CHAPTER

9 Meta-Analysis as a Validity Summary Tool

George C. Banks *and* Michael A. McDaniel

Abstract

The chapter discusses the role of meta-analysis in enhancing the understanding of employment test validity. We discuss the state of validity knowledge prior to the introduction of meta-analysis and summarize the gains in knowledge following the introduction of meta-analysis. We review the standards of systematic literature reviews, data typically reported in a meta-analysis of a personnel selection test, and how meta-analytic findings are interpreted. Furthermore, we consider the differences between the meta-analysis of selection tests that evaluate specific constructs and those that assess selection test methods that measure multiple constructs. We discuss issues to consider when evaluating the degree to which meta-analytic reviews of validity data have credibility and how to make decisions regarding the appropriateness of the application of a selection test. Finally, we discuss the need to improve reporting practices in meta-analytic reviews as well as the inconsistencies of the *Uniform Guidelines on Employee Selection Procedures* with scientific knowledge concerning meta-analysis and validity.

Key Words: validity generalization, meta-analysis, psychometric meta-analysis, validity, criterion-related validity, personnel selection

Introduction

The primary purpose of an employment test is to screen applicants based on inferences about future performance. Empirical evidence indicates that individual differences evaluated through personnel assessment methods have important implications for job performance and the financial value of the employees' performance for the organization (Hunter, Schmidt, & Judiesch, 1990). This chapter describes how researchers acumulate research results using meta-analysis and how this aggregation of research results can be used by organizations to better inform their use of selection procedures.

What Is Meta-Analysis and What Is Validity Generalization?

The term "meta-analysis" was first introduced by Gene Glass (1976) to "refer to the statistical analysis of a large collection of analysis results from individual studies for the purpose of integrating the findings" (p. 3). Validity generalization is the use of meta-analytic techniques to explore the generalizability of the correlation (validity) between employment test scores and outcome variables, such as job performance, performance in training, and tenure (Rothstein, McDaniel, & Borenstein, 2002), across various situations in which an employment test might be used. A validity generalization analysis estimates the mean population validity and the variance in the population validity. One therefore can conclude that a test demonstrates validity generalization when the large majority (typically 90% or more) of the validity estimates between the test and the criterion of interest (e.g., job performance) are above zero.

Note that the definition of validity generalization does not mean that all population validity

estimates are the same. Typically, there is some variability remaining that may be due to differences across studies on a third variable (i.e., a moderator) such as characteristics of the job or the situation. For example, the validity of cognitive ability for job performance typically shows validity generalization and some of the remaining variability in validity coefficients is due to the moderating effect of the cognitive complexity of the job (Hunter & Hunter, 1984). Likewise, conscientiousness typically shows validity generalization and some evidence indicates that the correlation between the test and job performance is higher for jobs with greater autonomy (Barrick & Mount, 1993). Depending on one's perspective, the autonomy moderator is either a characteristic of the job or a characteristic of the situation. The remaining variance may also be due to differences across studies in nonmoderator sources of variance that were not corrected in the validity generalization study (e.g., reporting errors in the studies that contributed data to the meta-analysis).

In this chapter, we use the phrase "employment test" to refer to any type of procedure used to screen job applicants. Thus, an employment test could be an interview or a résumé review in addition to a cognitive ability or personality test. We use the phrase "validity coefficients" to refer to correlations between an employment test and job performance.

Chapter Overview

We begin the chapter with a review of the status of personnel selection knowledge prior to validity generalization. This review provides background context for the objectives and challenges faced by meta-analytic researchers. In subsequent sections, we review the principles of a systematic literature review, data commonly reported in a meta-analysis of an employment test, and how these results are interpreted. We discuss the distinction between the meta-analyses of employment tests that assess specific constructs (e.g., cognitive ability, conscientiousness) and the analyses of employment test methods (e.g., interviews, assessment centers) that measure multiple constructs. This chapter also describes the interpretation of estimated mean population validities and variances. We review considerations when evaluating the extent to which meta-analytic summaries of validity data are credible. We also discuss issues to consider when using validity generalization results to make decisions about the appropriateness of a test in a particular application. We highlight reporting practices that need improvement and the inconsistency of the *Uniform Guidelines on Employee Selection Procedures* (Equal Employment Opportunity Commission, Civil Service Commission, Department of Labor, & Department of Justice, 1978) with current science as reflected in professional guidelines and peer-reviewed literature.

The Dark Ages: Life before Meta-Analysis

Beginning in the early twentieth century, research findings indicated that the same employment test could yield different validity estimates for predicting job performance, even when computed for the same job in similar settings (Schmidt & Hunter, 1998). The conclusion drawn was that there were as-yet-undiscovered attributes of the situations (e.g., the occupational contexts) that influenced the magnitude and direction of the correlations between employment tests and job performance. Detailed job analyses were unable to identify these situational attributes (Schmidt & Hunter, 1998). In other words, it was often observed that an employment test used to hire a teller for a bank branch on Main Street yielded a different validity coefficient than when used to hire a teller for a bank branch on Broad Street. This phenomenon came to be called the *situational specificity hypothesis.* The apparent implication of situational specificity was that organizations that wished to use employment tests had to conduct a validation study (e.g., examine the correlation between the employment test and job performance) for each job and in each setting for which they wished to use an employment test.

As late as the 1970s, the assumption of situational specificity was accepted as a fact (Schmidt & Hunter, 2003) and there was a strong reliance on small sample, local validity studies. Consequently, job analysts and researchers alike refrained from making firm statements about personnel selection methods. In addition, it was difficult to accumulate knowledge concerning the best selection procedures. As such, the notion of situational specificity retarded the growth and development of our knowledge of personnel selection for decades.

Let There Be Light: The Genesis of Meta-Analysis

In the late 1970s, Schmidt and Hunter (1977) challenged the situational specificity hypothesis by suggesting that the validity of personnel selection methods varied across studies due to statistical artifacts. They proposed that the validity of employment

tests is largely stable across organizations. This stability became evident when researchers corrected for variance in study results caused by statistical artifacts. The conclusion was that local validation studies are not routinely needed every time an organization wanted to apply selection methods.

One of the major contributions of this work was the observation that much of the variation across applications in the validity of the personnel selection methods was caused by simple random sampling error (sampling error is one type of statistical artifact). Random sampling error occurs when a study sample is not representative of the population from which the sample was drawn. The relatively small samples that had been used in past validity studies resulted in substantial random sampling error that caused validity coefficients to appear unstable across situations. Thus, even if the test had a constant validity in the population (e.g., the correlation between the employment test and job performance was always 0.50), random sampling error might result in a validity coefficient being 0.10 in one application of the employment test and 0.70 in another. Furthermore, there was variance in the findings of local validity studies that was caused by differences across studies in measurement error and range restriction. This artifactual variance contributed to the apparent instability of validity results across situations. Measurement error and range restriction also caused the observed validity coefficients to underestimate their population parameter (e.g., the "true" validity).

Schmidt and Hunter (1977) developed methods that could correct for variance across studies due to sampling error, measurement error, and range restriction. The methods also permitted the estimation of the population or true validity of employment tests. When the variability in population validity indicated that most validities would be positive in future applications, the employment test was considered to have validity generalization. This indicated that the validity would generalize across most applications in which the test might be used.

Early validity generalization studies demonstrated the value of validity generalization for several jobs. For example, Pearlman, Schmidt, and Hunter (1980) showed validity generalization of several predictors in the selection of clerical workers. This study showed that differences across job tasks had very little influence on the validity of employment tests. Another study that aided in the acceptance of the meta-analysis of employment tests was a validity generalization study conducted using the General Aptitude Test Battery (GATB) in the context of 12,000 jobs (Hunter, 1980; Hunter & Hunter, 1984). This finding demonstrated the robustness of general cognitive ability and psychomotor ability employment tests across jobs and established the value of validity generalization analyses in the accumulation of knowledge in personnel selection. These studies suggested that the emphasis on detailed job analyses, which was common in the field and incorporated into the *Uniform Guidelines on Employee Selection Procedures* (Equal Employment Opportunity Commission et al., 1978), was likely misguided.

A second major advancement that greatly assisted the acceptance of meta-analysis was an article by Schmidt, Hunter, Pearlman, and Hirsh (1985) that contained a question and answer dialogue. In the article titled "Forty Questions About Validity Generalization and Meta-Analysis," Schmidt et al. addressed major critiques directed at validity generalization and meta-analysis. This publication was a major turning point in the acceptance of the meta-analysis methods applied to employment test validity.

Two Major Methods for Conducting a Meta-Analysis

During the 1970s and 1980s, researchers across disciplines were working independently, and nearly at the same time, on the foundations of what has come to be known as meta-analysis (Glass & Smith, 1979; Hedges & Olkin, 1985; Rosenthal & Rubin, 1978; Schmidt & Hunter, 1977). The approach offered by Schmidt and Hunter (1977) became known as psychometric meta-analysis (Hunter & Schmidt, 1990) and its use in showing the magnitude and relative stability of validity across situations is called validity generalization.

