

Accepted for Publication
in
Public Personnel Management

Should Employers Rely on Local Validation Studies or Validity Generalization (VG) to
Support the Use of Employment Tests in Title VII Situations?

Abstract

Since the landmark U.S. Supreme Court case Griggs v. Duke Power (1971) employers have been subject to challenge by plaintiffs or government enforcement agencies when they use employment tests that have adverse impact. In such situations the 1991 Civil Rights Act requires employers to demonstrate that the test is “. . . job related for the position in question and consistent with business necessity.” Employers typically rally such a defense by addressing the Uniform Guidelines on Employee Selection Procedures, the federal treatise framed in 1978 for the express purpose of Title VII enforcement, along with relevant professional standards (Joint Standards, 1999; SIOP Principles, 2003). To address such requirements, employers are faced with several viable validation alternatives, ranging from conducting a local criterion-related validation study to relying on validity evidence from other studies for similar positions, employers, and tests (i.e., Validity Generalization). The strengths and limitations of both of these strategies in a Title VII context are reviewed, and employers are ultimately encouraged to select the local validation strategy whenever technically feasible for a variety of reasons discussed.

Introduction

There are two main reasons why employers conduct validation studies on pre-employment tests. The first reason is to assess whether the test is an effective tool for the desired situation (i.e., for the target job or group of jobs, how the test should be used in conjunction with other tests, what cutoffs may be appropriate, etc.). The second reason is for defensibility—for both low-stakes (challenges brought by individual applicants) and high-stakes (challenges under Title VII and government audits) situations.

Choices for validation strategies include both local techniques that strive to evaluate and/or evidence connections between the test and the target position and more global strategies that evaluate how tests seem to generalize across various settings and positions (as with Validity Generalization, or “VG”). Both federal (the Uniform Guidelines¹) and professional (the Joint Principles² and SIOP Principles³) standards permit techniques that investigate the local validity of the test as well as the more global techniques (e.g., either through transportability studies that attempt to connect existing validity evidence with the local situation).

With such choices available to practitioners, which techniques are likely to produce the most accurate and defensible results? The local validity techniques or the more global ones? What have the courts had to say about either technique in Title VII situations where an employer is being required to demonstrate validity to justify their testing practices? Answers to these questions and others will be provided in this review. While a number of test validation techniques are available to practitioners (e.g., content

validity, construct validity, etc.), this discussion will be limited to only two: local criterion-related validity and VG.⁴

Overview of Local Criterion-related Validity

A local criterion-related validity study is conducted by statistically correlating test scores with some measure of job performance (typically supervisor ratings or performance evaluation scores). Following the conventional practices for the social sciences, validity can be claimed if the correlation between test scores and some job performance metric (i.e., the *criterion*) has a corresponding probability value that is less than .05, which indicates that the correlation is a “beyond chance” occurrence. This type of validity study is typically conducted for tests that measure abstract traits (e.g., some types of cognitive ability, personality, etc.) that may not have obvious connections to the job (as contrasted with content validity, which seeks to demonstrate a more rational-type of connection between the test and the job with traits that are more concrete in nature).

The steps necessary to conduct this type of validation study are very straightforward. Under a predictive model, the researcher administers the test to the applicants and then correlates test scores with some subsequent measure of job performance. Under a concurrent model, the test is given to current job incumbents and simultaneously correlated with job performance metrics of some type. Under either model, having high reliability for both the test and job performance metrics is key for making sure that the results will be accurate and reliable.

Having an adequate sample size to maximize *statistical power* is also important when conducting a local study. Statistical power refers to the ability of the study to find a statistically significant finding if it exists in the target population. Validity studies that have large sample sizes (e.g., 300+ subjects) have high statistical power, and those with small samples have low statistical power. For example, assume that a researcher wanted to find out if a certain test had a validity coefficient of .25 or higher, and there were only 80 incumbents in the target position for whom test and job performance data was available. In this situation, they could be about 73% confident (i.e., have 73% power) of finding such a coefficient (if it existed to be found). With twice the sample size (160 subjects), power is increased to about 94%, which provides the researcher an almost certain ability to find out whether the test was valid for the target position.

Overview of Validity Generalization

VG studies rely on a research technique called meta-analysis. Meta-analysis seeks to combine the results of several similar research studies to form general theories about relationships between similar variables across different situations. As early as 1977, Schmidt & Hunter⁵ applied meta-analyses techniques to the field of personnel testing and framed it as VG. Prior to this time, meta-analyses in the personnel testing and psychological literature was very rare,⁶ but it has since grown to widespread use in the academic field.

The purpose for conducting VG studies in the personnel field is to evaluate the effectiveness (i.e., validity) of a particular type of personnel test (e.g., personality,

integrity, conscientiousness) and to describe what the findings mean in a broader sense.⁷

Practically speaking, VG studies are conducted by compiling several related local criterion-related validity studies into an aggregate analysis to determine the overall effectiveness of the test(s) included in the study for the jobs and settings involved. VG studies also make use of various statistical corrections (e.g., sampling error, range restriction, and criterion unreliability) that allow the researcher to forecast what the overall operational validity of the test(s) may in fact be if they were not hampered by these suppressors.

Some researchers that conduct VG studies apply the “75 Percent Rule” to determine whether validity can be generalized outside of the VG study to other situations. The 75 Percent Rule evaluates whether at least 75 percent of variance in the observed validities (in the VG study) can be accounted for by the correctable statistical artifacts (i.e., sampling error, criterion unreliability, predictor unreliability, and range restriction on the predictor), then the variance between validities is assumed to be zero because the uncorrected artifacts would likely account for the remaining 25 percent of variance. VG studies where at least 75 percent of the variance is explained by these correctable artifacts are said to generalize to other settings outside those included in the study.

