offer validity evidence, which may require the employer to conduct a local validation study, to prepare validity transport evidence, or to participate in a consortium validity study—all of which are activities that require substantial resources.

Addressing the Uniform Guidelines

The widespread acceptance of situational specificity in the 1970s, concerns over potential differential validity and differential prediction, and general uneasiness concerning racial and ethnic mean score differences all formed a part of the world view when the *Uniform Guidelines* were adopted. It is not surprising that the authors of the *Uniform Guidelines* incorporated a number of viewpoints into the regulations that are inconsistent with current scientific thinking. Specifically, the *Uniform Guidelines* are biased in favor of

- Conducting local validation studies
- Conducting differential validity and differential prediction studies
- Requiring detailed and costly job analysis data

In addition, when there are racial disparities in hiring as a result of mean racial differences in test performance, the *Uniform Guidelines* encourage a close examination of the validity of the test rather than viewing the racial difference in test performance as a common finding consistent with many years of cumulative data. Worse yet, they may influence employers to shun employment testing or to follow nonoptimal selection practices such as setting low

cutoff scores, banding, and other methods of gerrymandering employment test results in attempts to eliminate the disparity in hiring rates. Thus, the provisions of the *Uniform Guidelines* may encourage employers to make nonoptimal selection decisions that result in productivity losses. Specifically, if an employer engages in nonoptimal implementation of their selection process in order to hire minorities at about the same rates as Whites, they may avoid inquiries by the Equal Employment Opportunity Commission (EEOC) and the Office of Federal Contract Compliance Programs (OFCCP). Ironically, an employer's failure to screen effectively on cognitive skills in its hiring decisions can exacerbate mean racial differences in job performance (McKay & McDaniel, 2006) that may also trigger enforcement agency scrutiny. Use of nonoptimal selection practices can cause serious detriment to the productivity of organizations.

It should be noted that employers may follow nonoptimal selection practices without pressure from enforcement agencies. For example, an employer might seek to hire more minority employees than would be hired using selection tests in the most optimal way because they feel a social, public relations, or business need to do so. However, the *Uniform Guidelines* increase the pressure on employers to make such selection decisions.

Shortly after the adoption of the *Uniform Guidelines,* the Executive Committee of the Society of Industrial and Organization Psychology (SIOP) sent a letter to the authoring agencies of the *Uniform Guidelines* documenting how they were inconsistent with professional guidance. The letter did not result in a revision of the *Uniform Guidelines.* Thus, although the *Uniform Guidelines* state that they are intended to be consistent with professional standards, the federal agencies that are responsible for them have not called for their revision.

Professional Guidelines and Validity Generalization

In contrast to the *Uniform Guidelines,* professional associations have updated professional standards relevant to employment testing to recognize the significant merit of validity generalization analyses. The *Standards for Educational and Psychological Testing* are profes-

sional guidelines issued jointly by the American Educational Research Association, the American Psychological Association, and the National Council on Measurement in Education (1999). The *Principles for the Validation and Use of Personnel Selection Procedures* are published by the Society for Industrial and Organizational Psychology (2003). Both the *Standards* and the *Principles* were adopted by the APA and thus have "formal professional status" (Jeanneret, 2005, p. 49). These documents are important because they summarize the best judgment of the profession concerning the state of the science regarding acceptable validation research. They are also important because the *Uniform Guidelines* state:

> The provisions of these guidelines relating to validation of selection procedures are intended to be consistent with generally accepted professional standards for evaluating standardized tests and other selection procedures, such as those described in the Standards for Educational and Psychological Tests prepared by a joint committee of the American Psychological Association, the American Educational Research Association, and the National Council on Measurement in Education . . . and standard textbooks and journals in the field of personnel selection. (Section 5C)

In addition, the *Uniform Guidelines* (Section 5A) state that "New strategies for showing the validity of selection procedures will be evaluated as they become accepted by the psychological profession." Thus, to the extent that the *Uniform Guidelines* are in conflict with professional principles and standards as well as scientific knowledge in textbooks and journals, one can argue that they should be read in concert with those more current professional sources in developing validation evidence for employment tests.