The primary meta-analytic methods used today are (1) psychometric (Hunter & Schmidt, 1990) and (2) meta-analyses in the tradition of Hedges and Olkin (1985). More recent versions of these two approaches are represented by *Methods of Meta-Analysis: Correcting Error and Bias in Research Findings* (Hunter & Schmidt, 2004) and *Introduction to Meta-Analysis* (Borenstein, Hedges, Higgins, & Rothstein, 2009). Both of these methodological approaches focus on estimating the population distribution of studies. Both meta-analysis approaches recognize that correlations (and other effect sizes) vary from study to study due to random sampling

error. Psychometric meta-analysis also explicitly considers other statistical artifacts whereas meta-analyses in the Hedges and Olkin tradition typically do not. In the next section, we will review in greater depth the factors that limit inferences from primary validity studies such as sampling error, measurement error, and range restriction.

Factors That Limit Inferences from Primary Validity Studies

Understanding the effects of statistical artifacts on validity coefficients is the key to understanding validity generalization. Here, we present an overview of sampling error, measurement error, and range restriction.

What Is Random Sampling Error?

Random sampling error is the difference between a sample statistic and a population parameter from which the sample was derived. It is a major constraint on the usefulness of an individual study and influences our ability to estimate a population correlation. To determine the correlation between an employment test and job performance, one wants to know the population correlation (known as the population parameter, rho, which is symbolized as ρ). The preliminary estimate of the population correlation is obtained by calculating an "observed" correlation in a sample. Due to random sampling error, the observed correlation in the sample may overestimate or underestimate the population correlation. The nature of the relation between the size of a sample and random sampling error is such that as the size of a sample increases, the magnitude of the error of a sample decreases in an asymptotic manner. The decrease in sampling error is much more dramatic when increasing the sample size from 100 to 200 than when increasing the sample size from 1,000 to 1,100. Because of this relation between sample size and sampling error, larger samples are more representative of a population than smaller samples. That is, on average, large samples provide better estimates of the population correlation than smaller samples.

Our ability to identify a population correlation is also a function of the magnitude of that effect. The stronger a population effect (e.g., a correlation or standardized mean difference), the smaller the sample required to detect that effect can be. For example, a researcher would not need to sample a large number of people to discover that the majority of Americans would rather eat a slice of pizza than a handful of dirt.

Consider the following illustration of random sampling error. Imagine a bag that includes 21 red poker chips and nine white poker chips ($N = 30$). The ratio of red poker chips to white poker chips is 7:3. Now imagine that you shake the bag to mix the chips and remove a random sample of 10 poker chips ($n = 10$). This sample may yield the true ratio of seven red poker chips and three white poker chips. However, if you were to replace the 10 poker chips, shake the bag again, and remove a second random sample you might select six red poker chips and four white poker chips. In a third sample, you might select eight red poker chips and two white poker chips. The phenomenon that you would be witnessing is random sampling error.

With enough random samples, the mean of the distribution of all your samples would come to reflect the actual distribution of red to white poker chips in the bag (7:3), but there would be substantial variability in results across samples. This variability is entirely due to random sampling error. This poker chip example demonstrates the constraint imposed by random sampling error on any single, primary validity study. Some samples will underestimate the number of red poker chips and other samples will overestimate the number of red poker chips. Likewise, some samples will yield an observed correlation that underestimates the population correlation and other samples will yield correlations that overestimate the population correlation. This limits what can be concluded from any single study. However, one can see the benefit of the application of a validity generalization study that quantitatively aggregates primary sample results. Because sampling error is random, with enough samples, an average of results across samples will yield the correct population information (the correct ratio of red to white poker chips in the population). Likewise, with enough samples, an average of the observed validity coefficients will yield a mean observed validity coefficient that is not distorted by random sampling error. However, the mean observed validity coefficient will still be an underestimate of the population correlation due to measurement error and range restriction.

Formulas exist to estimate the amount of random sampling error in a study (Borenstein et al., 2009; Hunter & Schmidt, 2004). The formulas account for the size of the sample and the estimated magnitude of the population effect. Sampling error is summarized with a statistic called the standard error. A standard error is used in all statistical significance

tests and is also, perhaps more appropriately, used in creating confidence intervals around sample statistics.

What Is Measurement Error?

In addition to random sampling error, psychometric meta-analysis permits correction for measurement error. Measures do not have perfect reliability. For example, two interviewers will typically not have exactly the same evaluations of a job applicant. Schmidt and Hunter (1996) stated, "every psychological variable yet studied has been found to be imperfectly measured, as is true throughout all other areas of science" (p. 199). Measurement error always distorts the observed correlation, which results in an underestimate of the population correlation.[1] Therefore, it is important that research studies minimize the influence of measurement error on validity coefficients.

Consider a concurrent validity study in which an organization attempts to measure the relation between conscientiousness and job performance of current employees (e.g., job incumbents). The objective is to identify the correlation between conscientiousness and job performance in order to inform decisions about the use of a conscientiousness measure as a screening tool and predictor of future job performance of job applicants. Measures of conscientiousness are, however, not perfect. In addition, there is substantial measurement error in supervisory ratings (e.g., a mean interrater reliability of 0.52 for supervisors' rating of overall job performance; Viswesvaran, Ones, & Schmidt, 1996). Because it is not possible to have perfectly reliable measures of conscientiousness and job performance, the observed correlation will be underestimated due to measurement error.

The reason for this underestimation may be explained using classical true score theory:

$$O = T + E$$

This formula indicates that the observed score (O) is the result of the sum of the true score (T) and error (E). Similarly, the observed variance is the sum of the true score variance and the measurement error score variance σ^2:

$$\sigma^2_O = \sigma^2_T + \sigma^2_E$$

Because measurement error is random with a mean of zero, it will not affect the mean of an observed score, on average. However, it will result in increased variance in a set of observed scores because the measurement error variance component of the observed variance is always positive. Whereas the measurement error variance component of the observed variance is random, its correlation with other variables is zero, on average. This causes the correlation between two variables to underestimate the population correlation due to the random measurement error variance component of the observed variance. Thus, the observed correlation between two variables, for example cognitive ability and job performance, will underestimate its population correlation due to measurement error in both the cognitive ability measure and the job performance measure. The observed correlation is said to be attenuated by measurement error. To estimate the population correlation, it is necessary to correct for attenuation. As is the case with random sampling error, validity generalization allows for the correction of measurement error so that the population mean correlation can be estimated more accurately. Also, correction for measurement error removes variance due to differences in reliability across studies permitting an improved estimate of the population variance.

Note, however, that organizations are tasked with selecting applicants based on the operational validity of a predictor variable (Hunter, Schmidt, & Le, 2006). When a predictor (e.g., conscientiousness) is used to select applicants, it will not be free of error. Thus, in meta-analyses applied to employment tests, the population mean estimate does not include a correction for measurement error in the predictor, but is corrected for measurement error in the outcome measure (e.g., job performance).

What Is Range Restriction?

Another major artifact in personnel selection research is range restriction. Validity coefficients (e.g., correlation coefficients between an employment test and job performance) are influenced by the range of the variables. To calculate the correlation between an employment test and job performance, it is necessary to have a measure of job performance. Such a measure is available only for those who are hired. Those who are hired almost always have a smaller variance of employment test scores than the full application pool because the applicants with the lowest employment test scores are not hired. As a result, the correlation between the employment test and job performance will underestimate the correlation for the full applicant

pool. To estimate the value of an employment test in screening an applicant pool, it is necessary to correct the correlation based on those hired for range restriction in the employment test scores.

In addition, it must be determined whether the range restriction is direct or indirect. Direct range restriction occurs when an organization selects applicants based solely on their ranked performance on an assessment test. For instance, if an organization were to select applicants based solely on their cognitive ability scores, range restriction would be direct. In the instance of indirect range restriction, an organization selects applicants based on their performance on an assessment test as well as other variables. If an organization selects applicants based on their cognitive ability scores, as well as letters of recommendation, a job interview, and a résumé, indirect range restriction will occur on the cognitive ability test. In most validity generalization studies, range restriction will be indirect as organizations usually do not select applicants based on their ranked performance on any single measure. Because the traditional formula for indirect range restriction requires information usually unavailable, typically many researchers applied the correction for direct range restriction even in cases of indirect range restriction, which led to an underestimation of the population validity coefficient (Hunter et al., 2006). However, Hunter et al. (2006) provided a method for correcting correlations for indirect range restriction. As with corrections for measurement error, corrections for range restriction increase the accuracy of both the population correlation and its variance.