Another more contemporary tool used in VG research is the *credibility interval* which is used by some researchers to determine the extent to which validity can be generalized outside the VG study. The credibility interval is an estimate of the variability of individual correlations across studies and informs the researcher the percentage of correlations in the study that are “not likely to be zero.” For example, an 80% credibility interval indicates that 90% of the individual correlations in the VG study *excluded zero*.⁸

One of the major limitations of “corrected” VG studies (as will be discussed more in depth below) is that there is no guarantee that employers would find the level of validity promised by the result of a VG study if a study was performed in a new local setting. This is primarily because a host of situational factors exist in each and every new situation that may drastically impact the validity of a test. In addition, there are a number of limitations with typical VG studies that may further limit their relevance and reliability when evaluating test validity in new situations (see discussion below). However, VG studies offer useful insights into the strength of the relationship between the test and job performance in the studies included in the VG analysis and can be immensely useful in personnel research studies.

Federal and Professional Requirements Surrounding Validity Generalization

Because there is a high degree of overlap and agreement between the Uniform Guidelines and the professional standards regarding the basics involved in conducting and interpreting local criterion-related validity studies, they will not be reviewed here. Only the federal and professional standards relevant to VG are covered because the more recent version of the SIOP Principles (2003) provided additional content surrounding this topic than previous standards included.

Validity Generalization and the Uniform Guidelines

The Uniform Guidelines include two primary sections that describe the requirements for transporting validity evidence from either a VG study or a single validity study conducted elsewhere. Section 7B describes the requirements for *transporting* validity evidence from one (or more) studies to a new local situation, and requires that a job comparability study is conducted between both locations and that the original study includes a fairness study. Sections 7C and 7D direct specific attention to “variables that are likely to affect validity significantly” (called “moderators” in the context of VG studies) and if such variables exist, the user may not rely on the studies, but will be expected instead to conduct an internal validity study in their local situation.

Section 15E of the Uniform Guidelines provides additional guidance regarding transporting validity evidence from existing studies into new situations. Section 15E1(b) includes elements that pertain to the utility and effectiveness of the test and the mitigation of risk that is gained by using a test supported by local validity evidence. Section 15E1(c) cautions the researcher to ensure that extraneous variables are not operating in a way that negatively impacts test validity. Finally, Section 15E1(d) suggests evaluating how the test is used (e.g., ranked, banded, or used with a cutoff).

Validity Generalization and the Professional Standards

The Joint Standards (1999) include a one-page preamble (p. 15) and two standards (with interpretive comments) surrounding VG. While the complex issue of VG is given only a 2-page treatment in the entire manual, the discussion is relevant and to the point. These two standards (1.20 and 1.21) advise test users regarding the conditions

under which validity evidence can be inferred into a new situation based on evidence from other studies.

Standard 1.20. When a meta-analysis is used as evidence of the strength of a test-criterion relationship, the test and the criterion variables in the local situation should be comparable with those in the studies summarized. If relevant research includes credible evidence that any other features of the testing application may influence the strength of the test-criterion relationship, the correspondence between those features in the local situation and in the meta-analysis should be reported. Any significant disparities that might limit the applicability of the meta-analytic findings to the local situation should be noted explicitly.

Standard 1.21. Any meta-analytic evidence used to support an intended test use should be clearly described, including methodological choices in identifying and coding studies, correcting for artifacts, and examining potential moderator variables.

Assumptions made in correcting for artifacts such as criterion unreliability and range restriction should be presented, and the consequences of these assumptions made clear.

Validity Generalization and the SIOP Principles

The updated (2003) SIOP Principles provide a more extensive discussion on VG than the Joint Standards. The entire discussion relevant to VG is contained within pages 8-10 and 27-30 which outline three strategies that are neither mutually exclusive nor exhaustive: (a) transportability, (b) synthetic validity/job component validity, and (c)

meta-analytic validity generalization. The primary elements of this discussion are highlighted below:

- Applying professional judgment in interpreting and applying the results of meta-analytic research is critical.
- The statistical methods employed, including the underlying assumptions used and the statistical artifacts that may influence the results should be evaluated.
- Moderators (situational factors that affect validity findings across settings) should be researched.
- Study characteristics that may possibly impact the study results should be evaluated.
- The similarity of the constructs measured by the tests included in the meta-analysis and those in the local situation should be assessed.
- The similarity between the tests within the meta-analysis, or the situation into which validity will be transported, when the tests differ based upon the development process, content, or ways in which they are scored should be assessed.

Most of the requirements relevant to VG that are presented by the Uniform Guidelines, Joint Standards, and SIOP Principles can be grouped into two areas: evaluating the *internal quality of the VG study itself* (including factors such as study design features, similarity of the tests, jobs, and job performance criteria used in the study) and evaluating the *comparability between the VG study and the new local situation*. The Joint Standards

and SIOP Principles seem to include more on the first; whereas the Uniform Guidelines focus more on the latter.

Legal Requirements for Title VII Cases

When it comes to adverse impact and tests, Title VII operates in a very straightforward manner. In short, any employer that uses a practice, procedure, or test that has significant adverse impact is required to make a justification by making a *job relatedness defense*. The actual language from Title VII states that adverse impact discrimination occurs when “. . . a complaining party [e.g., plaintiffs or government enforcement agencies] demonstrates that a respondent uses a particular employment practice that causes a disparate⁹ impact on the basis of race, color, religion, sex, or national origin, and the respondent fails to demonstrate that the challenged practice is job related for the position in question and consistent with business necessity.”¹⁰ Note that the law requires employers to *make a demonstration* that the test is job related *for the position in question*. Thus, in its very definition, Title VII displays a disconnect between its objectives and the goals of VG. When employers stand face-to-face with federal EEO law, they are tasked with the burden of demonstrating that their particular test—and use of that test—is specifically related to the particular position in question. VG, on the other hand, has a central goal of understanding how various tests or traits *may generalize across different settings*. How these fundamental differences have played out in litigation settings is discussed next.

Validity Generalization in the U.S. Supreme Court

Griggs v. Duke Power (1971)¹¹ was the first major adverse impact / test validation case that was tried by the U.S. Supreme Court after the passage of the 1964 Civil Rights Act. The EEO field was struggling to understand the language of the 1964 Act with respect to its requirements surrounding the definition of “adverse effect” (the term used during that era for adverse impact) and the linkage and connection that employers must make between their use of a given test that caused such effect and the requirements of the job. The specific matter of contention that arose in the Griggs case involved their use of a speeded general intelligence test (the Wonderlic), a mechanical comprehension test (the Bennett), and a high school education requirement to screen employees who desired employment in any division outside the general labor department. None of these requirements were intended to measure the applicant’s ability to *perform specific job duties* of a particular job or category of jobs.