Both the *Standards* and the *Principles* endorse the use of validity generalization as a means for establishing the validity of an employment test. The *Standards* recognize that local validation studies are not preferred over validity generalization evidence. The *Principles* share this view and state:

> Meta-analysis is the basis for the technique that is often referred to as "validity generalization." In general, research has shown much of the variation in observed differences in obtained validity coefficients in different situations can be attributed to sampling error

> and other statistical artifacts (Barrick & Mount, 1991; Callender & Osburn, 1980; 1981; Hartigan & Wigdor, 1989; Hunter & Hunter, 1984). These findings are particularly well established for cognitive ability tests; additional recent research results also are accruing that indicate the generalizability of predictor-criterion relationships for noncognitive constructs in employment settings. (p. 28)

The *Principles* also acknowledge that validity generalization is often a better source of validity information than a local validity study (p. 29). In summary, both the *Standards* and the *Principles* have endorsed the usefulness of validity generalization for drawing inferences about the validity of an employment test.

Validity Generalization in Court

In contrast to the effects of validity generalization on the practice and knowledge of personnel selection, the methodology has received limited attention by the courts. Landy (2003) provides an overview of validity generalization in court proceedings and the following draws in part on his overview.

Although predating validity generalization, a case concerning screening for firefighters in Richmond, Virginia, addressed the extent to which validity results are transportable across employers (*Friend et al. v. Leidinger et al.,* 1977). The court found that the firefighter job in Richmond was very similar to the firefighter job in California, where the test had been the subject of a validation study, and that the California study was sufficient validity information for the test to be considered valid in Richmond. The Court of Appeals (*Friend et al. v. City of Richmond et al.,* 1978) agreed with the decision and noted that "To require local validation in every city, village, and hamlet would be ludicrous" (p. 65) (see Gibson & Caplinger (2007) of this volume).

In a reverse discrimination case in the City of Detroit, White applicants claimed reverse discrimination because the City did not use strict rank ordering on a test to select applicants for the fire academy (*Van Aken et al. v. Young and City of Detroit et al.,* 1982). The plaintiff's expert witness used validity generalization to argue that the test should have been used in a top-down manner to achieve maximum utility. The expert argued that cognitive ability tests are valid for all jobs. The court did not accept the argument and noted

that the expert knew nothing specific about the Detroit Fire Department and the academy curriculum. I note that the expert's testimony was consistent with the then (and still) current state of scientific knowledge, specifically that cognitive ability tests are valid for all jobs. Thus, it appears that in 1982 the situational specificity hypothesis still influenced some court decisions.

Validity generalization fared better in *Pegues v. Mississippi State Employment Service* (1983). The expert witness provided testimony on the validity generalization based on his analysis of the General Aptitude Test Battery (GATB) validity evidence. The judge cited research indicating that even gross changes in job duties do not destroy validity. He concluded that "Plaintiffs have not shown that the USES tests (in this case the GATB) were invalid because the tasks of the jobs in the research setting may have been different from those in Bolivar County" (p. 1136). Here the judge accepted the scientific conclusions of validity generalization research and clearly rejected situational specificity.

In *EEOC v. Atlas Paper Box Company* (1987), the expert witness used validity generalization research in support of the validity of a cognitive ability test. The trial judge concluded that the sample sizes at Atlas would have been too small to provide meaningful local validity studies. However, in 1989 the Sixth Circuit of Appeals rejected the validation generalization argument:

> The expert witness offered by the defendant, John Hunter, failed to visit and inspect the Atlas office and never studied the nature and content of the Atlas clerical and office jobs involved. The validity of the generalization theory utilized by Atlas with respect to this expert testimony under these circumstances is not appropriate. Linkage or similarity of jobs in this case must be shown by such on site investigation to justify application of such a theory. (p. 1490)

In a concurring but separate opinion, one of the judges stated:

> The first major problem with a validity generalization approach is that it is radically at odds with *Albermarle Paper Co v. Moody* . . . and *Griggs v. Duke Power* and the EEOC *Guidelines,* all of which require a showing that a test is actually predictive of performance of a specific job. (p. 1499)

The judged concluded that "As a matter of law, Hunter's validity generalization is totally unacceptable under relevant case law and professional standard" (p. 1501). Landy (2003) noted that as a Circuit Court of Appeals opinion, this opinion carries much more weight than a trial court opinion. It is binding in the Sixth Circuit Court and would be influential in other circuit courts.

Two cases have been decided with courts giving deference to validity generalization research. In *Bernard v. Gulf Oil Corp.* (1989), the Fifth Circuit upheld the district court decision, which found in favor of the defendant based in part on Drs. Hunter and Sharf's testimony that the cumulative knowledge of the validity of the Bennett Mechanical Comprehension Test supported generalizing validity for selecting craft workers at Gulf refineries. Likewise, in *Taylor v. James River Corp.* (1989), the court gave deference to Dr. Sharf's testimony and ruled the Bennett Mechanical Comprehension Test was valid based on validity generalization evidence.