Recognition of statistical artifacts and their effects. We have now discussed the three most common types of statistical artifacts that are corrected in psychometric meta-analysis (e.g., sampling error, measurement error, and range restriction). Artifacts fall into two categories: unsystematic sources of error that do not have a consistent biasing direction (i.e., random sampling error), and systematic sources of error that result in a downwardly biased estimate of the observed correlation (e.g., measurement error and range restriction). The result of both unsystematic and systematic sources of error is that observed validity coefficients are different from the value of the population parameter (the population correlation). Other types of artifacts include scale coarseness (Aguinis, Pierce, & Culpepper, 2009) for Likert scales, dichotomization of continuous variables, deviation from perfect construct validity, and reporting or transcriptional errors. The cumulative effect of the artifacts is that the distribution of observed validity coefficients almost always underestimates the population mean[2] and overestimates the population variance. Meta-analysis procedures correct for many of the artifacts to provide more accurate estimates of the population correlation and its variance.

Information Commonly Reported in a Meta-Analysis of an Employment Test

Although we argue that psychometric meta-analyses are superior to meta-analyses in the Hedges and Olkin (1985) tradition due to corrections for measurement error and range restriction, meta-analyses in the psychometric tradition typically have some nonoptimal reporting practices relative to meta-analyses in the Hedges and Olkin tradition. Improvements in psychometric meta-analyses are needed in several areas. We detail what should be, but typically is not, reported in validity generalization studies and other meta-analyses in the psychometric tradition. We note that the *Publication Manual of the American Psychological Association* (2010) incorporated reporting requirements for meta-analysis. We know of no past validity generalization paper or psychometric meta-analysis that is consistent with these APA style requirements. Thus, we detail these style requirements in our description of a systematic review with the desire to encourage psychometric meta-analyses consistent with the APA requirements.

Departures from Principles of Systematic Literature Reviews

A systematic review is a summary of literature that is organized in an objective manner so as to identify all relevant studies. Furthermore, a systematic review is one that documents its steps in such a fashion that it can be replicated by others. All validity generalization studies should be conducted as systematic reviews, which requires researchers to make use of and report the protocol used to conduct their study. A protocol is a plan for conducting a meta-analysis that states what is being done and why. Protocols include the decision rules that are used and are ideally created prior to the start of a meta-analysis, but may be updated as needed. The information from the protocol is included in the journal article. A protocol spells out the decisions taken during each of the steps of a meta-analysis (Cooper, 1982). These steps include (1) question formulation, (2) data collection, (3) data evaluation,

(4) data analysis, and (5) reporting of the results. When researchers do not use a protocol, they risk departing from the principles of a scientific study by limiting the ability of others to replicate the meta-analysis. Relative to meta-analysis in other disciplines, validity generalization studies often do a poor job in step two (data collection and its documentation) and are hampered in step three by limitations of the primary studies' reporting deficiencies (e.g., failure to report data needed to evaluate range restriction).

In the formulation stage, a researcher will specify the question that the meta-analysis is attempting to address. Thus, the researcher specifies the predetermined characteristics of the samples, the design, and the population to be investigated (Berman & Parker, 2002). This is often referred to as specifying the inclusion criteria or decision rules. It is important to report these decisions in the methods section of the meta-analysis. For example, if a researcher conducts a validity generalization study of conscientiousness as a predictor of job performance *only* in high complexity jobs, it is important for that researcher to specify this in the protocol (and the resulting journal article). This allows the readers of the paper to recognize the aim of the study and the limitations of the applicability of the findings. In other words, the results of the study may not generalize to lower complexity jobs. This recommendation is consistent with the *Principles for the Validation and Use of Personnel Selection Procedures* (Society for Industrial and Organizational Psychology, 2003), which emphasizes setting boundary conditions of the meta-analysis, particularly for meta-analyses of employment test methods (e.g., interviews; discussed in more detail later in this chapter).

The second step of a protocol guides researchers on how to collect studies (Cooper, 1982). For example, what search terms or key words should be used when searching electronic databases or when sending out calls for papers over email listservs. The types of terms used should be based on the question the study is seeking to address. Validity generalization studies should be replicable. Therefore, reporting the search terms used is very important. The following two excerpts can be used to contrast a poor example and a better example of reporting the steps taken to systematically search the literature.

The literature search for a meta-analysis by Hoffman, Blair, Meriac, and Woehr (2007) summarized its search in one sentence:

> We conducted a search of the OCB literature by using a number of online databases (e.g., Web of Science, PsycINFO) as well as by examining the reference lists of previous reviews (p. 557).

We consider this to be a poor description of the literature review. Compare that excerpt to a systematic literature search described by Williams, McDaniel, and Nguyen (2006), who wrote:

> We began with an automated search of PsycINFO (Psychological Abstracts) and ABI/Inform using the key words *compensation satisfaction, pay satisfaction, compensation equity, pay equity, compensation fairness, and pay fairness.* We also searched manually 12 journals for the years 1960 through 2003: *Academy of Management Journal, Administrative Science Quarterly, Human Relations, Industrial and Labor Relations Review, Industrial Relations, Journal of Applied Psychology, Journal of Management, Journal of Organizational Behavior, Journal of Occupational and Organizational Psychology, Journal of Vocational Behavior, Organizational Behavior and Human Decision Processes,* and *Personnel Psychology* (p. 396).

After reading the first excerpt the reader may not have initially recognized what information the authors failed to report. However, after reading the quote from Williams et al. (2006), the reader can clearly tell that the latter example is more transparent, explicit, and replicable. This comparison should highlight the importance of documenting all the steps taken to conduct a systematic review.

The third step of the protocol involves the coding of the studies included in the meta-analysis. Decision rules must be clearly stated in the protocol so that later, when the information is reported in the methods section of the study, readers can understand the processes used to aggregate the data. For example, primary studies often do a poor or inconsistent job of reporting their findings. For example, information related to reliability is often not reported. Or researchers often do not report the means and standard deviations for variables that are not of primary interest in their study (means and standard deviations of job performance by race or sex), but may be a primary interest for a particular meta-analysis (e.g., a meta-analysis of subgroup differences in job performance). It is also possible that primary authors alter items on validated scales or they do not administer a test in a manner that is consistent with instructions from the test vendor (Banks, Batchelor, & McDaniel, 2010). This issue makes it difficult to correct for range restriction because of the lack of population variance estimates

that apply to the measure administered in a nonstandard way.

The result is that the meta-analytic researcher is often unable to identify important artifact statistics needed to correct for measurement error and range restriction. If the researcher is unable to contact the author of the primary study to obtain the missing data, that researcher must make decisions (guided by the question being tested) regarding how to deal with the missing data. It is important that meta-analytic researchers report coding decisions and fully disclose their steps so that other researchers can critically evaluate their decisions.

The fourth step of a protocol involves the analysis of data (Cooper, 1982). Here, researchers need to report the steps used to analyze their data. Therefore, if a researcher uses the psychometric (Hunter & Schmidt, 2004) or Hedges and Olkin (1985) approach to conduct the meta-analysis, it is important to report the technique as well as any other analytic techniques used. In summary, it is critical to report these first four steps to ensure that the study can be replicated.

The APA reporting standards detail the expectations for the reporting of the results of a meta-analysis. The reporting of the results is equally important as describing the steps used to obtain the findings being reported. Table 9.1, published by the APA Publications and Communications Board Working Group on Journal Article Reporting Standards (2008) and later incorporated into the 6th edition of the *Publication Manual of the American Psychological Association* (2010), outlines the main paper sections and topics that should be considered when reporting a meta-analysis.

Each paper section and topic includes a description of the information that is recommended for inclusion in the manuscript. For instance, it is recommended that the title indicate that the study was a research synthesis or actually include "meta-analysis." Abstracts should summarize the major points of the paper, such as the research question, the eligibility criteria, and the types of primary studies included. The introduction should state the population under investigation, the rationale for exploring certain moderators, as well as the strengths and weaknesses of the study's design. Also, the meta-analytic researchers should report any funding sources to ensure transparency in the event there is a conflict of interest.

The methods section should describe the eligible research populations and design features as well as the operationalized characteristics of the predictor and outcome variables. It should also indicate how moderators and mediators are coded, and how the literature has been searched. This includes describing which keywords are used and which listservs are queried. Coding procedures should include information for how missing data (e.g., reliability estimates) were handled and how interrater agreement was evaluated. Statistical methods should address whether a fixed or random-effects-model was used, how the effect sizes were weighted, and how heterogeneity (i.e., variance not due to random sampling error) was assessed. Although unusual in a validity generalization study, the sixth edition of the *Publication Manual of the American Psychological Association* (2010) requires a listing of studies including, the correlation, and the sample size.

Multiple analyses should be run, if possible, under different scenarios to demonstrate the robustness of the results. These analyses are often characterized as sensitivity analyses (Borenstein et al., 2009). One can have greater confidence if the conclusions do not change as a function of the sensitivity analyses. As an example of a sensitivity analysis, a researcher with 10 effect sizes may repeat an analysis 10 times, each time excluding a different sample's data (Borenstein et al., 2009). Such a sensitivity analysis would evaluate the extent to which conclusions might change based on the inclusion of a single study. Also, results could be reported with and without certain studies. Such studies might be ones that appear nontypical (e.g., a large sample that has an outlier effect size) and it is necessary to determine if the study has an undue influence on the results. Or, one could report results with and without certain artifact corrections. For example, McDaniel, Whetzel, Schmidt, and Maurer (1994) reported estimates of validity with and without range restriction corrections due to concerns about the quality of range restriction information. Publication bias analyses, to be discussed later in the chapter, are also useful sensitivity analyses. Finally, the discussion section should state the major findings, consider multiple alternatives for the explanation of the results, as well as evaluate the generalizability of the conclusions and the general limitations of the study. Meta-analytic discussion sections are also quite valuable when they provide guidelines for future research.

