Impact of Assessments of Validity Generalization and Situational Specificity on the Science and Practice of Personnel Selection

Kevin R. Murphy*

The application of meta-analysis, in particular validity generalization (VG) analysis, to the cumulative literature on the validity of selection tests has fundamentally changed the science and practice of personnel selection. VG analyses suggest that the validities of standardized tests and other structured assessments are both higher and more consistent across jobs and organizations than was previously believed. As a result, selection researchers and practitioners can draw on the research literature to make reasonably accurate forecasts about the validity and usefulness of different tests in particular applications. Distinctions between tests of validity generalization and tests of situational specificity are described, and difficulties in demonstrating that validity is constant across the different settings where tests are used are outlined.

Introduction

The goal of applied psychology is to bring scientific knowledge about human behavior to bear to help solve important practical problems. Personnel selection represents one of the 'success stories' of applied psychology, in the sense that it presents a significant opportunity to apply empirical research directly to help solve the problems an organization faces whenever it must make choices between multiple applicants for a given job. The marriage between science and practice in applied psychology is not always an easy or productive one (Dunnette 1990; Murphy and Saal 1990), and there are good reasons to believe that current practices in personnel selection often lag far behind the best recommendations that empirical research has to offer. Nevertheless, scientific research on personnel selection (particularly studies of the validity of various tests and assessments used in selecting among applicants) offers clear guidance to the practitioner. The application of meta-analysis to the body of research on test validity has played a very important role in putting personnel selection on a firm scientific footing.

There have been hundreds, if not thousands, of studies examining the validity and utility of tests, interview methods, work samples, systems for scoring biodata, assessment centers, etc., and narrative reviews of these studies have often suggested that validities are relatively small and highly variable across situations (Ghiselli 1966, 1970). From the 1950s to the 1980s, this pattern of findings led to the widespread assumptions that (a) it would be difficult to predict or determine what sorts of tests might or might not be valid as predictors of performance in a particular job; (b) the validity of particular tests varied extensively across settings, organizations, etc., even when the essential nature of the job was held constant; and (c) the only way to determine whether a test was likely to be valid in a particular setting was to do a local validity study. One implication of this set of assumptions was the belief that efforts to develop and validate personnel selection tests or other methods of assessing job candidates would have to be customized and tailored to each job, organization, setting, etc., and that tests that had worked well in many other jobs or organizations could not be safely used without careful study of their validity in that particular setting.

Applications of meta-analysis, and particularly validity generalization analyses, to studies of the validity of tests, interviews, assessment centers, and the like has led to substantial changes in assumptions and beliefs about the validity and usefulness of selection tests. In particular, it is now widely accepted that (a) professionally developed ability tests, structured interviews, work samples, assessment centers and other structured assessment techniques are likely to provide valid predictions of future performance

* Address for correspondence: Kevin R. Murphy, Department of Psychology, Pennsylvania State University, University Park, PA 16802. E-mail krmurphy@psu.edu

across a wide range of jobs, settings, etc.; (b) the level of validity for a particular test can vary as a function of characteristics of the job (e.g., complexity) or the organizations, but validities are often reasonably consistent across settings; and (c) it is possible to identify abilities and broad dimensions of personality that are related to performance in virtually all jobs (for reviews of research supporting these points, see Hartigan and Wigdor 1989; Hunter and Hunter 1984; Reilly and Chao 1982; Schmidt and Hunter 1999; Schmitt, Gooding, Noe and Kirsch 1984; Wigdor and Garner 1982. For illustrative applications of VG methods, see Callender and Osburn 1981; Schmidt, Hunter, Pearlman, and Shane 1979.) Schmidt and Hunter (1999) reviewed 85 years of research on the validity and utility of selection methods and concluded that cognitive ability tests, work samples, measures of conscientiousness and integrity, structured interviews, job knowledge tests, biographical data measures and assessment centers all showed consistent evidence of validity as predictors of job performance.

There have been several factors that have contributed to changing beliefs about the validity and utility of selection tests, but the most important has probably been the widespread application of meta-analysis to validity studies. Meta-analysis has given researchers a set of methods for examining the broad body of evidence about the validity of these tests, and has given them a sophisticated set of procedures for evaluating and understanding the consistency of validity evidence across multiple settings.

Meta-Analysis and Validity Generalization Analyses

There are a number of different methods of meta-analysis. For example, Rosenthal (1984) developed methods of combining the *p* values (i.e. probability that experimental results represent chance alone) from several independent studies to obtain an estimate the likelihood that the particular intervention, treatment, etc. has some effect. Glass, McGaw and Smith (1981) developed methods of combining effect size estimates (e.g. the difference between the experimental and control group means, expressed in standard deviation units) from multiple studies to give an overall picture of how much impact treatments or interventions have on key dependent variables. Hedges and Olkin (1985) developed a mathematically rigorous formulation for combining the results of independent studies. All of these methods share the same key insight, that statistical methods can be used to pool and compare the results of multiple studies, and that when this is done, a good deal more order and regularity in findings is likely to be revealed than when impressionistic or subjective methods are used to review and evaluate the research literature.