In sum, some previous court decisions have taken note of what I would refer to as the situational specificity doctrine of the *Uniform Guidelines* and have been unpersuaded by more recent scientific findings. Other decisions have accepted validity generalization scientific findings. The existing and growing disjunction between the *Uniform Guidelines* and the state of the science raises the question of why the federal enforcement agencies have not undertaken to revise the *Uniform Guidelines.*

Why Haven't the Uniform Guidelines Been Revised?

In this section, I offer speculation on why the *Uniform Guidelines* have not been revised. Given that the *Uniform Guidelines* are inconsistent with scientific knowledge, they are not serving their original intention of providing "a framework for determining the proper use of tests and other selection procedures" (Section 3.1). The *Principles* and *Standards* provide guidance on test validity that is consistent with the scientific findings.

Why have the *Uniform Guidelines* not been revised to be consistent with professional standards? It appears to the author that a primary use of the *Uniform Guidelines* is to pressure employers into using suboptimal selection methods in order to hire minorities and

Whites at approximately the same rates. If employers do not hire minorities at about the same rates as Whites, the *Uniform Guidelines* are invoked by enforcement agencies and plaintiffs to require the employer to prepare substantial validity documentation.

It is noted that a revision of the *Uniform Guidelines* would likely cause uncertainty for employers. Currently, employers know what to expect. Revised rules would likely be followed by new litigation serving to clarify ambiguities. For a time, employers would have a less clear understanding of the new rules. Also, a revision of the *Uniform Guidelines* could make matters worse. A new set of *Uniform Guidelines* might continue to ignore current scientific findings and end up even more problematic than the current set. However, the current *Uniform Guidelines* are close to thirty years old and are substantially disparate from scientific findings and other professional principles and standards. In other areas of federal regulations and guidelines, such regulations and guidelines are often updated to reflect scientific knowledge and professional practice. It is well past the time for the *Uniform Guidelines* to be revised.

Validity Generalization and Suggestions for Interpreting Validity Generalization in Support of Test Validity

To use generalization as a test validation strategy, one must find or conduct a meta-analytic study that is applicable to one's needs. Locating such studies is not difficult. Validity generalization studies have been conducted for cognitive ability tests (Hunter, 1980; Hunter & Hunter, 1984; Pearlman, Schmidt, & Hunter, 1980), assessment centers (Winfred, Day, & McNelly, 2003; Gaugler, Rosenthal, Thornton, & Benson, 1987), personality tests (Barrick & Mount, 1991; Hurtz & Donovan, 2000; Salgado, 1997; Tett, Jackson, & Rothstein, 1991), college grades (Dye & Reck, 1989), psychomotor tests (Hunter & Hunter, 1984), short-term memory tests (Verive & McDaniel, 1996), biodata instruments (Carlson, Scullen, Schmidt, Rothstein, & Erwin, 1998; Gandy, Dye, & MacLane, 1994; Rothstein, Schmidt, Erwin, Owens, & Sparks, 1990), customer service tests (Frei & McDaniel, 1998), interviews (Huffcutt & Arthur, 1994; McDaniel, Whetzel, Schmidt, & Maurer, 1994), training and experience assessments (McDaniel, Schmidt, & Hunter, 1988a,

1988b), length of job experience (McDaniel, 1986; McDaniel, Schmidt, & Hunter, 1988b), integrity tests (McDaniel & Jones, 1986; Ones, Viswesvaran, & Schmidt, 1993, 1995, 2003), and job knowledge measures (Dye, Reck, & McDaniel, 1993). Analyses have also been conducted for specific jobs or classes of jobs, including clerical (Pearlman, Schmidt, & Hunter, 1980), police (Hirsh, Northrop, & Schmidt, 1986), fire fighter (Barrett, Polomsky, & McDaniel, 1999), petroleum occupations (Schmidt, Hunter, & Caplan, 1981a, 1981b), and computer programmers (Schmidt, Gast-Rosenberg, & Hunter, 1980).

Considerations in Applying Validity Generalization Results as Support for Test Validation

In this section, I review issues that should be considered when drawing validity inferences about the local use of a test from a validity generalization study. Some of the issues raised should be viewed as the responsibility of the validity generalization researcher in preparing the meta-analysis report. Other issues are better viewed as the responsibility of the employer seeking to use a validity generalization study as evidence supporting test use and interpretation in a local setting. I also refer the reader to Sackett (2003) and the "Suggestions for enhancing the contribution of meta-analysis" section of Rothstein (2003).