Rather, a vice president of the Company testified that these tests were instituted on the Company’s judgment that they “. . . generally would improve the overall quality of the workforce.” The court ruled that the tests failed to “bear a demonstrable relationship to successful performance of the jobs for which it was used.” Both tests (i.e., general intelligence and mechanical comprehension) were adopted, as the Court of Appeals noted, “without meaningful study of their relationship to job-performance ability.” The court further ruled that, “What Congress has commanded (citing then-current EEO law) is that *any tests used must measure the person for the job and not the person in the abstract* . . . The touchstone is business necessity. If an employment practice which

operates to exclude blacks cannot be shown to be related to job performance, the practice is prohibited (emphasis added).”

Duke Power was aware that these tests had high levels of adverse impact against minorities, but they nonetheless continued their use under the assumption that the subgroup differences exhibited by the tests were commensurate with differences that existed between groups on job performance. However, the Supreme Court in their 8-0 decision, ruled that the tests needed to measure abilities that had a demonstrated, proven relationship to the specific requirements of the specific job—rather than the person in the abstract.

Validity Generalization in the 6th Circuit EEOC v. Atlas Paper (1989) Case

In the Sixth Circuit EEOC v. Atlas Paper (1989)¹² case, the court addressed VG head-on and decided—as a matter of fact and conclusion of law—that VG was inherently at odds with the U.S. Supreme Court case Griggs v. Duke Power and Title VII at its core. The Sixth Circuit rejected the use of VG to justify a test purporting to measure general intelligence (the 12-minute Wonderlic test), which had adverse impact when used for screening clerical employees. Without conducting a local validity study, the defense expert testified regarding the generalized validity of the challenged cognitive ability test, stating that it was “valid for all clerical jobs.” The lower District Court had previously approved Atlas’ use of the Wonderlic test, but the Court of Appeals reversed this decision and rejected the use of VG evidence as a basis for justifying the use of the test by stating:

We note in respect to a remand in this case that the expert failed to visit and inspect the Atlas office and never studied the nature and content of the Atlas clerical and office jobs involved. The validity of the generalization theory utilized by Atlas with respect to this expert testimony under these circumstances is not appropriate. Linkage or similarity of jobs in dispute in this case must be shown by such on site investigation to justify application of such a theory.

It is interesting to note that the requirement applied by the judge (above) is precisely what is currently required by the Uniform Guidelines for transporting validity evidence into a new situation (Section 7B). Both require that a job comparability study be done between the job in the original validation study and the new local situation.

The Sixth Circuit continued its critique of VG by stating:

The premise of the validity generalization theory, as advocated by Atlas' expert, is that intelligence tests are always valid. The first major problem with a validity generalization approach is that it is radically at odds with Albemarle Paper v. Moody, Griggs v. Duke Power, relevant case law within this circuit, and the EEOC Guidelines, all of which require a showing that a test is actually predictive of performance at a specific job. The validity generalization approach simply dispenses with that similarity or manifest relationship requirement. Albemarle and Griggs are

particularly important precedents since each of them involved the Wonderlic Test . . . Thus, the Supreme Court concluded that specific findings relating to the validity of one test cannot be generalized from that of others.

In issuing final judgment and conclusion of law, the judge relied on the applicability of another landmark U.S. Supreme Court case (Albemarle Paper v. Moody, 1975)¹³ findings regarding the situational specific validity requirements of Title VII, stating:

The kind of potentially Kafkaesque result, which would occur if intelligence tests were always assumed to be valid, was discussed in Van Aken v. Young (451 F.Supp. 448, 454, E.D. Mich. 1982, aff'd 750 F.2d. 43, 6th Cir. 1984). These potential absurdities were exactly what the Supreme Court in Griggs and Albemarle sought to avoid by requiring a detailed job analysis in validation studies. *As a matter of law . . . validity generalization theory is totally unacceptable under the relevant case law and professional standards (emphasis added).*

The Atlas case marks the clear “danger zone” that employers enter when relying solely on VG evidence in Title VII situations (i.e., when their testing practices exhibit adverse impact). Frank Landy cautioned that even if the Uniform Guidelines were changed to adopt a more open stance toward VG, a constitutional challenge would likely follow because “. . . they would then be at odds with established law—in particular the Sixth

Circuit Atlas case that dismisses VG as inconsistent with Albemarle and impermissible as a matter of law.”¹⁴

The requirement that tests must be proven job related and consistent with business necessity for the requirements of the at-issue position was unanimously framed by the U.S. Supreme Court in the Griggs case and was endorsed by Congress when it passed the Equal Employment Opportunity Act of 1972 (which amended Title VII of the Civil Rights Act of 1964). It was subsequently re-affirmed by the drafting of the 1978 Uniform Guidelines and the 1991 Civil Rights Act.

Rather than relying on broad-scale VG studies that attempt to show, in general, how specific tests or traits may relate to the requirements of a given position, conducting Uniform Guidelines-style “transportability” studies (to address Section 7B) offers higher levels of defensibility. Conducting a local validation study perhaps offers even higher levels of defensibility.

Local Criterion-related Validity in the Courts

When the courts evaluate criterion-related validity evidence in Title VII situations, they are interested in evaluating the *strength of the relationship* (i.e., the correlation) between the test and job performance (or other job-relevant performance metric). During this evaluation, four aspects of this relationship are typically included in the review: (1) the statistical significance of the correlation, (2) the practical significance of the correlation, (3) the type and relevance of the job criteria predicted by the test, and (4) the evidence to support the *specific use* (i.e., ranking, banding, pass/fail, or weighted

and combined with other tests) of the testing practice. If any of these elements are missing or do not meet reasonable levels, judicial processes will sometimes conclude with a finding of discrimination because the adverse impact was not sufficiently justified by the validity evidence. Each of these elements is discussed in more detail below.