Suppose you were conducting a study, and you collected data from 125 subjects. One thing you would almost certainly do would be to compute some simple descriptive statistics (e.g., the mean and the standard deviation) to help you make sense of your findings. Meta-analysis is often little more that the application of this same strategy to the results of multiple studies. That is, if you wanted to make sense of the results of 125 different validation studies, one thing you would probably do would be to compute the mean and the standard deviation of the validities across studies. Many of the current methods of meta-analysis will take a more sophisticated approach to the problem than simply computing the average across all studies (e.g., they might weight for sample size), but the starting point for virtually all methods of meta-analysis is essentially to compute some descriptive statistics that summarize key facets of the research literature you hope to summarize and understand.

The descriptive statistics that might be obtained from a quantitative review of research on the validity of most selection tests would often appear to confirm the wisdom of the traditional view of the validity of selection tests. That is, if you apply many methods of meta-analysis to the accumulated literature on test validity, you might find that validities *are* often low, and *are* often substantially different in different settings, organizations, jobs, etc. However, applications of validity generalization (VG) analysis to this research literature has led to very different conclusions. For example, applications of VG to studies of the validity of cognitive ability tests as predictors of job performance and performance in training have led to the conclusion that validities are both (a) generalizable, in the sense that such tests appear to be at least minimally valid predictors in virtually all settings; and (b) consistent, in the sense that the level of validity is reasonably comparable across settings (Hunter and Hirsh 1987; Hunter and Hunter 1984; Schmidt 1992; Schmidt and Hunter 1977). Similarly, applications of the VG model to quantitative reviews of research on the validity of personality inventories as predictors of performance (e.g., Barrick and Mount 1991; Hough, Eaton, Dunnette, Kamp and McCloy 1990; Tett, Jackson and Rothstein 1991) has overturned long-held assumptions about the relevance of such tests for personnel selection. Personnel researchers now generally accept the conclusion that scores on personality inventories are related

to performance in a wide range of jobs. One way to understand the differences between the conclusions of earlier reviews of validity research and reviews that apply the VG model is to note that the difference between VG analyses and many other methods of meta-analysis is analogous to the difference between descriptive and inferential statistics. That is, many methods of meta-analysis simply describe the distribution of outcomes for different studies in a particular are of research; VG analyses attempt to go beyond simple description to draw inferences about the population values of the statistics obtained via meta-analysis.

Using VG Analysis to Draw Inferences about the Meaning and Implications of Validity Studies

More than 20 years ago, Schmidt and Hunter (1977) developed procedures for what is often referred to as 'psychometric meta-analysis' or 'validity generalization analysis'. There have been a number of developments and elaborations of this basic model (Burke 1984; James, Demaree, Mulaik and Ladd 1992; Raju and Burke 1983; Schmidt, Law, Hunter, Rothstein, Pearlman and McDaniel 1993), as well as developments and elaborations of alternative approaches for attacking similar problems (Hedges 1988; Raudenbush and Bryk 1985; Thomas 1990), and there is an extensive literature debating the validity generalization model (e.g., Hartigan and Wigdor 1989; James, Demaree, and Mulaik 1986; Kemery, Mossholder, and Roth 1987; Thomas 1990). Although there is still considerable discussion and controversy over specific aspects of or conclusions drawn from validity generalization analyses, the core set of ideas in this method are simple and straightforward. This model applies some basic ideas from psychometric theory and from statistical theory to try and draw inferences about what the data in a particular area of research mean (Schmidt 1992).

As noted earlier, if you collect validity coefficients from 100 studies of a particular assessment procedure (e.g., the situational interview), you are likely to find that the average validity coefficient is relatively small and that the validities vary considerably across studies. The VG model suggests that there are a variety of statistical artifacts that artificially depress the mean and inflate the variability of validity coefficients, and further that the effects of these artifacts can be easily estimated and corrected for. It is useful to discuss two broad classes of corrections separately, corrections to the mean and corrections to the variability in the distribution of validity coefficients that would be found in a descriptive meta-analysis.

Corrections to the Mean

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

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

Performance ratings are normally collected in settings where range restriction is ubiquitous, especially when ratings are used to make administrative decisions about ratees (e.g., salary, promotion; see Murphy and Cleveland 1995). For example, Bretz, Milkovich and Read (1992: 333) conclude: 'the norm in U.S. industry is to rate employees at the top end of the scale'. Evidence of leniency and range restriction in performance ratings is so pervasive that several commentators (e.g., Jawahar and Williams 1997; Ilgen, Barnes-Farrell and McKellin 1993) have urged caution in using ratings as criteria in validation studies. Range restriction can also artificially depress validity coefficients

Suppose, for example, that an organization used a performance appraisal form with a 9-point rating scale. If ratings were normally distributed throughout the entire scale, they would have a

mean of about 5 and a standard deviation of about 1.5. It is more likely, however, that ratings will be concentrated at the top end of the scale (Bretz *et al*. 1992). If raters restrict their ratings to the top half of the scale, you might find a mean of about 7 and a standard deviation of about 1. If they concentrate their ratings in the top third of the scale, the mean might be closer to 8 and the standard deviation less than .50. If raters use a restricted range when filling out performance evaluations, the effects on the correlations test scores and ratings can be substantial.