Are the jobs in the meta-analysis comparable to the job for which the employer seeks test validity support? There are two issues here. The first issue is whether the job content matters in drawing conclusions about the validity of the test. The second is whether there are some jobs or classes of jobs for which the test has no validity.

If the validity generalization study targets a specific job or class of jobs and the job for which one seeks test validity support is not one of those jobs, one might question the usefulness of the validity generalization study. For example, Pearlman, Schmidt, and Hunter (1980) reported a meta-analysis of tests for clerical jobs. It would not be prudent to rely on that study to support the validity of a test for petroleum engineers. However, there is little evidence that small differences in the tasks performed across jobs within the same job family (Schmidt, Hunter, & Pearlman, 1980) result in sub-

stantive differences in relationships with relevant criteria. Many meta-analyses do not report validity differences between jobs or job classes (Gaugler et al., 1987; Hurtz & Donovan, 2000; McDaniel et al., 1994; McDaniel et al., 1988a, 1988b; McDaniel, Morgeson, Finnegan, Campion, & Braverman, 2001). Here, the task would be to argue that the local job is comparable to the jobs in the meta-analysis or that the content of the job is not relevant to the validity of the test.

Some meta-analyses have categorized jobs by the job's characteristics rather than its content. For example, some studies have grouped jobs by the degree of cognitive complexity, where a cognitively complex job is one that places substantial cognitive demands on the incumbent (Hunter & Hunter, 1984; Hunter, 1980; McDaniel, 1986). Thus, the job of rocket scientist has high cognitive complexity whereas the job of poultry eviscerator has lower cognitive complexity. Tests of general cognitive ability are more valid for jobs higher in cognitive complexity, although the validity of cognitive ability tests for jobs at the lowest level of cognitive complexity is still positive. The author knows of no validity generalization studies showing that a test of cognitive ability is valid for one job but not valid for another.

When determining the relevance of a validity generalization study for a local test validation, it would be useful if meta-analysis authors specified the types of jobs contributing data to the meta-analysis. The *Principles* have argued for clear specification of the scope of the meta-analysis. An example is provided of a meta-analysis of interviews, which concludes that "a cumulative database on interviews where content, structure, and scoring are coded could support generalization to an interview meeting the same specification" (p. 30). The *Principles* encourage that the boundary conditions of the meta-analysis be specified with respect to the content, structure, and scoring of the tests. It is suggested that it would be reasonable also to describe the scope of the jobs covered in the meta-analysis.

Is the test examined in the meta-analysis comparable to the test for which one seeks test validity support? Often it is clear that the tests analyzed in the meta-analysis are similar to the test for which one seeks test validation evidence. For example, Ones, Viswesvaran, and Schmidt (1993) listed the names of tests they analyzed in their

meta-analysis of integrity tests. Short of listing the names of tests in a meta-analysis, there might be clear evidence that the meta-analysis is relevant to an employer's validation needs. For example, measures of general cognitive ability are highly correlated with each other. Thus, if one seeks validation support for a measure of general cognitive ability, there are several validity generalization studies that document the validity of tests of general cognitive ability (Hunter, 1980; Hunter & Hunter, 1984; McDaniel, Schmidt, & Hunter, 1988a; Pearlman, Schmidt, & Hunter, 1980), so an employer could offer evidence that their test is a cognitive ability test—and that past validity generalization research has demonstrated that cognitive ability tests are valid for all jobs.

The answer to the question of whether a given employer's test is comparable to the tests in the validity generalization study is different when the meta-analysis addresses the validity of a method (for example, an interview, a situational judgment test, an assessment center) than when the meta-analysis addresses the validity of a construct (for example, cognitive ability, conscientiousness). In the case of a method, validity generalization has shown that following a specific procedure, such as a procedure to develop a structured interview, results in a valid measure. The *Principles* argue that inferences from a validity generalization study of a method are more complicated because the interviews may measure different constructs. Another point of view is that validity inferences are not dependent on knowing the constructs assessed; rather, the validity generalization study reflects the validity of a measure developed by following a procedure, such as a procedure for developing a job-related structured interview. From this perspective, the inference from the validity generalization of a method is not more constrained than the inferences drawn from validity generalization of a construct.