The use of visual displays should be increased in validity generalization studies. The majority of research published in industrial/organizational

Table 9.1 Meta-Analysis Reporting Standards: Information Recommended for Inclusion in Articles Reporting Meta-Analyses.

Article Section and Topic	Description
Title	Make it clear that the report describes a research synthesis and include "meta-analysis," if applicable Footnote funding source(s)
Abstract	The problem or relation(s) under investigation Study eligibility criteria Type(s) of participants included in primary studies Meta-analysis methods (indicating whether a fixed or random model was used) Main results (including the more important effect sizes and any important moderators of these effect sizes) Conclusions (including limitations) Implications for theory, policy, and/or practice
Introduction	Clear statement of the question or relation(s) under investigation Historical background Theoretical, policy, and/or practical issues related to the question or relation(s) of interest Rationale for the selection and coding potential moderators and mediators of results Types of study designs used in the primary research, their strengths and weaknesses Populations to which the question or relation is relevant Hypotheses, if any
Method inclusion and exclusion criteria	Operational characteristics of independent (predictor) and dependent (outcome) variable(s) Eligible participant populations Eligible research design features (e.g., random assignment only, minimal sample size) Time period in which studies needed to be conducted Geographical and/or cultural restrictions
Moderator and mediator analyses	Definition of all coding categories used to test moderators or mediators of the relation(s) of interest
Search strategies	Reference and citation databases searched Registries (including prospective registries) searched: Key words used to enter databases and registries Search software used and version Time period in which studies needed to be conducted, if applicable Other efforts to retrieve all available studies: Listservs queried Contacts made with authors (and how authors were chosen) Reference lists of reports examined Method of addressing reports in languages other than English Process for determining study eligibility: Aspects of reports that were examined (i.e., title, abstract, and/or full text) Number and qualifications of relevance judges Indication of agreement How disagreements were resolved Treatment of unpublished studies
Coding procedures	Number and qualifications of coders (e.g., level of expertise in the area, training) Intercoder reliability or agreement Whether each report was coded by more than one coder, and if so, how disagreements were resolved

(*continued*)

Table 9.1 (Continued)

Article Section and Topic	Description
	Assessment of study quality: If a quality scale was employed, a description of criteria and the procedures for application If study design features were coded, what these were How missing data were handled
Statistical methods	Effect size metric(s): Effect size calculating formulas (e.g., *M*s and *SD*s, use of univariate *F* to *r* transform) Corrections made to effect sizes (e.g., small sample bias, correction for unequal *n*s) Effect size averaging and/or weighting method(s) How effect size confidence intervals (or standard errors) were calculated How effect size credibility intervals were calculated, if used How studies with more than one effect size were handled Whether fixed and/or random effects models were used and the model choice justification How heterogeneity in effect sizes was assessed or estimated *M*s and *SD*s for measurement artifacts, if construct-level relationships were the focus Tests and any adjustments for data censoring (e.g., publication bias, selective reporting) Tests for statistical outliers Statistical power of the meta-analysis Statistical programs or software packages used to conduct statistical analyses
Results	Number of citations examined for relevance List of citations included in the synthesis Number of citations relevant on many but not all inclusion criteria excluded from the meta-analysis Number of exclusions for each exclusion criterion (e.g., effect size could not be calculated), with examples Table giving descriptive information for each included study, including effect size and sample size Assessment of study quality, if any Tables and/or graphic summaries: Overall characteristics of the database (e.g., number of studies with different research designs) Overall effect size estimates, including measures of uncertainty (e.g., confidence and/or credibility intervals) Results of moderator and mediator analyses (analyses of subsets of studies): Number of studies and total sample sizes for each moderator analysis Assessment of interactions among variables used for moderator and mediator analyses Assessment of bias including possible data censoring
Discussion	Statement of major findings Consideration of alternative explanations for observed results: Impact of data censoring Generalizability of conclusions: Relevant populations Treatment variations Dependent (outcome) variables Research designs General limitations (including assessment of the quality of studies included) Implications and interpretation for theory, policy, or practice Guidelines for future research

Adapted and reproduced with permission from APA Publications and Communications Board Working Group on Journal Article Reporting Standards (2008).

psychology journals use only a tabular form to display their findings and therefore do not benefit from graphic displays of results (Rothstein et al., 2002). Rothstein et al. (2002) recommend stem and leaf plots and forest plots to complement the commonly used tabular forms in order to demonstrate the direction and magnitude of effect sizes. Forest plots are designed for the presentation of meta-analytic data, in that these plots display the effect size for each study as well as the confidence intervals or credibility values around the effect size (Rothstein, 2003). An example of the value added by such visual displays is demonstrated in an article by Rothstein et al. (2002) that visually presented results previously reported in Kubeck et al. (1996) indicating the study names next to the corresponding effect size, sample size, correlation coefficient, and confidence interval. Such practices should be used more commonly in the reporting of validity generalization results so that the interpretation of the study's findings is clearer to both researchers and a more general audience.

In validity generalization studies, corrections are typically made to the observed effect sizes to estimate the population effect more accurately. Situations exist in which researchers disagree on the techniques used and the extent to which observed effects are corrected. In keeping with the aim of transparency and full disclosure of results, it is important for researchers to report both the (uncorrected) mean observed correlations and the (corrected) estimated mean population correlations.

Table 9.2, adapted from McDaniel et al. (1994), demonstrates the manner in which such information can be displayed. The first column identifies the groups from which the data being presented are drawn. In the example in Table 9.2, the information presented is based on an analysis of all the interview data with the criterion of job performance. The *N* column shows how many individuals were included in each analysis. This *N* is the sum of the sample sizes across all the correlations in the analysis. In this article, the number of correlations is denoted with No. *rs* (although in more recent meta-analyses, the symbol *k* is typically used to indicate the number of samples). The mean *r* and Obs σ columns indicate the observed mean correlation (the mean correlation, which has not been corrected for measurement error and range restriction) and standard deviation. The observed σ has not been corrected for sampling error or for differences across studies in measurement error and range restriction. Next, results are presented for analyses that were corrected for measurement error, but did not include range restriction corrections. Results were then presented for the data that included corrections for range restriction. Reporting results with and without range restriction corrections is not a requirement. It was done in this case because range restriction data used in this particular study were scant and the authors sought to demonstrate that the analyses supported the validity of the employment interview, regardless of whether range restriction corrections were used.

The last three columns are used to report the summary statistics for the population distribution. The symbol ρ (rho) is the estimate of the population distribution mean (the "true" relation), σ_ρ is the estimated standard deviation of the population distribution, and the 90% credibility interval is the bottom 10th percentile of the population distribution. Given that the 90% credibility interval (CV) is a positive value, one would assert that the validity of employment interviews generalizes across situations. In other words, when one uses employment interviews, one could expect positive validities in more than 90% of the applications of the test. The advantage of reporting the results in this manner is that it allows the reader to review the observed validity, the validity corrected for measurement error, and the validity corrected for both measurement error and range restriction.

Table 9.2 Information to Be Reported in a Meta-Analysis.

					Without Range Restriction Corrections			With Range Restriction Corrections		
Interview distribution	*N*	No. *rs*	Mean *r*	Obs σ	ρ	σ_ρ	90% CV	ρ	σ_ρ	90% CV
All interviews	25,244	160	0.20	0.15	0.26	0.17	0.04	0.37	0.23	0.08

Reproduced from McDaniel et al. (1994) with permission of Elsevier.

The Detection of Moderating Variables

Variance in the estimated population validity suggests that there are other variables at play that may influence the validity. To the extent that there is true variance in a population, it influences our ability to interpret the estimated mean of the population validity distribution. For example, Hunter and Hunter (1984; Hunter, 1980) documented the validity of general cognitive ability tests that varied by the cognitive complexity of the job. When validity coefficients from jobs of varying complexity are combined, the population variance is larger than if all jobs contributing data were of the same level of cognitive complexity. When population variance estimates are large, the population mean estimate becomes less informative. For example, the degree to which a job is cognitively complex influences the validity of general cognitive ability (Hunter & Hunter, 1984). In other words, cognitive ability has a validity of 0.56 in highly complex jobs, but only 0.23 in low complexity jobs (Schmidt & Hunter, 1998). Without the consideration of the moderator, the validity would have been reported as being some value between 0.23 and 0.56 and thus, would have been an inaccurate estimate of the validity for both low complexity jobs and high complexity jobs.