Assume, for example, that the observed correlation between scores on a test and performance ratings is .25. If raters in this organization use only the top half of the scale, the correlation corrected for range restriction will be .32. If raters in this organization use only the top third of the scale, the correlation corrected for range restriction .45. In corrections are made for both range restriction and unreliability, it is likely that the corrected mean correlation will be a good deal larger that the average of the uncorrected *r* values. For example, McDaniel, Whetzel, Schmidt and Maurer (1994) analyzed the validity of the interview as a predictor of job performance; their study included more than 150 validity coefficients. The mean validity was .20, but when corrected for attenuation and range restriction, this estimate nearly doubled (the corrected mean validity was .37).

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

Corrections to the Variance

Meta-analyses of validity coefficients have sometimes shown that the validity for the same type of test or measure varies considerably across jobs, organizations, settings, etc. This variability in validity coefficients is one of the chief reasons for the long-held assumption that it was necessary to validate tests in each setting where they were used. The validity generalization model suggests that some, and perhaps all of the variability in validity coefficients might be explained in terms of a few simple statistical artifacts, and that once the effects of these artifacts are removed, you are likely to conclude that the validity of tests is substantially similar across settings. Many potential explanations for variability in test validity have been put forth (e.g., the reliability of performance measures is higher in some organizations than in others, which can lead to apparent differences in validity), but much of the literature dealing with validity generalization has focused on the simplest and probably the most important explanation for differences in validity coefficients across studies, i.e., simple sampling error. Many validity studies, particularly studies from the early 1970s and earlier, used small samples, and it is well known that statistical results of all sorts, including validity coefficients, are highly unstable when samples are small. Corrections for sampling error and for other artifacts that artificially inflate the variability in test validities often suggest that much of the apparent instability of validities is a reflection of weaknesses of validity studies (small samples, variation in the degree of unreliability and range restriction) rather than a reflection of true differences in the validity of tests across settings. For example, in McDaniel *et al.*'s (1994) analysis of the validity of situational interviews the standard deviation of the validities they reviewed was .14. After applying statistical corrections based on the VG model, this value shrank to .05.

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

Validity Generalization versus Situational Specificity of Validities

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

In the early years of validity generalization research (e.g., late 1970s to mid-1980s), researchers often talked about validity as if it were a dichotomous variable, i.e., tests are either valid or not valid. This way of thinking closely

mirrors the treatment of validity in the legal system, in which tests that led to adverse impact were held to be illegal unless they were shown to be 'valid'. If such a showing was made, it did not matter much whether the test was just above the minimum threshold for defining validity or if it was a highly sensitive predictor of performance. Early research on validity generalization focused largely on the question of whether test validities exceeded some minimum level in most validity studies. Later research has focused more strongly on the consistency of validity across situations, in particular on the hypothesis that the level of validity achieved by a test might be situationally specific.

Distinguishing between Validity Generalization and Situational Specificity

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

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

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

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

Gaugler, Rosenthal, Thornton and Bentson (1987) conducted a meta-analysis of assessment center validities; results from this study can be used to illustrate the procedures used to evaluate validity generalization vs. situation specificity. Their review included 44 correlations (from 29 separate studies) between assessment center ratings and measures of job performance. The mean and the standard deviation of these validity coefficients were .25 and .15, respectively. After correcting for sampling error, unreliability, and other statistical artifacts, Gaugler *et al*. (1987) reported that (a) the best estimate of assessment center validity was given by a corrected mean validity of .36; (b) the corrected validities varied substantially across studies (i.e., a corrected standard deviation of .14); and (c) 90% of the corrected validities were greater than .18. This set of results led them to conclude that the assessment center method was at least minimally valid in virtually all reported applications (i.e. assessment center validity generalized), but that the level of validity was *not* consistent across studies, suggesting that characteristics of the jobs, organizations, assessment exercises, etc. could substantially affect the validity of assessment center ratings.

In principle, there is no necessary relationship between tests of situational specificity and tests of validity generalization. The most common finding, at least in the area of ability testing, has been that validities are both (a) generalizable, in

the sense that such tests appear to be at least minimally valid predictors in virtually all settings; and (b) consistent, in the sense that the level of validity is reasonably comparable across settings (Hunter and Hirsh 1987; Hunter and Hunter 1984; Schmidt 1992). However, it is also possible to conclude that validities are generalizable, but not consistent. That is, tests might show *some* validity in virtually all settings, but might be substantially more useful in some jobs, organizations, etc. than in others.

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

Difficulties in Assessing Situational Specificity

For most of the history of personnel selection research, it was assumed that tests, interview protocols, or other assessment procedures that were valid predictors of performance in one setting might be useless in other, apparently similar settings. Although there have been few attempts to develop a compelling theory *why* tests might be highly valid in some settings and not in others (James *et al*. 1992), the data seemed compelling, and for many years it was believed that no matter how well the test worked in other settings, a new validity study would be needed whenever the test was introduced into a new situation.