Can the reliability and range restriction corrections be defended as accurate? Hartigan and Wigdor (1989) reviewed the use of the GATB and validity generalization. Although validity generalization was accepted by this National Academy of Science Committee, they expressed some concern regarding assumed values of range restriction corrections. Hartigan and Wigdor also expressed some concerns over the assumed values of reliabilities used in corrections. I encourage meta-analytic researchers to specify the details of any assumed val-

ues of reliability and range restriction so that an employer using the study as a test validity defense can point to the reasonableness of the corrections. As noted by Sackett (2003), there is still some controversy about the appropriate use of some reliability estimates in meta-analytic corrections.

Are the criteria used in the meta-analysis appropriate for drawing validity inferences regarding a local test? For most meta-analyses, the answer to this question is straightforward because the criterion used in most meta-analyses is job performance, typically assessed through supervisor ratings. However, in some validity generalization analyses based on small numbers of studies, criteria may be grouped in combinations that might serve to hinder efforts to generalize the validity to a local test. For example, Hurtz and Donovan (2000) combined performance rating criterion data with objective sales data in their meta-analysis of personality test validity. Thus, if an employer wanted to use the Hurtz and Donovan data to support the validity of a personality test for a nonsales job, an adversary might argue that the data do not permit an inference to a nonsales job.

Is the meta-analysis sufficiently credible to use as a validity defense? Some meta-analyses are better than others. Consider, for example, an early meta-analysis of integrity tests by McDaniel and Jones (1986). The criteria were almost entirely self-report theft measures. Many would find self-report theft to be an inadequate criterion. Fortunately for those who seek validity generalization studies of integrity tests, more comprehensive studies have been conducted (Ones, Viswesvaran & Schmidt, 1993, 2003). The number of studies summarized in the meta-analysis might also limit its credibility. McDaniel, Schmidt, and Hunter (1988a) reviewed the validity of methods of evaluating training and experience. One type of training and experience (the task method) was based on only ten underlying studies. Although the meta-analytic summary provided the best available validity evidence regarding the task method at that time, an adversary might point out that the validity estimate is based on limited information. A rebuttal argument is that the ten studies provide ten times as much information as a local validation study.

Are the studies in the meta-analysis representative of all studies? Publication bias occurs when the effect sizes (correlations in the case of validity data) analyzed by the researcher are not representative

of all available effect sizes. Statistical methods for evaluating potential publication bias have been developed in the medical literature (Rothstein, Sutton, & Borenstein, 2005), but instances of their application to validity data are few. Vevea, Clements, and Hedges (1993) conducted a publication bias analysis of the GATB validity data and concluded that any bias present did not meaningfully affect the conclusion about the validity of cognitive ability tests. However, emerging research results on the employment interview are more unsettling. Readers of the validity generalization studies on the employment interview will likely conclude that structured interviews are substantially more valid than unstructured interviews. Duval (2005) analyzed the data from the McDaniel et al. (1994) meta-analysis of interviews and concluded that there was evidence of publication bias in the distribution of structured interviews such that the validity of structured interviews was overestimated. Although the Duval (2005) analysis is not the final word for the validity of structured employment interviews, it does suggest that it would be prudent for meta-analyses to incorporate publication bias analyses. McDaniel, Rothstein, and Whetzel (in press) applied publication bias methods to data in the technical manuals of three employment test vendors. Evidence of publication bias was found in the analyses of two of the four test vendors. Validity data had been selectively reported by the test vendors and served to overestimate the validity of the test. This unfortunate circumstance could be used by an adversary to question the value of any meta-analysis that has incorporated the vendor's data. I suggest that the validity generalization studies that have evaluated and ruled out publication bias in their data will offer more compelling evidence of validities than those that do not.

Summary and Recommendations

In sum, validity generalization provides reasonable and scientifically defensible evidence for the validity of an employment test and finds support in professional guidelines and research literature. However, the *Uniform Guidelines* were written prior to the development and acceptance of meta-analytically based validity generalization. In addition to encouraging the revision of the *Uniform Guidelines* to be consistent with the scientific knowledge, one can

argue that the they can and should be more broadly interpreted. As was noted in discussing the role of professional standards, to the extent that the *Uniform Guidelines* are in conflict with those *Principles* and *Standards,* they should be read in concert. Practitioners must exercise prudent professional judgment in interpreting both the regulatory and the scientific issues. The *Uniform Guidelines* state that they are intended to be consistent with professional guidelines and scientific knowledge. Employers can argue that because the validity inferences they draw from validity generalization studies are consistent with professional principles and scientific knowledge, the validity evidence should be judged as consistent with the underlying requirements of the *Uniform Guidelines.*

References

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education (1999). *Standards for Educational and Psychological Testing.* Washington, DC: American Educational Research Association.