There are several methods to detect the presence of moderators. One method advocated by Hunter and Schmidt (2004) evaluates whether sampling error and other statistical artifacts account for at least 75% of the observed variance. If artifactual variance does not account for at least 75% of the observed variance, it is necessary to look for moderators if theoretical, logical, or knowledge-based justification is available. Another approach for identifying moderators is the *Q*-statistic, which is a chi-square difference test that is used to test the statistical significance of potential variance due to moderators. Finally, one might use the estimated standard deviation of the population distribution as a means to detect the presence of moderators. The credibility interval uses the estimated standard deviation of the population distribution to express the variance that might be attributable to moderators.

One validity generalization study that demonstrated the importance of moderators examined the validity of employment interviews (McDaniel et al., 1994). The findings of this study demonstrated that interview validity was moderated by whether the interviews were structured or unstructured and the content of the interview (e.g., situational, job related, or psychological). For instance, the mean validity of structured interviews was 0.44, as compared to 0.33 for unstructured interviews. The validities of the interview content that were situational, job related, and psychological were 0.50, 0.39, and 0.29, respectively.

Consideration in Relying on Validity Generalization to Support Test Use

This section examines issues to be considered when interpreting the meta-analysis results of employment tests to support test use. There are two sets of issues. One concerns the extent to which the meta-analysis results are conducted sufficiently well to accept their conclusions. The second set of issues concern the extent to which it is possible to draw inferences from the meta-analysis to guide decisions about the use of a test in a specific situation.

Meta-Analysis Credibility as a Function of the Reasonableness of Nonsampling Error Artifact Data

A primary advantage of psychometric meta-analysis is that it permits a more accurate estimate of population correlations by correcting for the statistical artifacts of measurement error and range restriction. Concerns may arise when the reliability data or range restriction data are missing from one or more of the primary studies contributing correlations to the analysis. Meta-analyses often impute the missing reliability data based on knowledge of the reliability of the scale as reported in other studies. For example, if the reliability of scale A is reported in the range of 0.78 to 0.82 in studies that reported reliabilities, it would appear reasonable to assume a reliability of 0.80 for studies that did not report the reliability.

An exception to this approach would occur when the "scale" consists of a single item. Here, one might use the Spearman–Brown formula to estimate the reliability of a single item based on the reliability of a multi-item scale that measures the same construct. Also, it is possible to rely on meta-analyses of the reliability of a measure. For example, there are meta-analyses of the reliability of interviews (Conway, Jako, & Goodman, 1995) and supervisor ratings (Rothstein, 1990; Viswesvaran et al., 1996).

When imputing statistical artifact data, confidence in the meta-analysis can be increased through sensitivity analyses. For example, the analysis could be conducted with and without corrections for measurement error and/or range restriction to determine if conclusions change. McDaniel et al. (1994)

estimated the population validities of the employment interview with and without corrections for range restriction.

Also, it is possible to compare the mean observed correlations of studies that did and did not report artifact data. If the mean observed correlations in the two groups are similar, one might have increased confidence in the imputation by arguing that the two sets of correlations were likely to be drawn from the same population and are subject to about the same level of attenuation, as evidenced by the similar means. For example, McDaniel (2005) compared studies with and without range restriction data reported and observed that the mean observed correlations in the two sets of studies were similar. Thus, McDaniel (2005) was able to use this analysis to support the assertion that range restriction data that were reported could be used to impute range restriction data that were not reported.

When most studies report data needed for artifact corrections, one will typically want to conduct a meta-analysis in which correlations are corrected individually (Hunter & Schmidt, 2004). In validity generalization studies when there is measurement error and indirect range restriction, measurement error would be corrected in both the predictor and criterion variables. Next, indirect range restriction corrections would be made (Hunter et al., 2006). Finally, each correlation would be uncorrected for measurement error in the predictor in order to estimate the operational relation between predictor and criterion variables.

If most studies do not report data needed for artifact corrections, one will typically conduct a meta-analysis using artifact distributions (Hunter & Schmidt, 2004). The first step in this analysis is to create four distributions: one composed of the reported correlations, one containing estimates of the reliabilities of the predictor, a third containing estimates of the reliabilities of the criterion, and a fourth distribution consisting of estimated values needed to correct for range restriction. The result is four distributions each with four means and four variances (Hunter & Schmidt, 2004). This artifact distribution meta-analysis assumes that artifact distributions reflect the artifacts that are attenuating the observed correlation and the credibility of the meta-analysis rests on the accuracy of the assumption. A series of research articles, which included Monte Carlo studies, reported evidence supporting the accuracy of this approach (Hunter & Schmidt, 1994; Law, Schmidt, & Hunter, 1994a, 1994b).

Meta-Analysis Credibility as a Function of Data Source

When considering the extent to which meta-analysis results are credible, one consideration is the source of the data. Concerns about the source of the data reflect at least three issues. The first issue concerning data source is whether the data are primarily from one author or organization. Consider the meta-analysis of the validity of ratings of training and experience (T&E; McDaniel, Schmidt, & Hunter, 1988). A review of the data listed in Appendix A in McDaniel et al. indicates that 10 of the 15 validity coefficients analyzed for the behavioral consistency method (better known as the "accomplishment record" method) were from studies by Leaetta Hough and all were from a prestigious consulting firm in which she worked. The mean of the estimated population validity was 0.45. It may be that the rigor of the Hough et al. studies is higher, perhaps substantially higher, than many applications of the behavioral consistency method. Some anecdotal evidence available to the authors indicates that sometimes the behavioral prompts to which the applicants respond in some applications of the method are less well developed than in the case of the Hough validity studies. Likewise, raters may not always be as well trained as in the Hough studies and the reliability of the ratings may, therefore, be less. Thus, it is possible that the validity estimate reported by McDaniel et al. may be an overestimate of the typical validity for this T&E method. We might have greater confidence in the validity estimate offered by McDaniel et al. if they had obtained other validity studies from a more diverse set of authors. In the same study, the estimated mean population validity of the point method of T&E evaluation was 0.11. However, 51 of the 91 coefficients analyzed were from a single paper (Molyneaux, 1953) and the mean validity of those studies was only 0.06. Perhaps there is something unique about the Molyneaux data that made it unrepresentative of typical applications of the point method. We note that the McDaniel et al. article remains the most comprehensive review of the T&E validity literature and we are not arguing that it is incorrect. However, we do argue that confidence in the conclusions of a meta-analysis should be greater, on average, when data are obtained from a diverse set of sources than when data are mostly from one or a few authors.

A second issue with respect to source of the data is whether we trust the data. When a test vendor offers an employment test for sale, the test vendor

has a financial motivation to make available only the most favorable results. McDaniel, Rothstein, and Whetzel (2006) found evidence consistent with the inference that some test vendors suppressed validity data that were not supportive of their product. When meta-analyses are based almost entirely on test vendor supplied data (integrity tests; Ones, Viswesvaran, & Schmidt, 1993), some might have serious concerns about the credibility of the meta-analysis results. Note that we do not argue that all test vendor data are subject to suppression. Rather, we argue that one should consider how much the data can be trusted when evaluating the credibility of a meta-analysis.

The third issue concerning data source relates to the outlet in which the data became available. That is, were the data from a journal article or from another outlet such as a conference paper or a dissertation? As noted by Lipsey and Wilson (1993), the magnitude of effect sizes (e.g., correlations) from dissertations are typically smaller than results from published articles. Thus, a meta-analysis of the Big-Five (Hurtz & Donovan, 2000) drawn solely from English-language journal articles and a few conference papers might be less credible than a paper that draws data from additional sources (e.g., dissertations, technical reports, non-English journals). We do not wish to disparage the Hurtz and Donovan study. We do argue that their results would be more credible if data had been obtained from other sources in addition to journals and conferences.

We have offered concerns about three issues related to data sources used in meta-analysis: (1) the majority of data are from only a few sources, (2) data are from sources one may not trust, and (3) data are from selected outlets such as restricting data to published studies. All of these concerns could be framed with respect to our next category of concerns, publication bias.

Meta-Analysis Credibility as a Function of Evidence Concerning Publication Bias

A key consideration in judging the credibility of a meta-analysis is whether the conclusions of the study are robust to potential publication bias. Publication bias is present when the set of studies summarized in the meta-analysis is not representative of all the studies (Banks & McDaniel, 2011; McDaniel et al., 2006; Rothstein, Sutton, & Borenstein, 2005). Publication bias is better referred to as availability bias because studies can be unavailable for a variety of reasons. However, we will use the term publication bias to be consistent with the literature. Reasons for publication bias often include practices in the editorial and review processes, language barriers (e.g., studies published in foreign language journals), behaviors by authors, and the proprietary nature of research completed within some organizations.