Statistical significance. The courts, Uniform Guidelines, and professional standards share the same threshold when it comes to determining whether a statistical relationship between the test and job performance is “significant.” Unless a statistical test results in a finding that is below the .05 probability threshold, the evidence cannot even enter into the courtroom. This standard is applied to both sides of adverse impact litigation: for determining statistically significant adverse impact (using hypergeometric probability distributions) as well as determining the statistical significance of the correlation coefficient obtained in the validation study.

Practical significance. Like statistical significance, the concept of practical significance has also been applied to both the adverse impact and validity aspects of Title VII cases. When adverse impact is deliberated in Title VII cases, the courts have sometimes evaluated the practical significance or “stability” and effect size of the adverse impact.¹⁵ For example, some courts have re-evaluated statistically significant findings after hypothetically changing two applicants in the disadvantaged group from “failing” to “passing” status. If this hypothetical process changes the statistically significant finding from “significant” (<.05) to “non-significant” (>.05), the finding is not practically significant.

With criterion-related validity studies, practical significance relates to the *strength* of the validity coefficient—its practical utility in the specific setting. This is important in

litigation settings because the square of the validity coefficient represents the percentage of variance explained on the criterion used in the study. For example, a validity coefficient of .20 explains only 4% of the criterion variance, whereas coefficients of .40 explains 16%. Looking at this practically, if the test had a .40 correlation to a single job production metric (e.g., number of units produced per hour), this would mean that 16% of what makes that performance metric go up and down (high/low performance) is explained by whatever is being measured by the test—because they have 16% in common. A few cases that have deliberated the practical significance of validities are provided below.

- Dickerson v. U. S. Steel Corporation (1978)¹⁶: Regarding the validity coefficients in the case, the judge noted, “a low coefficient, even though statistically significant, may indicate a low practical utility” and further stated, “. . . one can readily see that even on the statistically significant correlations of .30 or so, only 9% of the success on the job is attributable to success on the (test) batteries. *This is a very low level, which does not justify use of these batteries, where correlations are all below .30.* In conclusion, based upon the guidelines and statistical analysis . . . the Court cannot find that these tests have any real practical utility. The Guidelines do not permit a finding of job-relatedness where statistical but not practical significance is shown. On this final ground as well, therefore, the test batteries must be rejected” (emphasis added).
- NAACP Ensley Branch v. Seibels (1980)¹⁷: The judge rejected statistically significant correlations of .21, because they were too small to be meaningful.

- EEOC v. Atlas Paper (1989)¹⁸: The judge weighed the decision heavily based on the strength of the validity coefficient: “There are other problems with (the expert’s) theory which further highlight the invalidity of the Atlas argument. Petty computed the average correlation for the studies to be .25 when concurrent and .15 when predictive. A correlation of .25 means that a test explains only 5% to 6% of job performance. Yet, Courts generally accept correlation coefficients above .30 as reliable . . . This Court need not rule at this juncture on the figure that it will adopt as the bare minimum correlation. Nonetheless, the Court also notes that higher correlations are often sought when there is great disparate impact (Clady v. County of Los Angeles, Guardians Assn of New York City v. Civil Service, 630 F.2d at 105-06). Thus, despite the great disparate impact here, the correlations fall significantly below those generally accepted.”
- U.S. v. City of Garland (2004)¹⁹: The court thoroughly debated the level of the validity coefficients: “As discussed supra at n. 25, whether the correlation between the Alert (test) and performance should be characterized as ‘low’ or ‘moderate’ is a matter of earnest contention between the parties. (See D.I. 302 at p. 11, 35-40.) In a standard statistical text cited at trial, correlations of .10 are described as ‘low’ and correlations of .30 described as ‘moderate.’”

In addition to the courts, the Uniform Guidelines (15B6), U.S. Department of Labor (2000, p. 3-10), and SIOP Principles (p. 48) all provide a discussion regarding the importance of taking the strength of the validity coefficient into practical consideration.

Type and relevance of the job criteria. Evaluating just exactly what is measured by the test (i.e., what aspect of performance is predicted by the test) is also crucial to weighing the overall quality of the criterion-related validity study. Criterion-related validity studies can correlate test scores to objective (e.g., number of units produced per hour, sales, etc.) or subjective factors (e.g., supervisory ratings, peer ratings). In addition to the cases above (each of which discussed the area of job performance predicted by the test), the Uniform Guidelines (15B6) and SIOP Principles (p. 16) include discussion on this topic. The courts will obviously more carefully scrutinize tests that have high levels of adverse impact, but are only correlated to minor aspects of job performance.

Considering the validity coefficient and the specific use of the testing practice. In most situations where written tests are used for selection or promotion purposes, the way in which test scores are used will make a significant difference on the degree of adverse impact that will occur. In fact, when adverse impact is calculated, it matters more how the scores are used—i.e., how many of each group pass the test versus fail—than how far apart the average score differences are between groups. This is because even a test with a large gap between the average scores of two groups will not produce adverse impact until the scores are *used* (e.g., ranked, banded, or used on a pass/fail basis with certain results for each group) in a way that tips the scales of a statistical test (such as the Fisher Exact Test) below the .05 significance level based upon the difference in success rates between the groups. For example, a written test that has a standardized mean group difference (*d*) value of 1.0 (indicating a 44.6% overlap of the scores between the two groups) will still not produce statistically detectable adverse impact until enough applicants are actually hired that cause the adverse impact to cross the .05 significance threshold.