Barrett, G. V., Polomsky, M. D., & McDaniel, M. A. (1999). Selection tests for firefighters: A comprehensive review and meta-analysis. *Journal of Business and Psychology, 13,* 507–514.

Barrick, M. R., & Mount, M. D. (1991). The Big Five personality dimensions and job performance: A meta-analysis. *Personnel Psychology, 44,* 1–26.

Bartlett, C. J., Bobko, P., Mosier, S. B., & Hannan, R. (1978). Testing for fairness with a moderated multiple regression strategy: An alternative to differential analysis. *Personnel Psychology, 31,* 233–241.

Bernard v. Gulf Oil Corporation, 890 F.2d 735, 744 (5th Cir. 1989).

Boehm, V. R. (1972). Differential prediction: A methodological artifact? *Journal of Applied Psychology, 62,* 146–154.

Bray, D. W., & Moses, J. L. (1972). Personnel selection. *Annual Review of Psychology, 23,* 545–576.

Carlson, K. D., Scullen, S. E., Schmidt, F. L., Rothstein, H., & Erwin, F. (1999, Fall). Generalizable biographical data validity can be achieved without multi-organizational development and keying. *Personnel Psychology, 52*(3), pp. 731–755.

Duval, S. J. (2005). The "trim and fill" method. In H. Rothstein, A. J. Sutton, & M. Borenstein (Eds.), *Publication bias in meta analysis: Prevention, assessment, and adjustments* (pp. 127–144). New York: Wiley.

Dye, D. A., & Reck, M. (1989). College grade point average as a predictor of adult success. *Public Personnel Management, 18,* 235–241.

Dye, D. A., Reck, M., & McDaniel, M. A. (1993). The validity of job knowledge measures. *International Journal of Selection and Assessment, 1,* 153–157.

Equal Employment Opportunity Commission, Civil Service Commission, Department of Labor, & Department of Justice. (1978). *Uniform guidelines on employee selection procedures. Federal Register, 43*(166), 38290–39315.

Equal Employment Opportunity Commission v. Atlas Paper Box Company, 680 F. Supp. 1184 (U.S. Dist. 1987).

Fiedler, A. M. (2001). Adverse impact on Hispanic job applicants during assessment center evaluations. *Hispanic Journal of Behavioral Sciences, 23,* 102–110.

Frei, R., & McDaniel, M. A. (1998). The validity of customer service orientation measures in employee selection: A comprehensive review and meta-analysis. *Human Performance, 11,* 1–27.

Friend et al. v. City of Richmond et al., 588 F.2d 61 (4th Cir. 1978).

Friend et al. v. Leidinger et al., 446 F. Supp. 361 (U.S. Dist. 1977).

Gandy, J., Dye, D., & MacLane, C. (1994). Federal government selection: The individual achievement record. In G. Stokes, M. Mumford, & G. Owens (Eds.), *Biodata handbook: Theory, research, and use of biographical information in selection and performance prediction* (pp. 275–310). Palo Alto, CA: CPP Books.

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

Gibson, W. M., & Caplinger, J. A. (2007). Transportation of validation results. In S. M. McPhail (Ed.), *Alternative validation strategies: Developing new and leveraging existing validity evidence* (pp. 29–81). San Francisco: Jossey-Bass.

Goldberg, L. R., Sweeney, D., Merenda, P. F., and Hughes, J. E. (1998). Demographic variables and personality: The effects of gender, age, education, and ethnic/racial status on self-descriptors of personality attributes. *Personality and Individual Differences, 24,* 393–403.

Goldstein, H. W., Yusko, K. P., & Braverman, E. P. (1998). The role of cognitive ability in the subgroup differences and incremental validity of assessment center exercises. *Personnel Psychology, 51,* 357–374.

Goldstein, H. W., Yusko, K. P., & Nicolopoulos, V. (2001). Exploring Black-White subgroup differences of managerial competencies. *Personnel Psychology, 54,* 783–807.

Guion, R. M. (1975). Recruiting, selection and job placement. In M. D. Dunnette (Ed.), *Handbook of industrial and organizational psychology* (pp. 777–828). Chicago: Rand McNally.