Research in the medical literature has indicated that publication bias is typically a function of an author decision (Dickersin, 2005). A common scenario for publication bias stems from small sample studies in which the results are statistically nonsignificant. The author of such a study may give priority to working on other studies that have a higher likelihood of being published. The study with insignificant findings may never be published or otherwise made available. As a result, researchers who conduct a meta-analysis might have easier access to studies that are statistically significant than to studies that are not significant. An additional example is bias due to selective publication in a new and rapidly developing literature. Medical research indicates that the earliest effect sizes (e.g., correlation coefficients) are often larger than effect sizes obtained in later time periods (Ioannidis, 1998, 2005; Trikalinos & Ioannidis, 2005). This phenomenon may be due to a time-lag bias, such that the time to publication is shorter for statistically significant effects than for statistically insignificant effects (Ioannidis, 1998, 2005; Stern & Simes, 1997; Trikalinos & Ioannidis, 2005). The time-lag bias could also include the Proteus effect (i.e., studies with large effects are published earlier because they are more dramatic and more interesting; Trikalinos & Ioannidis, 2005). Under either explanation, validity studies in relatively new literatures (e.g., conditional reasoning tests; Banks, Kepes, & McDaniel, 2011) may be subject to a bias, such that initial findings overestimate the validity of a test.

We view publication bias as a major concern for meta-analyses of validity data because there are a large number of unpublished studies in the area of personnel selection (Rothstein et al., 2002). It is extremely rare for a meta-analysis of validity data to include publication bias analyses. We are more hopeful for future studies because the *Publication Manual of the American Psychological Association* (2010) now encourages publication bias analyses in a meta-analysis. It is our hope that our journals start enforcing this requirement and that past validity generalization studies be examined for publication bias.

Meta-Analysis Credibility as a Function of Unexplained Variance

Interpretations of meta-analytic summaries of validity data often focus on the mean of the estimated population validity distribution. Unexplained variance in the estimated population validities may be a function of errors that the author cannot correct (e.g., data reporting errors in studies contributing data to the meta-analysis) and may not be of concern for the credibility of the meta-analysis. However, the unexplained variance may also be a function of a moderator that has implications for the credibility of the results. Sometimes the authors of a meta-analysis may be asked to remove a moderator from a paper due to an editor's legitimate concerns over journal space. For example, McDaniel, Whetzel, Schmidt, and Maurer (1994) originally submitted the interview meta-analysis paper showing that the job performance validities were substantially smaller for police occupations than for other occupations. Faced with page limit constraints, the editor had the authors remove the discussion and tables associated with the police–not police moderator. As a result of that decision, the knowledge of the moderator did not enter the scientific literature. We are not seeking to criticize the editor; in his position, we may have made the same decision. We do believe that journals need to consider the substantial impact that meta-analyses can have and try to balance the need for full reporting of moderator analyses with page constraints. Many journals have moved in the direction of permitting additional information and analyses to be placed on the journal's web site. We suggest that journals in our research literatures adopt this practice.

A more difficult concern is moderators that have not yet been discovered and reported. For example, the McDaniel et al. (1994) meta-analysis of employment interviews did not consider study design (predictive versus concurrent) as a moderator of validity. It was 10 years later before it was noted that concurrent validity studies of the interview yielded validities 0.10 higher than predictive studies (Huffcutt, Conway, Roth, & Klehe, 2004). For validity coefficients, a difference of 0.10 is a large moderator. We suggest that predictive versus concurrent design can be an important moderator for any test, primarily noncognitive (e.g., personality, integrity, situational judgment) tests, in which applicant faking is likely to be an issue. Unfortunately, most validity data in our field are concurrent data. Meta-analyses of predictors relying on concurrent validity data may be the best available estimates of validity, but may eventually be shown to overestimate, perhaps substantially, the validity of the measures.

Meta-Analysis Credibility as a Function of the Number of Effect Sizes

The final consideration we offer in judging the credibility of a meta-analysis of validity data is the number of coefficients. Estimates of the mean population validity are more credible, on average, when they are based on a large number of studies. Also, meta-analyses of new predictors (e.g., conditional reasoning tests) may not initially yield a large number of studies for analysis and the studies may be subject to a time-lag bias (Banks et al., 2011; Ioannidis, 1998, 2005; Stern & Simes, 1997; Trikalinos & Ioannidis, 2005) that may result in an overestimate of the mean population validity.

Summary of Credibility of Meta-Analysis Issues

Our discussion of issues to consider when judging the credibility of a meta-analysis are not meant to discourage the conduct of meta-analyses in personnel selection research or the use of meta-analyses in drawing conclusions concerning the validity of a test. A meta-analytic review is far more useful in drawing conclusions about the validity of a test than any primary study or narrative review of studies. However, we do argue that it is necessary to evaluate the credibility of a meta-analysis before accepting its conclusions. We also argue that meta-analysts should consider drawbacks of past meta-analyses of validity data and design their meta-analyses to avoid these.

Considerations in Using Meta-Analyses to Draw Conclusions about a Test in a Specific Application

Assuming one has decided that the results of a meta-analysis are credible for drawing conclusions about the validity of an employment test, one must consider the usefulness of the meta-analysis for making decisions about a specific test in a specific application. One consideration rests on whether the meta-analysis summarized the validity of a predictor *construct* or a *method* (Arthur & Villado, 2008; Hunter & Hunter, 1984). An example of an employment test that measures only one construct is a measure of cognitive ability. Employment interviews, assessment centers, and situational judgment tests can best

be classified as methods that can and typically do measure multiple constructs. Thus, an employment interview is a method because it may be designed to measure both oral communication ability and conscientiousness.

When deciding on whether the validity reported in a meta-analysis is a good estimate of the validity that can be expected with a specific test in a specific application, the decision is easiest when the meta-analysis addressed a single construct and the test being considered measures the same construct. Thus, if the meta-analysis considered measures of general cognitive ability and the test under consideration is a measure of general cognitive ability, the applicability of the meta-analysis findings for the application of the test is clear.

Broader constructs may create some inference problems. A variety of personality-based measures claim to measure customer service (Frei & McDaniel, 1998). Unlike measures of general cognitive ability that can be shown to be highly correlated, there is less evidence of this for customer service tests. A decision maker considering the use of a specific customer service test may wish to compare the content of the test under consideration to the tests summarized in the meta-analysis. To the extent that the test is similar to those summarized in the meta-analysis, one could rely on the meta-analysis in drawing inferences about the likely validity of the test.

In brief, meta-analyses of measurement methods, such as employment interviews, are less straightforward to apply to specific testing decisions (*Principles*, p. 30). Inferences are complicated by different applications of a method (e.g., an employment interview) that may measure different constructs. For example, one interview may primarily assess, among other topics, a knowledge construct (e.g., auto-repair) and another interview may primarily assess conscientiousness and agreeableness. Even if two employment interviews were designed to measure the same constructs (e.g., conscientiousness and agreeableness), one interview evaluation may weigh one construct more heavily than another. Thus, although the meta-analyses of the interview may help identify characteristics of interviews that enhance validity (Huffcutt & Arthur, 1994), the validity of the interview in a specific application will likely be more approximately estimated from a meta-analysis than a validity for a general cognitive ability test derived from a validity generalization study of general cognitive ability tests.

Future Directions

There are two final issues that we consider critical for the advancement of validity generalization. These include poor reporting practices by primary researchers and, finally, yet another call for the revision or abolishment of the *Uniform Guidelines*. This discussion serves as a review of the future directions of validity generalization given the current state of the literature.

Poor reporting practices by primary study authors. Advances have been made in the data analysis techniques of meta-analysis. However, this does not mean that data analysis techniques can overcome poor reporting practices in primary studies. In fact, it is often a contribution of a meta-analytic study to provide constructive guidance on how primary researchers in a literature area can improve their research methodology and reporting practices.

Researchers who engage in primary studies can also improve their reporting of results and research methods to aid meta-analytic researchers. In general, primary researchers and journals need to make sure that their studies adhere to the *Publication Manual of the American Psychological Association* (American Psychological Association, 2010). However, there are several specific items that should be highlighted for their importance for validity generalization studies.

First, primary researchers need to report correlations regardless of statistical significance, direction, or magnitude, consistent with the *Principles for the Validation and Use of Personnel Selection Procedures* (see p. 52; Society for Industrial and Organizational Psychology, 2003). Failing to report this information is a common practice in some test vendor manuals as the vendors may wish to present the tests as more valid than they are. Also, primary researchers should report a correlation matrix with *all* the variables used in their study. Correlation matrices should include the sample size, as often participants drop out of studies or do not provide full data.

Second, primary researchers need to report the appropriate reliability for each of their measures. For instance, coefficient alphas often are reported as the reliability of situational judgment tests, despite the fact that these tests are construct heterogeneous and coefficient alpha is, therefore, inappropriate (McDaniel, Hartman, Whetzel, & Grubb, 2007). Likewise, coefficient alpha is an inappropriate reliability for supervisor ratings of job performance (Schmidt & Hunter, 1996).

Third, primary researchers should clearly describe the measures that they use in their study. This includes reporting the full citation for their measures, the exact number of items used, the exact response scale (e.g., a 7-point Likert scale for a personality measure), and the reliability of the measure identified in their study. In short, primary researchers need to describe carefully all the measures that they use.

Fourth, primary researchers should report information needed to identify range variation in the applicant pool. This includes the means and standard deviations of their variables.