Because of this phenomenon, test usage, adverse impact, and test validity are issues that are simultaneously considered in litigation settings. A test that has high validity (e.g., a correlation exceeding .30), high reliability, and good score dispersion (characterized by having differentiating scores between applicants, rather than the scores being bunched together) will be more likely to withstand legal scrutiny when compared to a test that is not marked by these characteristics.²⁰ Generally speaking, the higher the adverse impact caused by the use of the test (typically with ranking being the highest and pass/fail cutoffs being the lowest), the higher level of validity evidence will be required.²¹ Because ranking typically exhibits higher levels of adverse impact, employers that practice ranking are typically subjected to a more stringent validity standard than when pass/fail cutoffs are used, as demonstrated by the following cases:

- Brunet v. City of Columbus (1993)²²: This case involved an entry-level firefighter Physical Capacities Test (PCT) that had disparate impact against women. The court stated, “The correlation coefficient for the overall PCT is .29. Other courts have found such correlation coefficients to be predictive of job performance, thus indicating the appropriateness of ranking where the correlation coefficient value is .30 or better.”
- Boston Chapter NAACP Inc. v. Beecher (1974)²³: This case involved an entry-level written test for firefighters. Regarding the correlation values, the court stated: “The objective portion of the study produced several correlations that were statistically significant (likely to occur by chance in fewer than five of one

hundred similar cases) and practically significant (correlation of .30 or higher, thus explaining more than 9% or more of the observed variation).”

- Clady v. County of Los Angeles (1985)²⁴: This case involved an entry-level written test for firefighters. The court stated: “In conclusion, the County’s validation studies demonstrate legally sufficient correlation to success at the Academy and performance on the job. Courts generally accept correlation coefficients above .30 as reliable ... As a general principle, the greater the test’s disparate impact, the higher the correlation which will be required.”
- Zamlen v. City of Cleveland (1988)²⁵: This case involved several different entry-level firefighter physical ability tests that had various correlation coefficients with job performance. The judge noted that, “Correlation coefficients of .30 or greater are considered high by industrial psychologists” and set a criteria of .30 to endorse the City’s option of using the physical ability test as a ranking device.

The Uniform Guidelines (3B, 5G, and 15B6) and SIOP Principles (p. 49) also advise taking validity levels into consideration when considering how to use a test in a selection process. This issue of test usage is an important one because validity has to do more with *actual test scores* than it does for tests as a whole. Consider a typing test for example. A typing test, per se, is not valid by itself because validity depends on two major factors: the position a test is used for and the way the test is used (i.e., the particular scores under consideration). A typing test may be valid for the position of Personnel Analyst and Legal Secretary alike, but minimum cutoff scores of 45 WPM and 80 WPM will mean different things for each of these positions. As such, specific scores may or may not be

valid given how closely they are aligned with the true needs of the job, and the qualification level to which they are aligned.

In Title VII situations, employers who have relied solely on VG evidence to “infer validity” will not have the information necessary to show the court that these four critical factors have been properly supported. In fact, if the only evidence offered is VG evidence, they will have no solid evidence to offer the courts with respect to *any* of these four factors (because VG relies essentially on inferring validity based on other studies). This is because there is no way to tell if a local study would result in a validity coefficient that is statistically significant, if such validity coefficient would be practically significant, if the job criteria predicted by the test was relevant given the needs of the particular position, or if the validity coefficient would sufficiently justify the specific use of the testing practice. This presents a major challenge for employers who opt for VG-only defenses in Title VII situations.

Benefits of Conducting Local Criterion-related Validity Studies: An Example Using a Mock Trial

Local criterion-related validity studies are more defensible than VG studies in litigation settings. While there are several reasons for this, the most basic one is that local and specific evidence of adverse impact requires some type of *local and specific justification*. This is because there are two bodies of evidence that are weighed on the balance scale in Title VII adverse impact litigation—adverse impact and validity. Judges are forced to weigh the degree of adverse impact caused by the test against the level of

adverse impact stacked on the other side of the scale. In virtually all adverse impact cases, the specific level of adverse impact (e.g., the probability value of the inference test, the difference in success rates between the two groups, etc.) will be known. Given this, should employers be allowed to just “borrow” VG evidence from outside the local situation to justify the degree of adverse impact observed locally? Should *generic* evidence be allowed on only one side of the scale? On the flipside, one can only imagine how outraged employers would be if a plaintiff group filed Title VII charges against them based on adverse impact evidence found at 20 similar employers and there was no adverse impact in the local situation.

Producing local validity evidence is a straight-forward process. If employers are using a type of test that cannot be content validated (e.g., an aptitude or personality oriented test), they can either conduct a local criterion-related validity study (if sample sizes are sufficient) or transport validity from another criterion-related validity study by conducting a job comparability study (as outlined in 7B of the Uniform Guidelines).

Employers, plaintiff groups, and federal enforcement agencies (e.g., EEOC, OFCCP) spend hundreds of millions every year in litigation issues surrounding adverse impact and test validity. Because written tests typically show the highest levels of adverse impact against minorities,²⁶ they are typically the target of much of this type of litigation. When such a challenge is made against an employer’s written test, they wind up in court deliberating the finer points of the sophisticated topic of test validity. Plaintiff and defense experts wrestle over whether the test sufficiently addressed the technical and complex requirements of federal (the Uniform Guidelines) and professional (the Joint Standards and SIOP Principles) standards. Judges (and juries when involved) are

sometimes overwhelmed by statistical terms of art and mathematically complex explanations that they sometimes haven't visited since college.

When it comes to making a decision to side for or against the employer, judges will sometimes rely on the parts of the validation battle they understood, or those that made the best sense to them along practical lines. Sometimes courts rely on previous cases that have already wrestled with technical issues surrounding complex issues like banding test scores or how to determine an acceptable pass/fail cutoff for a test. While some judges may have a better background than others for dealing with complex issues like “standard deviations,” validity coefficients, and “range restriction” it is safe to say that technical spin is not appreciated by the courts, and their preference typically lies along the lines of “*show me* how the test fits the needs of *this particular job*.”

Given this context, consider the example courtroom dialogue below for a Title VII case involving a written test that exhibited adverse impact for a certain position. Each question is asked by the federal judge with a hypothetical response provided from a testing expert who conducted a local validation study and one who conducted a VG study. The federal or professional requirement related to the questions and responses is provided, as well as some examples of court cases where the factor has been deliberated.