Validity generalization research suggests that much of the apparent variability in levels of test validity is probably due to sampling error and other statistical artifacts, and that the evidence of true situational specificity (i.e. the hypothesis that a test is in fact more valid in some settings than in others) is weak. Schmidt, Hunter and their colleagues have gone further, claiming that the validity of some tests (specifically, cognitive ability tests) is essentially constant, at least within some very broad groupings of jobs (Hunter and Hirsch 1987; Schmidt *et al.* 1993). For example, Schmidt, Hunter and Raju claimed that validity generalization 'studies have concluded that there is no situational specificity for cognitive ability tests' (1988: 666). The only potential moderator of the validity of cognitive ability tests that has been accepted by some VG researchers is job complexity; ability tests seem to show higher validities in more cognitively demanding jobs (Gutenberg, Arvey, Osburn and Jenneret 1983; Hunter and Hunter 1984; Murphy 1989). With this potential exception, Schmidt, Hunter and their colleagues have made a strong claim for the invariance of cognitive test validities over situations.

The claim that there is no situational specificity is controversial (recent papers dealing with situation specificity include Oswald and Johnson 1998; Erez, Bloom and Wells 1996); questions regarding this claim can be grouped under three main headings (a) the power of VG procedures to detect validity differences; (b) the role of sample size in tests of situational specificity; and (c) the lack of useful models to explain how situational variables might or might not moderate validity.

Power of Tests of Situational Specificity

A frequent criticism of validity generalization procedures is that they are biased in favor of the hypothesis that validity is consistent. In part, this criticism has to do with the statistical minutiae of various VG formulas (see James *et al.* 1992 for a statistical critique of VG formulas). The broader issue, however, relates to the statistical power of VG procedures to detect true differences in test validity across jobs, organizations, or situations. In this context, power refers to the probability that VG procedures will detect true and meaningful differences in validity. Research on the statistical power of VG procedures suggests that this probability can be disappointingly low.

The statistical power of VG procedures is affected by a number of variables; the two that have received the most attention are the number of studies included in the meta-analysis (k) and the number of subjects included in each study (N). Many validity generalization studies, particular in the area of cognitive ability testing, include large numbers of studies (i.e. large k), most of which are based on small samples (i.e. small N). Osburn, Callender, Greener and Ashworth (1983) found that the power of VG procedures was unacceptably low for detecting small to moderate differences in true validity when the average N was less than 100. Sackett, Harris and Orr (1986) found that when when there were small differences in true validity (e.g. differences of .10 in actual test validity over situations), power was low regardless of N or k. They also reported that power was unacceptably low for detecting larger differences in true validity (e.g., .20) when N or k was small. These studies suggest that when the actual level of test validity varies by as much as .10 to .20 over situations, VG procedures may nevertheless lead

one to conclude that validity is constant across situations.

The inability of some VG procedures to reliably detect validity differences in the .10-.20 range is not necessarily a serious problem; in many contexts, one might argue that differences that small, even if detected, would not constitute meaningful variation in validity. However, a study by Kemery, Mossholder and Roth (1987) suggests that VG procedures can show low power even in situations where validity differences are large. Their study considered the situation in which a test has essentially no validity in most applications, but has a high level of validity (e.g,. .60) in some. They found low levels of power for detecting validity differences in situations where as many as 10–30% of the true validities were .60 (and the rest were .00).

The Role of Sample Size in Tests of Situational Specificity

Schmidt (1992) suggests that approximately 80-90% of the variability in ability test validities is due to statistical artifacts. While this figure seems impressive, it does not by itself disclose much about situational specificity. A close analysis of VG studies shows that the percentage of variance accounted for by statistical artifacts is strongly affected by the sample sizes of validity studies, and that figures such as 80% of the variance accounted for are found only when VG methods are applied to small-sample studies (McDaniel, Hirsh, Schmidt, Raju and Hunter 1986; Murphy 1993; Schmitt *et. al*. 1984).

A comparison of meta-analyses of small-sample studies versus large-sample studies leads to two important conclusions. First, because of sampling error, the *amount* of variability in test validities depends substantially on *N*. Validities obtained from large-sample studies tend to be relatively consistent across situations, whereas (because of sampling error) small-sample validities vary substantially across situations. Extensive and unsystematic variability in test validities across situations seems to be restricted to small-sample validity studies. Second, the *percentage* of variance due to statistical artifacts such as sampling error varies inversely with *N*. In meta-analyses based primarily on small samples, sampling error alone often accounts for 70-80% of the variance in test validities. In VG analyses that include larger samples, the percentage of variance accounted for might be as small as 15-30% (McDaniel *et al*. 1986; Schmitt *et al.* 1984).

It seems clear that decision rules defined in terms of the percentage of variance accounted for are deficient. If *N* is small (e.g., less than 100), the percentage of variance accounted for by statistical artifacts will tend to be large. If *N* is large (e.g., greater than 200), the percentage of variance accounted for by artifacts will be small. If *N* is large enough, this percentage will tend to zero, and can even appear to become negative (i.e. when the corrected variance in validities is larger than the observed variance; see Murphy 1993). In the long run, the percentage of variance accounted for by statistical artifacts (particularly sampling error) might turn out to be little more than a roundabout estimate of the number of subjects included in the studies analyzed.