Revision or abolishment of *Uniform Guidelines.* Near the dawn of the age of meta-analysis, the *Uniform Guidelines* were published. The *Uniform Guidelines* were written in a period in which the situational specificity theory was still accepted as true by some, and concerns regarding differential validity and differential prediction influenced its formation (McDaniel, 2007). For example, the *Uniform Guidelines* advocated (1) local validation studies, (2) differential validity and prediction studies, and (3) detailed and costly job analysis data (McDaniel, 2007).

The *Uniform Guidelines* fail to acknowledge the science that has been accumulated for decades indicating mean racial differences on some assessment tests. Rather than acknowledging mean racial differences in employment tests as an unfortunately common occurrence, the *Uniform Guidelines* view this as a call for a local validation study. The resulting impact of these flawed *Uniform Guidelines* is that employers are encouraged to use less valid selection tests to avoid potential adverse impact in their hiring procedures. The result can be a substantial loss of human capital due to poor job performance.

The inefficiencies and inadequacies of the *Uniform Guidelines* did not go unnoticed. Not long after the implementation of the guidelines, the Society for Industrial and Organizational Psychology (SIOP) wrote a letter to the agencies responsible for the guidelines indicating how the guidelines were flawed (McDaniel, 2007). To this day, the *Uniform Guidelines* have not been revised and remain flawed and inconsistent with professional practice and accepted scientific evidence.

Unlike the *Uniform Guidelines,* professional associations composed of both scientists and practitioners have provided the field with updated guidance that is current and recognizes the overwhelming evidence supporting validity generalization (McDaniel, 2007). Both the *Standards for Educational and Psychological Testing* (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 1999) and the *Principles for the Validation and Use of Personnel Selection Procedures* (Society for Industrial and Organizational Psychology, 2003) provide relevant and up-to-date guidance. Both the *Standards* and *Principles* recognize that validity generalization can provide useful validity information. The *Uniform Guidelines* do not recognize the scientific advances in meta-analysis and validity generalization, and substantially overestimate the value of local validity studies. Unfortunately, federal regulations still require both public and private sector organizations to abide by these archaic guidelines and the guidelines still carry weight in employment litigation (McDaniel, 2007).

One of the reasons, perhaps, for the persistent use of the *Uniform Guidelines* is to encourage employers to hire racial minorities at close to the same rate as whites regardless of the validity of the selection measure (McDaniel, 2007). For instance, research indicates that there are mean racial differences in cognitive ability (Jensen, 1998). Cognitive ability is the most valid predictor of job performance. However, if cognitive ability is used to select applicants, organizations will likely hire racial minorities and whites at disparate rates. When an organization does not hire at the same rate, the *Uniform Guidelines* require the employer to provide extensive evidence of the validity of a selection measure to avoid a law suit or fines by enforcement agencies. Thus, organizations may ignore more valid selection methods to avoid negative ramifications. The result is a loss in the competitive advantage that can be gained by hiring only the best employees.

Conclusions

In this chapter, we have discussed how meta-analysis can be used to estimate the validity of employment tests. We began with a review of personnel selection before meta-analysis. In particular, we discussed the notion of situational specificity, such that the validity of an employment test in one organization did not appear to generalize to another organization. Thus, prior to validity generalization studies, there was an emphasis on local validity studies. As a result, it was difficult to accumulate knowledge and advance theory concerning the relation between predictors and employee work outcomes.

In the mid-to-late 1970s, meta-analysis was introduced into different research areas. Schmidt and Hunter (1977), in particular, introduced validity generalization, which is an application of meta-analysis to employment test validity data. With the introduction of meta-analysis, researchers were able to consider and correct for the effects of artifacts such as sampling error, measurement error, and range restriction. Researchers can now correct for artifacts and more accurately estimate the validity of employment tests.

Also in this chapter, we reviewed data that are commonly reported in the meta-analysis of an employment test. This includes the importance of reporting the protocol a researcher used to conduct a systematic review and accurately reporting results in a manner in which a reader can understand. We described the difference between employment test constructs and employment test methods. We also discussed issues when interpreting mean validities and estimating population variances. These issues relate to the limitations in the inferences that can be made based on corrections due to the presence of artifacts, publication bias, and potential moderating variables. We concluded this chapter with a discussion of the importance of improving the reporting of meta-analytic protocols and results. Finally, we have added yet another voice to the call for the revision or abolishment of the *Uniform Guidelines*.

In conclusion, validity generalization has contributed a great deal to the advancement of both theory and practice related to personnel selection. Prior to the introduction of validity generalization, researchers were unable to accumulate knowledge and advance theory related to the validity of employment tests. Although there is room for improvement in the way that validity generalization studies of personnel selection are conducted and reported, meta-analysis remains a very powerful and valuable tool in understanding the validity of employment tests.

Notes

1. Although random measurement error always operates to cause the observed correlation to underestimate the population correlation, a given observed correlation may be lower or higher than the population correlation due to the influence of another artifact. For example, random sampling error will, about half the time, cause a correlation to be an overestimate of the population correlation. Thus, for a given correlation, random measurement error will bias the correlation to underestimate the population correlation, but an opposing sampling error may cause the observed correlation to overestimate the population correlation. Also, certain study designs result in range enhancement resulting in an upward bias on the observed correlation coefficient.

2. The primary exception to this general finding occurs when researchers draw a sample containing only those with the very highest job performance and those with the very lowest job performance. This creates a situation in which the employment test variance is larger than the applicant pool variance. This situation is known as range enhancement and the resulting validity coefficient is biased in the direction of overestimating the population correlation. This practice is seen among unethical test vendors who seek to make their test appear more valid than its population validity.

References

Aguinis, H., Pierce, C. A., & Culpepper, S. A. (2009). Scale coarseness as a methodological artifact. *Organizational Research Methods, 12*, 623–652. doi: 10.1177/1094428108318065.

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.

American Psychological Association. (2010). *Publication manual of the American Psychological Association* (6th ed.). Washington, DC: American Psychological Association.

APA Publications and Communications Board Working Group on Journal Article Reporting Standards. (2008). Reporting standards for research in psychology: Why do we need them? What might they be? *American Psychologist, 63*, 848–849. doi: 10.1037/0003-066X.63.9.839.

Arthur, W., & Villado, A. J. (2008). The importance of distinguishing between constructs and methods when comparing predictors in personnel selection research and practice. *Journal of Applied Psychology, 93*, 435–442. doi: 10.1037/0021-9010.93.2.435.

Banks, G. C., Batchelor, J. H., & McDaniel, M. A. (2010). Smarter people are (a bit) more symmetrical: A meta-analysis of the relationship between intelligence and fluctuating asymmetry. *Intelligence, 38*, 393–401. doi: 10.1016/j.intell.2010.04.003.

Banks, G. C., Kepes, S., & McDaniel, M. (2011). *Publication bias and the validity of conditional reasoning tests*. Paper presented at the 26th Annual Conference of the Society for Industrial and Organizational Psychology. Chicago.

Banks, G. C., & McDaniel, M. A. (2011). The kryptonite of evidence-based I-O psychology. *Industrial and Organizational Psychology: Perspectives on Science and Practice, 4*, 40–44.

Barrick, M. R., & Mount, M. K. (1993). Autonomy as a moderator of the relationships between the Big Five personality dimensions and job performance. *Journal of Applied Psychology, 78*, 111–178. doi: 10.1037/0021-9010.78.1.111.

Berman, N. G., & Parker, R. A. (2002). Meta-analysis: Neither quick nor easy. *BMC Medical Research Methodology, 2*, 10–19. doi: 10.1186/1471-2288-2-10.

Borenstein, M., Hedges, L. V., Higgins, J. P. T., & Rothstein, H. R. (2009). *Introduction to meta-analysis*. West Sussex: John Wiley & Sons. Ltd.

Conway, J. M., Jako, R. A., & Goodman, D. F. (1995). A meta-analysis of interrater and internal consistency reliability of selection interviews. *Journal of Applied Psychology, 80*, 565–579. doi: 10.1037/0021-9010.80.5.565.

Cooper, H. (1982). Scientific guidelines for conducting integrative research reviews. *Review of Educational Research, 52*, 291–302. doi: 10.3102/00346543052002291.

Dickersin, K. (2005). Recognizing the problem, understanding its origins and scope, and preventing harm. In H. Rothstein, M. Borenstein, & A. J. Sutton (Eds.), *Publication bias in meta analysis: Prevention, assessment and adjustments* (pp. 11–34). West Sussex: John Wiley & Sons.

Equal Employment Opportunity Commission, Civil Service Commission, Department of Labor, & Department of Justice. (1978). Uniform Guidelines on Employee Selection Procedures. *Federal Register, 43*, 38290–39315.

Frei, R., & McDaniel, M. (1998). The validity of customer service orientation measures in employee selection: A comprehensive review and meta-analysis. *Human Performance, 11*, 1–27. doi: 10.1207/s15327043hup1101_1.