Federal Judge: In this case, 100 minorities and 100 whites (200 total) took the written test. Of the 100 applicants that passed (an overall 50% passing rate for both groups combined), 40 were minorities (a 40% success rate) and 60 were whites (a 60% success rate). This difference in passing rates results in an adverse impact finding of 2.69 standard deviations (which exceeds the 1.96 threshold applied in Title VII cases), which

equates to a probability value of .004. The odds of this probability value are only 1 in 280, which is well beyond the “1 chance in 20” (or 5%) associated with the 1.96 standard deviations needed for proving statistically significant adverse impact.

What type of study or research was conducted to address this employer’s legal burden of demonstrating that this test is “*job related for the position in question and consistent with business necessity*” that will lead this court to justify (or condemn) the shortfall of 10 minorities that resulted by its use?²⁷

VG Expert: To justify the use of this test, we conducted a VG study by gathering data from 22 separate validation studies that were conducted at other employers for similar positions that used similar tests for pre-employment testing. These studies spanned about 10 years and included a total of 2,950 subjects. We are relying on the results of this study to give us insight regarding how the test may be related to the performance of this job at this employer.

Local Validity Expert: This test is justified based on an empirical study we conducted at this employer’s location that involved the at-issue test and the incumbents who work in the at-issue position. The study involved gathering and analyzing two types of data that were obtained at this employer: test scores and job performance ratings. We gave the test to 130 incumbents and compared these scores to job performance ratings given by supervisors of the target position. By statistically correlating these two data sets, we were able to determine *if* and *to what extent* this test was related to job performance.

Federal Judge: Based on the study you conducted, what is the *strength of the relationship* (i.e., validity coefficient) between this test and the performance of this job?²⁸

VG Expert: We don't know for sure what the validity coefficient is or would be in this *specific situation*. However, the results of my VG study indicate that the *operational validity coefficient* (corrected for range restriction and criterion reliability) for this type of test and positions similar to this one is .42. Using the conventional practices for VG studies, the results of our study indicates that 90% of the true validity coefficients in the population (for all similar settings) exceeds .33. This credibility interval demonstrates that validity generalizes across situations (including this one) because the 90% interval does not contain zero.²⁹

Local Validity Expert: The results of our local study revealed a .25 validity coefficient between test scores and job performance. This level of validity is classified as "likely to be useful" by the U.S. Department of Labor standards and demonstrates that the test captures about 6.3% of job performance (.25 squared) for this position at this employer. However, when corrected for range restriction and criterion unreliability, the operational validity is .42. This level of observed validity is close to what other federal courts have deemed acceptable under similar Title VII situations.

Federal Judge: Are scores from this test correlated with the performance of the target position at a level that is *statistically significant* under the federal and professional requirements of the testing field (e.g., the Uniform Guidelines and professional standards)?³⁰

VG Expert: I don't know for certain whether this test is significantly related to the performance of this job; however, the test has shown this type of relationship in other

similar situations. Of the 22 validation studies included in my VG analysis, 17 were above the threshold required for statistical significance, and 5 were not.

Local Validity Expert: The statistical significance value of this validity coefficient is .002 (using a one-tail test for a directional hypothesis) or .004 (using a two-tail test), which is statistically significant under the Uniform Guidelines and professional standards. A probability value of .005 indicates that the odds of this finding being a “chance occurrence” (rather than a legitimate relationship) are only 1 in 500 (or 1 in 250 for the two-tail probability value). Because this finding indicates that the test is significantly correlated with an important aspect of job performance, we believe the test is sufficiently valid and is acceptable for use.

Federal Judge: How did you know what correction levels were appropriate to use for estimating “operational” validity in this situation?³¹

VG Expert: We based our calculations on the range restriction and criterion reliability data that came from some of the studies in our sample. The range restriction values ranged from .58 to .87, with an average of .67 (which is common in the testing field³²) and were based on other employers that used similar tests in a variety of testing circumstances. We based our corrections for criterion unreliability the same way—using the criterion reliabilities that were available from some of the studies in our VG analysis, which ranged from .70 to .90, and averaged about .80.³³ The criteria reliabilities varied based on whether they were objective or subjective criteria, and the degree to which raters were consistent with giving their ratings. We did not base our VG study on the range restriction or criterion reliability values that exist in this situation.

Local Validity Expert: We based our corrections on the range restriction and criterion reliability data from this local situation, for the test of interest. The test was used with a cutoff score of 70%, which resulted in a range restriction ratio of .67, and we used this value in our computations. The reliability for the supervisor ratings we used as the criteria for our study was .79 and was based on the consistency between the supervisors who gave job performance ratings on the incumbents who took the test. Correcting the .25 validity coefficient we attained (between test scores and overall job performance) for these two factors increased our reported validity coefficient from .25 to .42.

Federal Judge: What *aspect of job performance* is correlated with the test? How important is this part of job performance relative to other parts of the job? What negative consequences are likely to occur on the job if that part of the job is not performed adequately?³⁴

VG Expert: We are not sure what aspect of job performance would specifically be predicted by this test, or the negative outcomes associated with them if performed poorly. However, the test would likely predict some facet of job performance similar to those included in the VG study, which range from subjective supervisory ratings to objective measures such as productivity.

Local Validity Expert: We had supervisors provide job performance ratings on five distinct aspects of job performance for each of the incumbents that were included in our study. We also asked supervisors to provide an overall rating of job performance, for a total of six ratings. Of the five different aspects of job performance, the test was significantly correlated with two (teamwork with a validity coefficient of .32 and

productivity with a validity coefficient of .29), as well as with the overall rating ($r = .25$). The test did not result in statistically significant correlations with the other three job performance criteria, although it was close with one, and the other two were not negative. This indicates that the test was positively and significantly correlated with two specific areas of job performance (with the strongest relationship with teamwork), as well as overall job performance. All three of these areas are critical to the effective functioning of our work units. Incumbents who practice effective teamwork behaviors foster a more conducive and productive workforce, work productivity is obviously highly critical, and the overall job performance ratings predicted by this test assures us that incumbents will be quality workers (as perceived and rated by their supervisors).