Lack of a Situational Model

The belief that test validities varied across situations was long part of the common wisdom of personnel researchers; the belief that validities do *not* vary across situations seems to now be part of that common wisdom. Given all the attention that has been devoted to situational specificity, both by its advocates and its critics, there have been remarkably few attempts to articulate just how situational variables might affect validity, which situational variables might be important, or even what exactly constitutes a situation (James *et al*. 1986; James *et al.* 1992). Gutenberg *et al*. (1983) suggested that validities would be higher for more jobs with more complex information-processing demands than for simpler jobs. Murphy (1989) suggested that validities might be higher when organizations were in a turbulent environment, where new tasks, technologies, and responsibilities were constantly being added, than when the environment was stable and the process of production was routine. James *et al*. (1992) articulated how the restrictiveness of an organization's climate might affect the validities of tests.

Restrictive climates are characterized by highly structured work environments, an emphasis on standardization, formalization, and control, end extensive centralization of authority. Non-restrictive climates are characterized by decentralization, autonomy, innovation, and a relative lack of structure. James *et al*. (1992) predicted that tests will be more useful predictors of performance in non-restrictive than in restrictive environments. Similar predictions have been made by other researchers (e.g. Kemery *et al*. 1987; Murphy 1989).

James *et al*. (1992) note that the effects of variation in organizational climate, situational constraints on performance, etc. on the validity of tests will be obscured by traditional VG procedures. In particular, these situational characteristics may lead to substantive effects that are dismissed as statistical artifacts in VG procedures, and are 'correct for'. For example, they note that an important goal of restrictive

climates is to reduce differences in employees' output, which means that there should be less true variability in performance in restrictive than in nonrestrictive climates. This range restriction is not a statistical artifact, and any 'correction' for the fact that some climates restrict variance in output whereas others enhance this variance would be misleading and inappropriate.

Similarly, James *et al*. (1992) note that the reliability of the criterion (here, the performance measure employed) should be substantively affected by the restrictiveness of the climate. If there are small true differences in performance in restrictive climates, it follows that the reliability of performance measures will also tend to be small. In contrast, performance measures should be relatively reliable in non-restrictive climates, where there are meaningful differences in performance to measure. The net result is that there should be variability in criterion reliabilities as a function of the restrictiveness of the situation. Once again, VG procedures that correct for this variability will produce misleading results. This variability is not a statistical artifact unrelated to 'true validity', but rather is a meaningful source of validity differences across climates that vary in their restrictiveness.

Suppose that you used the same test to predict performance in four large organizations, two of which were characterized by highly restrictive climates and two by climates that put no restrictions on individual variation, innovation, etc. In the first two organizations, you would expect little variability in performance, which implies low reliability and low validity (both because of low reliability and range restriction). In the second two organizations, you would expect larger variation in performance, more reliable performance measures, and larger validities. Traditional VG procedures would lead you to regard the differences in validity across situations as the result of a set of statistical artifacts (i.e., differences in range restriction and unreliability across organizations), and would lead you to correct for those artifacts. In this case, statistical corrections would lead to the misleading conclusion that the 'true validity' of the test was the same in all four organizations. This conclusion is correct only in the trivial sense, where 'true validity' is defined in terms that have no relationship to the contexts in which the tests are actually used. In this example, the tests really are more valid and useful in some contexts than in others, and analytic procedures that lead to the conclusion that there are no differences in test validity are simply misleading.

James *et al.* (1992) called for systematic research on potential moderators of validity, with particular attention to identifying the *processes* by which validities might be altered as situational variables change. During the long period when situational specificity was assumed, there was virtually no serious attempt to explain *why* validities might vary. James *et al*. (1992) present a clear example of how such a theory might be developed, and how this theory might provide an alternative explanation of substantive phenomena that are dismissed in VG analyses as statistical artifacts.

VG as an Inferential Method: Implications for Tests of Situational Specificity

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

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

First, it is important to ask whether a VG analysis provides credible evidence about situational specificity. Analyses that are based on relatively weak studies (e.g., studied with small N and unreliable criteria) may not allow you to convincingly sort out variance due to

statistical artifacts from variance due to meaningful changes in validity across jobs, organizations, or settings. For example, many early validity generalization analyses featured average sample sizes of approximately 60-75 (see Table 1 in McDaniel *et al.* 1986), whereas more recent studies often have sample sizes ranging from approximately 600–750, depending on the criterion (Schmitt *et al.* 1984). Meaningful inferences about situational specificity depend first and foremost on the quality of the database that supports those inferences, and even studies that include a very large number of studies (as has been the case in some meta-analyses of the ability-performance relationship) may not provide a firm basis for making inferences about situational specificity if most of the underlying studies are poorly designed.

Second, it is important to ask whether the sample of studies included in a meta-analysis spans the population of potential applications of the test. VG analyses that are based on a small number of validities, or that are based on studies taken from a very restricted range of potential applications may not provide a useful basis for making inferences about the consistency of test validity. For example, McDaniel *et al.* (1994) drew inferences about the validity of situational interviews on the basis of 16 validity coefficients, and about psychological interviews on the basis of 14 coefficients. They were appropriately cautious in interpreting these findings, and potential consumers of meta-analysis must also be cautious about over-interpreting consistency in a small set of validity studies. They must be even more cautious about drawing broad inferences when the sample of validity studies spans only a small part of the range of situations in which a test might be used. For example, validity studies are more likely to be done in lower-level jobs (e.g., clerical jobs, semi-skilled labor) than in managerial or professional jobs. When drawing inferences from a VG analysis, it is important to have detailed information about the range of jobs, situations, etc. represented by the set of validity studies examined. Unfortunately, this sort of information is virtually never presented in the publications describing meta-analyses or VG analyses, and it is often necessary to go back to the original validity studies to determine what sorts of populations have actually been studied.