Hartigan, J. A., & Wigdor, A. K. (Eds.). (1989). *Fairness in employment testing: Validity generalization, minority issues, and the General Aptitude Test Battery.* Washington, DC: National Academy Press.

Hirsh, H. R., Northrop, L., & Schmidt, F. L. (1986). Validity generalization results for law enforcement occupations. *Personnel Psychology, 39,* 399–420.

Hough, L. M., Oswald, F. L., & Ployhart, R. E. (2001). Determinants, detection and amelioration of adverse impact in personnel selection procedures: Issues, evidence and lessons learned. *International Journal of Selection and Assessment, 9,* 152–194.

Huffcutt, A. I., & Arthur, W. (1994). Hunter and Hunter revisited: Interview validity for entry jobs. *Journal of Applied Psychology, 79,* 184–190.

Hunter, J. E. (1980). *Validity generalization for 12,000 jobs: An application of synthetic validity and validity generalization to the General Aptitude Test Battery (GATB).* Washington, DC: U.S. Department of Labor, Employment Service.

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. (2004). *Methods of meta-analysis: Correcting error and bias in research findings* (2nd edition). Newbury Park, CA: Sage.

Hurtz, G. M, & Donovan, J. J. (2000). Personality and job performance: The Big Five revisited. *Journal of Applied Psychology, 85,* 869–879.

Jeanneret, R. (2005). Professional and technical authorities and guidelines. In F. J. Landy (Ed.), *Employment discrimination litigation: Behavioral, quantitative, and legal perspectives* (pp. 47–100). San Francisco: Jossey-Bass.

Landy, F. J. (2003). Validity generalization: Then and now. In K. R. Murphy (Ed.), *Validity generalization: A critical review* (pp. 155–195). Mahwah, NJ: Erlbaum.

Kirkpatrick, J. J., Ewen, R. B., Barrett, R. S., & Katzell, R. A. (1968). *Testing and fair employment.* New York: New York University Press.

McDaniel, M. A. (1986). *The evaluation of a causal model of job performance: The interrelationships of general mental ability, job experience, and job performance.* Doctoral dissertation, The George Washington University.

McDaniel, M. A., & Jones, J. W. (1986). A meta-analysis of the employee attitude inventory theft scales. *Journal of Business and Psychology, 1,* 31–50.

McDaniel, M. A., Morgeson, F. P., Finnegan, E. B., Campion, M. A., & Braverman, E. P. (2001). Use of situational judgment tests to predict job performance: A clarification of the literature. *Journal of Applied Psychology, 86,* 730–740.

McDaniel, M. A., Rothstein, H., & Whetzel, D. L. (in press). Publication bias: A case study of four test vendors. *Personnel Psychology.*

McDaniel, M. A., Schmidt, F. L., & Hunter, J. E. (1988a). A meta-analysis of methods for rating training and experience in personnel selection. *Personnel Psychology, 41,* 283–314.

McDaniel, M. A., Schmidt, F. L., & Hunter, J. E. (1988b). Job experience correlates of job performance. *Journal of Applied Psychology, 73,* 327–330.

McDaniel, M. A., Whetzel, D., Schmidt, F. L., & Maurer, S. (1994). The validity of the employment interview: A comprehensive review and meta-analysis. *Journal of Applied Psychology, 79,* 599–616.

McKay, P., & McDaniel, M. A. (2006). A re-examination of Black-White mean differences in work performance: More data, more moderators. *Journal of Applied Psychology, 91,* 538–554.

Nguyen, N., McDaniel, M. A., & Whetzel, D. L. (2005, April). *Subgroup differences in situational judgment test performance: A meta-analysis.* Paper presented at the 20th Annual Conference of the Society for Industrial and Organizational Psychology, Los Angeles.

Ones, D. L., Viswesvaran, C., & Schmidt, F. L. (1993). Comprehensive meta-analysis of integrity test validities: Findings and implications for personnel selection and theories of job performance. *Journal of Applied Psychology, 78,* 679–703.

Ones, D. L., Viswesvaran, C., & Schmidt, F. L. (1995). Integrity tests: Overlooked facts, resolved issues, and remaining questions. *American Psychologist, 50,* 456–457.

Ones, D. L., Viswesvaran, C., & Schmidt, F. L. (2003). Personality and absenteeism: a meta-analysis of integrity tests. Personality and industrial, work and organizational applications. *European Journal of Personality, 17*(Suppl. 1), S19–S38.