Glass, V. G. (1976). Primary, secondary, and meta-analysis of research. *Educational Researcher, 5*, 3–8. doi: 10.3102/0013189X005010003.

Glass, V. G., & Smith, M. L. (1979). Meta-analysis of research on class size and achievement. *Educational Evaluation and Policy Analysis, 1*, 2–16. doi: 10.3102/01623737001001002.

Hedges, L. V., & Olkin, I. (1985). *Statistical methods for meta-analysis*. San Diego: Academic Press.

Hoffman, B. J., Blair, C. A., Meriac, J. P., & Woehr, D. J. (2007). Expanding the criterion domain? A quantitative review of the OCB literature. *Journal of Applied Psychology, 92*, 555–566. doi: 10.1037/0021-9010.92.2.555.

Huffcutt, A. I., & Arthur, W. (1994). Hunter and Hunter (1984) revisited: Interview validity for entry-level jobs. *Journal of Applied Psychology, 79*, 184–190. doi: 10.1037/0021-9010.79.2.184.

Huffcutt, A. I., Conway, J. M., Roth, P. L., & Klehe, U. (2004). The impact of job complexity and study design on situational and behavior description interview validity. *International Journal of Selection & Assessment, 12*, 262–273. doi: 10.1111/j.0965-075X.2004.00282.x.

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

Hunter, J. E., & Hunter, R. F. (1984). Validity and utility of alternative predictors of job performance. *Psychological Bulletin, 96*, 72–98. doi: 10.1037/0033-2909.124.2.262.

Hunter, J. E., & Schmidt, F. L. (1990). *Methods of meta-analysis: Correcting error and bias in research findings*. Newbury Park, CA: Sage Publications.

Hunter, J. E., & Schmidt, F. L. (1994). Estimation of sampling error variance in the meta-analysis of correlations: Use of average correlation in the homogeneous case. *Journal of Applied Psychology, 79*, 171–177. doi: 10.1037/0021-9010.79.2.171.

Hunter, J. E., & Schmidt, F. L. (2004). *Methods of meta-analysis: Correcting error and bias in research findings*. Thousand Oaks, CA: Sage Publications.

Hunter, J. E., Schmidt, F. L., & Judiesch, M. K. (1990). Individual differences in output variability as a function of job complexity. *Journal of Applied Psychology, 75*, 28–42. doi: 10.1037/0021-9010.75.1.28.

Hunter, J. E., Schmidt, F. L., & Le, H. (2006). Implications of direct and indirect range restriction for meta-analysis methods and findings. *Journal of Applied Psychology, 91*, 594–612. doi: 10.1037/0021-9010.91.3.594.

Hurtz, G. M., & Donovan, J. J. (2000). Personality and job performance: The Big Five revisited. *Journal of Applied Psychology, 85*, 869–679. doi: 10.1037//0021-9010.85.6.869.

Ioannidis, J. P. (1998). Effect of the statistical significance of results on the time to completion and publication of randomized efficacy trials. *Journal of the American Medical Association, 279*, 281–286. doi: 210.1001/jama.1279.1004.1281.

Ioannidis, J. P. (2005). Differentiating biases from genuine heterogeneity: Distinguishing artifactual from substantive effects. In H. Rothstein, A. J. Sutton, & M. Borenstein (Eds.), *Publication bias in meta-analysis: Prevention, assessment, and adjustments* (pp. 287–302). West Sussex, UK: John Wiley & Sons.

Jensen, A. R. (1998). *The g factor: The science of mental ability*. Westport, CT: Praeger.

Kubeck, J. E., Delp, N. D., Haslett, T. K., & McDaniel, M. A. (1996). Does job-related training performance decline with age? *Psychology and Aging, 11*, 92–107. doi: 10.1037/0882-7974.11.1.92.

Law, K. S., Schmidt, F. L., & Hunter, J. E. (1994a). Nonlinearity of range corrections in meta-analysis: Test of an improved procedure. *Journal of Applied Psychology, 79*, 425–438. doi: 10.1037/0021-9010.79.3.425.

Law, K. S., Schmidt, F. L., & Hunter, J. E. (1994b). A test of two refinements in procedures for meta-analysis. *Journal of Applied Psychology, 79*, 978–986. doi: 10.1037/0021-9010.79.6.978.

Lipsey, M. W., & Wilson, D. B. (1993). The efficacy of psychological, educational, and behavioral treatment: Confirmation from meta-analysis. *American Psychologist, 48*, 1181–1209. doi: 10.1037/0003-066X.48.12.1181.

McDaniel, M. A. (2005). Big-brained people are smarter: A meta-analysis of the relationship between in vivo brain volume and intelligence. *Intelligence, 33*, 337–346. doi: 10.1016/j.intell.2004.11.005.

McDaniel, M. A. (2007). Validity generalization as a test validation approach. In S. M. McPhail (Ed.), *Alternative validation strategies: Developing new and leveraging existing validity evidence* (pp. 159–180). San Francisco, CA: John Wiley & Sons.

McDaniel, M. A., Hartman, N. S., Whetzel, D. L., & Grubb, W. L., III. (2007). Situational judgment tests, response instructions, and validity: A meta-analysis. *Personnel Psychology, 60*, 63–91. doi: 10.1111/j.1744-6570.2007.00065.x.

McDaniel, M. A., Rothstein, H. R., & Whetzel, D. L. (2006). Publication bias: A case study of four test vendors. *Personnel Psychology, 59*, 927–953. doi: 10.1111/j.1744-6570.2006.00059.x.

McDaniel, M. A., Schmidt, F., & Hunter, J. (1988). A meta-analysis of the validity of methods for rating training and experience in personnel selection. *Personnel Psychology, 41*, 283–314. doi: 10.1111/j.1744-6570.1988.tb02386.x.

McDaniel, M. A., Whetzel, D. L., 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. doi: 10.1037/0021-9010.79.4.599.

Molyneaux, J. W. (1953). An evaluation of unassembled examinations. Unpublished master's thesis. The George Washington University, Washington, DC.

Ones, D. L., Viswesvaran, C., & Schmidt, F. L. (1993). Comprehensive metaanalysis of integrity test validities: Findings and implications for personnel selection and theories of job performance. *Journal of Applied Psychology, 78*, 679–703. doi: 10.1037/0021-9010.78.4.679.

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. doi: 10.1037/0021-9010.65.4.373.

Rosenthal, R., & Rubin, D. (1978). Interpersonal expectancy effects: The first 345 studies. *Behavioral and Brain Sciences, 3*, 377–415. doi: 10.1017/S0140525X00075506.

Rothstein, H. R. (1990). Interrater reliability of job performance ratings: Growth to asymptote level with increasing opportunity to observe. *Journal of Applied Psychology, 75*, 322–327. doi: 10.1037/0021-9010.75.3.322.

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. In K. R. Murphy (Ed.), *Validity generalization: A critical review* (pp. 115–154). Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

Rothstein, H. R., McDaniel, M. A., & Borenstein, M. (2002). Meta-analysis: A review of quantitative cumulation methods. In F. Drasgow & N. Schmitt (Eds.), *Measuring and analyzing behavior in organizations* (pp. 401–445). San Francisco: Jossey-Bass.

Rothstein, H. R., Sutton, A. J., & Borenstein, M. (2005). Publication bias in meta-analysis. In H. R. Rothstein, A. J. Sutton, & M. Borenstein (Eds.), *Publication bias in meta-analysis: Prevention, assessment, and adjustments* (pp. 1–7). West Sussex, UK: John Wiley & Sons.

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. doi: 10.1037/0021-9010.62.5.529.

Schmidt, F. L., & Hunter, J. E. (1996). Measurement error in psychological research: Lessons from 26 research scenarios. *Psychological Methods, 1*, 199–223. doi: 10.1037/1082-989X.1.2.199.

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. doi: 10.1037/0033-2909.124.2.262.

Schmidt, F. L., & Hunter, J. E. (2003). History, development, evolution, and impact of validity generalization and meta-analysis methods, 1975–2001. In K. R. Murphy (Ed.), *Validity generalization: A critical review* (pp. 31–65). Mahwah, NJ: Lawrence Erlbaum Associates.

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. doi: 10.1111/j.1744-6570.1985.tb00565.x.

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

Stern, J. M., & Simes, R. J. (1997). Publication bias: Evidence of delayed publication in a cohort study of clinical research projects. *British Medical Journal, 315*, 640–645.

Trikalinos, T. A., & Ioannidis, J. P. A. (2005). Assessing the evolution of effect sizes over time. In H. R. Rothstein, A. J. Sutton, & M. Borenstein (Eds.), *Publication bias in meta analysis: Prevention, assessment and adjustments* (pp. 111–126). Chichester, UK: John Wiley & Sons.

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. doi: 10.1037/0021-9010.81.5.557.

Williams, M. L., McDaniel, M. A., & Nguyen, N. T. (2006). A meta-analysis of the antecedents and consequences of pay level satisfaction. *Journal of Applied Psychology, 91*, 392–413. doi: 10.1037/0021-9010.91.2.392.