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

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

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

Implications of VG for Selection Research and Practice

Schmidt notes that: 'Meta-analysis has been applied to over 500 research literatures in employment selection, each one representing a predictor-job performance pair' (1992: 1177). The most frequent application of these methods has been in research on the relationship between scores on cognitive ability tests and measures of overall job performance; representative examples of this type of validity generalization analysis include Pearlman, Schmidt and Hunter (1980), Schmidt, Gast-Rosenberg and Hunter (1980) and Schmidt, Hunter and Caplan (1981). However, applications of meta-analysis and validity generalization analysis have not been restricted to traditional test validity research. Hunter and Hirsh (1987) reviewed meta-analyses spanning a wide range of areas in applied psychology (e.g., absenteeism, job satisfaction). These methods have been applied to assessments of the relationship between personality traits and job performance (Barrick and Mount 1991), assessments of race effects in performance ratings (Kraiger and Ford 1985) and assessments of the validity of assessment center ratings (Gaugler *et al*. 1987). Finally, a number of authors (e.g., Hom, Carnikas-Walker, Prussia and Griffeth 1992) have combined meta-analysis with structural modeling to assess the appropriateness of competing theories or models, based on the cumulative literature in particular areas of research.

In virtually every research literature in which these methods are applied, similar conclusions have been reached, i.e., that validity coefficients are usually larger than and more consistent than a casual review of the literature (or even a careful meta-analysis that does not apply the corrections that are at the core of VG analyses) would suggest. Indeed, this conclusion is virtually foregone, because so many validity studies are so badly done. Many validity studies combine small samples, unreliable measures, and substantial levels of range restriction, making the results of these studies very difficult to interpret. Meta-analytic methods help substantially, because they focus the researcher's attention on the general trends in a research literature rather than the results of a specific study, but meta-analysis by itself does not directly solve the problem of how to interpret the results of studies that share the weaknesses noted above. The application of statistical and psychometric corrections to validity studies helps substantially in sorting out the effects of statistical artifacts on the mean and the variability of the distribution of validities. Corrected estimates of validity are not always substantially higher or more consistent than uncorrected estimates – Murphy (1993) discusses conditions under which VG corrections can lead to lower rather than higher levels of consistency in validity estimates – but they usually are.

The application of meta-analysis and validity generalization analysis to the research literature on selection validity has accomplished three things. First, it has helped researchers to identify general trends in the research literature. Prior to the development of meta-analytic methods, literature reviews often yielded little more than subjective and incomplete summaries of validity, the effects of interventions, etc., and they often failed to detect important trends in the literature. The application of meta-analytic methods has helped researchers identify these trends, and in particular has helped to identify consistent links between basic abilities and personality characteristics and job performance. For example, it seem likely that intelligence or cognitive ability is relevant for predicting performance in virtually every job studied (Hunter and Hunter 1984, Nathan and Alexander 1988; McHenry, Hough, Toquam, Hanson and Ashworth 1990; Ree and Earles 1994; Schmidt, Hunter, and Outerbridge 1986). The more demanding or complex the job, the stronger the relationship between cognitive ability and successful job performance. Similarly, analyses of the cumulative research literature on personality and job performance suggests there are broad personality traits that appear important in a wide range of jobs. For example, Barrick and Mount (1991) suggested that measures of conscientiousness are a valid predictor of performance across many jobs. Other analyses (e.g., Tett *et al*. 1991) have suggested that other similarly broad personality attributes might also show generalizable validity, and have confirmed the finding that individual differences in conscientiousness appear to be consistently related to job performance. Prior to the application of meta-analytic techniques, there was no clear consensus about attributes or traits that might be good predictors of performance, almost regardless of the job. Now there is substantial consensus, at least about these specific attributes.

Second, corrections to the mean of the distribution of validities obtained when you perform a meta-analysis of this literature have often led to the conclusion that tests and other assessment methods work much better (i.e., show higher average levels of validity) than was previously believed. As Schmidt and Hunter (1999) note, psychological tests and other structured assessments are extremely useful in helping organizations identify the job candidates who are most likely to succeed. Structured assessment procedures often combine relatively high levels of reliability and validity with relatively low costs. As a result, selection tests are often extremely cost effective. This is particularly true for standardized tests, but even

expensive and labor-intensive methods of assessment (e.g. work samples, assessment centers) are usually very cost-effective.

Third, corrections to the variance of the distribution of validities have often led to the conclusion that validities are much more consistent than they at first appear. Proponents of VG note that much of the apparent variability in validity coefficients is probably due to factors such as the use of small samples in validity studies. Once you take statistical artifacts into account, you are likely to conclude that the operational validity of a test is reasonably consistent across settings. As a result, tests or assessments that work well in other settings or organizations are likely to also work well in your setting. One implication is that hiring organizations do not have to 'reinvent the wheel' every time a new selection system is developed. Rather, they can often use the accumulated literature on the validity of selection tests to make reasonably accurate forecasts of how well particular tests or methods of assessment will work for them.