Federal Judge: How does the validity evidence justify how the test was used? What evidence exists to demonstrate the validity of the *cutoff score* used? How does the cutoff relate to the *normal expectations of acceptable proficiency in the workforce*?³⁵

VG Expert: The collection of evidence in the VG study we conducted demonstrates that the type of test used by this employer typically has a strong relationship with job performance for a wide variety of positions, including the one at-issue in this case. Because of the extensive validity evidence provided by our study, a wide degree of cutoff choices can be used with varying levels of utility as a result. Because our study shows that there is generally a linear relationship between test scores and job performance, higher job performance levels can be expected when higher cutoff levels are used.

Local Validity Expert: Our validation study shows that the validity coefficient of .25 provides a high degree of utility and effectiveness in helping the company screen in qualified workers. Assuming an applicant base rate of 50% (the percentage of applicants who “show up qualified” for the position prior to being screened by any testing), the 70% cutoff will result in 64% of the applicant group being qualified for the at-issue position (a 14% improvement over using no test or lower cutoff scores). Using the corrected validity of .42, this rises to 74% of the applicants being qualified—24% higher than would be hired when compared to not using a test. This cutoff screens in only the top 20% of applicants; however the return. Further, this cutoff will likely result in increasing the performance levels of our current workforce by 7% (12% when based on the operational validity of .42).

Federal Judge: What controls were used to insure the criterion ratings were free from bias and “criterion contamination”?³⁶

VG Expert: Because no study was conducted at this employer’s site for this position, I cannot reply to this. However, I can describe the protocols followed to insure safeguards against these concerns for some of the studies in my VG analysis.

Local Validity Expert: The survey we administered to supervisors to gather job performance ratings included behaviorally-anchored rating scales that defined the five dimensions of job performance as objectively as possible. Examples were given of high- and low-performance in each dimension. Before the supervisors completed the surveys we provided instructions to limit their ratings to these specific areas of job performance, and to be aware of various rating errors that could possibly bias their ratings. Further, we

made safeguards to insure that supervisors not have access to test score information on the people they rated. This reduced our chances of contaminating the job performance (criterion) data.

Federal Judge: Is the test a fair predictor for all groups?³⁷ Was a fairness study conducted to be sure that the test operated similarly for both groups on predicting job success outcomes?

VG Expert: Some of the studies included in our VG analysis included a fairness study to evaluate if the test operated differently between subgroups; however, no study was conducted for this particular test, employer, or position.

Local Validity Expert: To evaluate the test for fairness, we conducted a statistical analysis that compared how the correlation between test scores and job performance operated when taking ethnicity into account. By creating interaction terms between these variables, we were able to determine whether the test acted in an unfair way between various subgroups. No significant differences were found.

Which of these experts sounds most convincing? In a courtroom filled with plaintiff class members—all waiting to see what solid evidence exists to justify the disproportionate passing rate on the test—which series of responses seems most compelling? The outcome seems obvious—without conducting a local validity study, employers cannot *really know* if a test is valid. This example is not far fetched. In fact, several of the court cases (footnoted) have already had lengthy deliberations on each of

these topics. Given these limitations, recommendations are provided below when considering integrating VG into a validation defense in Title VII situations.

Recommendations

There are two main reasons for conducting validation research: effectiveness and defensibility. Using a test that is not effective is a waste of applicant time and the employer's money. Using a test that is not defensible (even if it is effective) can ruin the employer's reputation in the marketplace and cost hundreds of thousands (or more) in litigation costs. Based on why the test was not defensible, it can also be tied back to a judicial finding of discrimination, a ruling that no employer wants levied.

For these reasons, employers should carefully choose their validation strategy. While VG studies can provide useful insights to how a given test or trait may work across various settings, applicant groups, and employers, it may not present the strongest validity defense in Title VII situations. As such, employers that are faced with Title VII situation when they are using tests that cannot be content validated, have a couple of sound options. First, they can conduct a local criterion-related validity study if the sample sizes are sufficient (see discussion below). Second, they can conduct a transportability study to address Section 7B of the Uniform Guidelines. And, if the employer desires to rely on VG evidence to supplement either of these two approaches, the following guidelines are presented:

When VG evidence is evaluated in a Title VII situation:

1. Address the evaluation criteria provided by the Uniform Guidelines, Joint Standards, and SIOP Principles regarding an *evaluation of the internal quality of the VG study*.³⁸ This will help insure that the VG study itself can be relied upon for drawing inferences.

2. Address the evaluation criteria provided by the Uniform Guidelines, Joint Standards, and SIOP Principles regarding the *similarity between the VG study and the local situation*.³⁹ These will help insure that the VG study itself can be relied upon and the research is in fact relevant to the local situation (e.g., similarities between tests, jobs, job criteria, etc.). Perhaps the most critical factor evaluated by courts when considering VG-types of evidence in litigation settings is the similarity between jobs in the VG study and the local situation (see also 7B of the Uniform Guidelines). VG evidence is strongest when there is clear evidence that the work behaviors between the target position and those in the positions in the VG study are highly similar as shown by a job analysis in both situations (as suggested by the original authors of VG).⁴⁰

3. Only use VG evidence to supplement other sources of validity evidence (e.g., content validity or local criterion-related validation studies) rather than being the sole source. Supplementing a local criterion-related validity study with evidence from a VG study may be useful if an employer has evidence that statistical artifacts (such as range restriction)—not situational moderators—suppressed the actual validity of the test in the local situation.

Further, employers with only limited subjects available for a local criterion-related validity study may benefit from supplementing their local validation research with

VG evidence (provided that their local study demonstrates at least minimal levels of validity with respect to statistical significance, practical significance, the use of relevant criteria, and the test is used appropriately given this evidence and the levels of adverse impact observed).

For example, an employer wishes to supplement the validity evidence of their test for the at-issue position, and only has 70 subjects available to conduct a local validation study (i.e., has low statistical power for conducting a study). The study returns only a moderate (but significant) correlation between test scores and relevant job performance criteria and it is likely that this moderate result is due to sampling error, criterion unreliability, and range restriction (rather than legitimate situational differences between those included in the VG study and the new local situation). In these circumstances, it may be useful to draw inferences from professionally conducted VG studies that may show that higher levels of validity could be expected after accounting for these three statistical suppressors.