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

Pegues v. Mississippi State Employment Service, 488 F. Supp. 239 (N.D. Miss. 1980), aff's 699 F.2d 760 (5th Cir. 1983), cert denied, 464 U.S. 991, 78 L. Ed. 2d 679, 104 S. CT. 482 (1983).

Roth, P. L., BeVier, C. A., Bobko, P., Switzer, F. S., III, & Tyler, P. (2001). Ethnic group differences in cognitive ability in employment and educational settings: A meta-analysis. *Personnel Psychology, 54,* 297–330.

Roth, P. L., Van Iddekinge, C. H., & Huffcutt, A. I. (2002). Corrections for range restriction in structured interview ethnic group differences: The values may be larger than researchers thought. *Journal of Applied Psychology, 87,* 369–376.

Rothstein, H. R. (2003). Progress is our most important product: Contributions of validity generalization and meta-analysis to the development and communication of knowledge in I/O Psychology. 115–154. In K. R. Murphy (Ed.), *Validity generalization: A critical review.* Mahwah, NJ: Erlbaum.

Rothstein, H. R., Schmidt, F. L., Erwin, F. W., Owens, W. A., & Sparks, C. P. (1990). Biographical data in employment selection: Can validities be made generalizable? *Journal of Applied Psychology, 75,* 175–184.

Rothstein, H. R., Sutton, A. J., & Borenstein, M. (2005). *Publication bias in meta-analysis: Prevention, assessment and adjustments.* West Sussex, England: Wiley.

Sackett, P. R. (2003). The status of validity generalization research: Key issues in drawing inferences from cumulative research findings. In K. R. Murphy (Ed.), *Validity generalization: A critical review* (pp. 91–114). Mahwah, NJ: Erlbaum.

Sackett, P. R., Schmitt, N., & Ellingson, J. E. (2001). High-stakes testing in employment, credentialing, and higher education: Prospects in a post-affirmative-action world. *American Psychologist, 56,* 302–318.

Salgado, J. F. (1997). The five factor model of personality and job performance in the European community. *Journal of Applied Psychology, 82,* 30–43.

Schmidt, F. L. (1988). The problem of group differences in ability test scores in employment selection. *Journal of Vocational Behavior, 33,* 272–292.

Schmidt, F. L., Gast-Rosenberg, I. F., & 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,* 529–540.

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

Schmidt, F. L., & Hunter, J. E. (1998). 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. (2003). History, development, evolution and impact of validity generalization and meta-analysis methods, 1975–2001.

Schmidt, F. L., Hunter, J. E., & Caplan, J. R. (1981a). *Selection procedure validity generalization (transportability) results for three job groups in the petroleum industry.* Washington, DC: American Petroleum Institute.

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

Schmidt, F. L., Hunter, J. E., & Pearlman, K. (1980). Task difference and validity of aptitude tests in selection: A red herring. *Journal of Applied Psychology, 66,* 166–185.

Schmidt, F. L., Hunter, J. E., Pearlman, K., & Hirsh, H. R. (1985). Forty questions about validity generalization and meta-analysis (with commentary by P. R. Sackett, M. L. Tenopyr, N. Schmitt, J. Kehoe, & S. Zedeck). *Personnel Psychology, 38,* 697–798.

Schmidt, F. L., Pearlman, K., & Hunter, J. E. (1980). The validity and fairness of employment and educational tests for Hispanic Americans: A review and analysis. *Personnel Psychology, 33,* 705–724.

Society for Industrial and Organizational Psychology. (2003). *Principles for the validation and use of personnel selection procedures* (4th ed.). Bowling Green, OH: Society for Industrial and Organizational Psychology.

Taylor v. James River Corp., 1989 WL 165953 (S.D. Ala. 1989).

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.

Van Aken et al. v. Young and City of Detroit et al., 541 Supp. 448 (U.S. Dist. 1982).

Verive, J. M., & McDaniel, M. A. (1996). Short-term memory tests in personnel selection: Low adverse impact and high validity. *Intelligence, 23,* 15–32.

Vevea, J. L., Clements, N. C., & Hedges, L. V. (1993). Assessing the effects of selection bias on validity data for the General Aptitude Test Battery. *Journal of Applied Psychology, 78,* 981–987.

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

Winfred Jr., A., Day, E. A., & McNelly, T. L. (2003). A meta-analysis of the criterion-related validity of assessment center dimensions. *Personnel Psychology, 56,* 125–154.