The greatest single contribution of validity generalization analyses to personnel selection is the demonstration that the results of empirical research can indeed be applied, with considerable success, to help solve important and practical problems. When faced with the problem of selecting among applicants for the job of computer programmer, for example, practitioners can draw from a large body of scientific research to determine which attributes are likely to be related to success (e.g., cognitive ability and conscientiousness are very likely to be relevant, and an analysis of the specific job performed by programmers is likely to suggest other relevant attributes), to predict about how well tests of different sorts will work, and to predict the concrete consequences (e.g. in terms of productivity or dollar value) of applying different methods and models of selection. Validity generalization analyses suggest that most professionally developed selection tests and assessment methods work, that they work better than most people would think, and that work more consistently than most people thing. This set of conclusions suggests that there is a solid scientific basis for the use of standardized tests and other structured assessment methods in personnel selection, and that it is possible to base practical decisions about how to hire people on a firm empirical foundation.

References

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

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

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

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

Cronbach, L.J., Gleser, G.C., Nanda, H. and Rajaratnam, N. (1972) *The Dependability of Behavioral Measurements: Theory of Generalizability for Scores and Profiles.* New York: Wiley.

Dunnette, M.D. (1990) Blending the science and practice of industrial and organizational psychology: Where are we and where are we going? In M. Dunnette and L. Hough (eds.), *Handbook of Iindustrial and Organizational Psychology.* Second edn. (Vol. 1, pp. 1–27). Palo Alto, CA, Consulting Psychologists Press.

Erez, A., Bloom, M.C. and Wells, M.T. (1996) Using random rather than fixed effects models in meta-analysis: Implications for situational specificity and validity generalization. *Personnel Psychology*, **49**(2), 275–306.

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

Ghiselli, E. E. (1966). *The Validity of Occupational Aptitude Tests*. New York: Wiley.

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

Glass, G.V., McGaw, B. and Smith, M.L. (1981) *Meta-analysis in Social Research*. Beverly Hills, CA, Sage.

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

Hartigan, J.A. and Wigdor, A.K. (1989) *Fairness in Employment Testing: Validity Generalization, Minority Issues, and the General Aptitude Test Battery*. Washington, DC, National Academy Press.

Hedges, L.V. (1988) Meta-analysis of test validities. In H. Wainer and H. Braun (eds.) *Test Validity.* Hillsdale, NJ, Erlbaum, 191–212.

Hedges, L.V. and Olkin, I. (1985) *Statistical Methods for Meta-analysis.* New York, Academic Press.

Hom, P.W., Carnikas-Walker, F., Prussia, G.E. and Griffeth, R.W. (1992) A meta-analytical structural equations analysis of a model of employee turnover. *Journal of Applied Psychology*, **77**, 890–909.

Hough, L.M., Eaton, N. K., Dunnette, M.D., Kamp, J.D. and McCloy, R.A. (1990) Criterion-related validities of personality constructs and the effect of response distortion on those validities. *Journal of Applied Psychology*, **75**, 581–595.

Hunter, J.E. and Hirsh, H.R. (1987). Applications of meta-analysis. In C.L. Cooper and I.T. Robertson

(eds.), *International Review of Industrial and Organizational Psychology*. Chichester, Wiley, 321–357.
Hunter, J.E. and Hunter, R.F. (1984). Validity and utility of alternative predictors of job performance. *Psychological Bulletin*, **96,** 72–98.
Hunter, J.E. and Schmidt, F.L. (1990*). Methods of Meta-analysis: Correcting Error and Bias in Research Findings*. Newbury Park, CA, Sage.
Ilgen, D.R., Barnes-Farrell, J.L. and McKellin, D.B. (1993) Performance appraisal process research in the 1980s: What has it contributed to appraisals in use? *Organizational Behavior and Human Decision Processes*, **54,** 321–368.
James, L.R. Demaree, R.G. Mulaik, S.A. (1986) A note on validity generalization procedures. *Journal of Applied Psychology*, **71**, 440–450.
James, L.R., Demaree, R.G., Mulaik, S.A. and Ladd, R.T. (1992) Validity generalization in the context of situational models. *Journal of Applied Psychology*, **77**, 3–14.
Jawahar, I.M. and Williams, C.R. (1997) Where all the children are above average: The performance appraisal purpose effect. *Personnel Psychology*, **50**, 905–926.
Kemery, E.R., Mossholder, K.W. and Roth.L. (1987) The power of the Schmidt and Hunter additive model of validity generalization. *Journal of Applied Psychology*, **72**, 30–37.
Kraiger, K. and Ford, J.K. (1985) A meta-analysis of race effects in performance ratings. *Journal of Applied Psychology*, **70**, 56–65.
Lumsden, J. (1976) Test theory. *Annual Review of Psychology*, **27**, 251–280.
McHenry, J.J., Hough, L.M., Toquam, J.L., Hanson, M.A. and Ashworth, S. (1990) Project A validity results: The relationship between predictor and criterion domains. *Personnel Psychology*, **43**, 335–355.
McDaniel, M.A., Hirsh, H.R., Schmidt, F.L., Raju, N. and Hunter, J.E. (1986) Interpreting the results of meta-analytic research: A comment on Schmitt, Gooding, Noe, and Kirsch (1984). *Personnel Psychology*, **39**, 141–148.
McDaniel, M.A., Whetzel, D.L., Schmidt, F.L. and Maurer, S.D. (1994) The validity of employment interviews: A comprehensive review and meta-analysis. *Journal of Applied Psychology*, **79**, 599–616.
Murphy, K.R. (1989) Is the relationship between cognitive ability and job performance stable over time? *Human Performance*, **2**, 183–200.
Murphy, K.R. (1993).The situational specificity of validities: Correcting for statistical artifacts does not always reduce the trans-situational variability of correlation coefficients. *International Journal of Selection and Assessment*, **1**, 158–162.
Murphy, K. (1994) Advances in meta-analysis and validity generalization. In N. Anderson and P. Herriot (eds.), *Internationa Handbook of Selection and Appraisal*. Second edition. Chichester, Wiley, 322–342.
Murphy, K.R. and Cleveland, J.N. (1995) *Understanding Performance Appraisal: Social, Organizational and Goal-Based Perspectives*. Thousand Oaks, CA, Sage.
Murphy, K.R. and Davidshofer, C.O. (1998) *Psychological testing: Principles and Applications* Fourth edition. Englewood Cliffs, NJ, Prentice-Hall.
Murphy, K.R. and Saal, F.E. (1990) *Psychology in Organizations: Integrating Science and Practice.* Hillsdale, NJ, Erlbaum.
Nathan, B.R. and Alexander, R.A. (1988) A comparison of criteria for test validation: a meta-analytic investigation. *Personnel Psychology*, **41**, 517–535.
Osburn, H.G., Callender, J.C., Greener, J.M. and Ashworth, S. (1983) Statistical power of tests of the situational specificity hypothesis in validity generalization studies: A cautionary note. *Journal of Applied Psychology*, **68**, 115–122.
Oswald, F.L. and Johnson, J.W. (1998) On the robustness, bias, and stability of statistics from meta-analysis of correlation coefficients: Some initial Monte Carlo findings. *Journal of Applied Psychology*, **83**, 164–178.
Pearlman, K., Schmidt, F.L. and Hunter, J.E. (1980) Validity generalization results for tests used to predict job proficiency and training success in clerical occupations. *Journal of Applied Psychology*, **65**, 373–406.
Raju, N.S. and Burke, M.J. (1983) Two new approaches for studying validity generalization. *Journal of Applied Psychology*, **68**, 382–395.
Raudenbush, S.W. and Bryk, A.S. (1985) Empirical Bayes metaanalysis. *Journal of Educational Statistics,* **10**, 75–98.
Ree, M.J. and Earles, J.A. (1994) The ubiquitous productiveness of g. In M.G. Rumsey, C.B. Walker and J.H. Harris (eds.), *Personnel Selection and Classification.* Hillsdale, NJ, Erlbaum, 127–136.
Reilly, R.R. and Chao, G.T. (1982) Validity and fairness of some alternate employee selection procedures. *Personnel Psychology*, **35**, 1–67.
Rosenthal, R. (1984) *Meta-Analysis Procedures for Social Research*. Beverly Hills, CA: Sage.
Sackett, P.R. Harris, M.M. and Orr, J.M. (1986) On seeking moderator variables in the meta-analysis of correlational data: A Monte Carlo investigation of statistical power and resistance to Type I error. *Journal of Applied Psychology*, **71**, 302–310.
Schmidt, F.L. (1992) What do data really mean? Research findings, meta-analysis, and cumulative knowledge in psychology. *American Psychologist*, **47**, 1173–1181.
Schmidt, F.L., Gast-Rosenberg, I. and Hunter, J.E. (1980) Validity generalization results for computer programmers. *Journal of Applied Psychology*, **65**, 643–661.
Schmidt, F.L. and Hunter, J.E. (1977) Development of a general solution to the problem of validity generalization. *Journal of Applied Psychology*, **62,** 643–661.
Schmidt, F.L.and Hunter, J.E. (1996) Measurement error in psychological research: Lessons from 26 research scenarios. *Psychological Methods*, **1**, 199–223.
Schmidt, F.L. and Hunter, J.E.(1999) The validity and utility of selection methods in personnel psychology: Practical and theoretical implications of 85 years of research findings. *Psychological Bulletin*, **124**, 262–274.
Schmidt, F.L., Hunter, J.E. and Caplan, J.R. (1981)

Validity generalization results for two groups in the petroleum industry. *Journal of Applied Psychology*, **66**, 261–273.

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

Schmidt, F.L., Hunter, J.E. and Raju, N.S. (1988) Validity generalization and situational specificity: A second look at the 75% rule and Fisher's *z* transformation. *Journal of Applied Psychology*, **73**, 665–672.

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

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

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

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

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

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

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

Wigdor, A.K. and Garner, W.R. (1982) *Ability Testing: Uses, Consequences, and Controversies.* Washington, DC, National Academy Press.