4. Evaluate the test fairness evidence from the VG study using the methods outlined by the Uniform Guidelines, Joint Standards, and SIOP Principles.

5. Evaluate and consider using “alternate employment practices” that are “substantially equally valid” (as required by the 1991 Civil Rights Act⁴¹ and Section 3B of the Uniform Guidelines).

¹ Uniform Guidelines – Equal Employment Opportunity Commission, Civil Service Commission, Department of Labor, and Department of Justice (August 25, 1978), Adoption of Four Agencies of Uniform Guidelines on Employee Selection Procedures, 43 Federal Register, 38,290-38,315.

² Joint Standards — American Educational Research Association, the American Psychological Association, and the National Council on Measurement in Education (1999), *Standards for Educational and Psychological Testing*. Washington DC: American Educational Research Association.

³ SIOP Principles — SIOP (Society for Industrial and Organizational Psychology, Inc.) (1987, 2003), *Principles for the Validation and Use of Personnel Selection Procedures* (3rd and 4th eds). College Park, MD: SIOP.

⁴ It is noted, however, that the professional standards (i.e., the Joint Standards and SIOP Principles) regard validity as a more universal concept, with five various sources of evidence that can be used in its evaluation (e.g., see the Principles, 2003, p. 4).

⁵ 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.

⁶ Hunter, J. E., & Schmidt, F. L. (2004). *Methods of meta-analysis: Correcting error and bias in research findings* (2nd ed.). Newbury Park, CA: Sage.

⁷ Murphy, K. R. (2003). The logic of validity generalization. In K. R. Murphy (Ed.) *Validity generalization: a critical review*. Mahwah, NJ: Erlbaum.

⁸ An 80% credibility interval indicates that 90% of the studies will not include a “zero validity” value because it uses a one-tail interval around the mean of the “true validity” population using the standard deviation of infinite-sample effect sizes. So, for positive correlations, 10% are zero or less and 10% lie beyond the upper bound of the interval. And, when the 90% interval does not contain zero, it is said that one can have confidence that a relationship generalizes across those situations examined in the study and can be generalized to new situations.

⁹ The terms “disparate impact” and “adverse impact” are used interchangeably in this document.

¹⁰ 1991 Civil Rights Act (42 U.S.C. §2000e-2[k][1][A][i]).

¹¹ *Griggs v. Duke Power Company*, 401 U.S. 424 (1971).

¹² *EEOC v. Atlas Paper*, 868 F.2d. 487, 6th Cir., cert. denied, 58 U.S. L.W. 3213, (1989).

¹³ *Albemarle Paper v. Moody*, 422 U.S. at 423, 95 S.Ct. 2362 (1975).

¹⁴ Landy, F. J. (2003). Validity generalization: then and now. In K. R. Murphy (Ed.), *Validity generalization: a critical review* (pp. 155-195). Mahwah, NJ: Erlbaum.

¹⁵ See *Contreras v. City of Los Angeles* (656 F.2d 1267, 9th Cir., 1981), *US v. Commonwealth of Virginia* (569 F.2d 1300, CA-4 1978, 454 F. Supp. 1077), *Waisome v. Port Authority* (948 F.2d 1370, 1376, 2d Cir., 1991).

¹⁶ *Dickerson v. U. S. Steel Corporation*, 472 F. Supp. 1304, E.D. Pa. (1978).

¹⁷ *NAACP Ensley Branch v. Seibels*, 13 E.P.D. 11,504 at pp. 6793, 6803, 6806, N.D.Ala. (1977), *aff'd* in relevant part, *rev'd* in other part, 616 F.2d 812,818 and note 15 (5th Cir.), *cert. den.*, 449 U.S. 1061 (1980).

¹⁸ *EEOC v. Atlas Paper*, *supra*, footnote #24.

¹⁹ *U.S. v. City of Garland*, WL 741295, N.D.Tex. (2004).

²⁰ See, for example: *Guardians Association of the New York City Police Dept. v. Civil Service Commission*, 630 F.2d 79, 88, 2d Cir., 1980; *cert. denied*, 452 U.S. 940, 101 S.Ct. 3083, 69 L.Ed.2d 954 (1981).

²¹ See the Uniform Guidelines (1978), Section 15B6.

²² *Brunet v. City of Columbus*, 1 F.3d 390, U.S. Ct. App. Sixth Cir. (1993).

²³ *Boston Chapter, NAACP Inc. v. Beecher*, 504 F.2d 1017, 1021, 1st Cir. (1974).

²⁴ *Clady v. County of Los Angeles*, 770 F.2d 1421, 1428, 9th Cir. (1985).

²⁵ *Zamlen v. City of Cleveland*, 686 F.Supp. 631, N.D. Ohio (1988).

²⁶ Sackett, P. R., Schmitt, N., Ellingson, J. E., Kabin, M.B. (April, 2001). High-stakes testing in employment, credentialing, and higher education: prospects in a post-affirmative-action world. *American Psychologist*, 56 (4), pp. 302-318.

²⁷ 1991 Civil Rights Act, *supra*.

²⁸ Uniform Guidelines (1978), Sections 14B5, 15B8. Principles (2003), p. 19. Standards (1999), Standard 1.18. See also *Boston Chapter NAACP Inc.*; *Brunet*; *Clady*; *Zamlen*, *supra*.

²⁹ Hunter, J. E., & Schmidt, F. L. (1990). *Methods of meta-analysis: correcting error and bias in research findings*. Newbury Park, CA: Sage.

³⁰ Uniform Guidelines (1978), Sections 14B5, 15B8. Principles (2003), p. 19. It is widely accepted in litigation settings that the .05 threshold applies where adverse impact or validity is evaluated.

-
- ³¹ Uniform Guidelines (1978), Section 15B8. Principles (2003), p. 19. Standards (1999), Standard 1.21. U.S. v. State of Delaware (WL 609331, D.Del., 2004); Dickerson, supra, footnote 36; Bernard v. Gulf Oil Corporation, 841 F.2d 547, C.A.5 (Tex.), 1988.
- ³² Hunter, J. E., & Schmidt, F. L. (2004). *Methods of meta-analysis: correcting error and bias in research findings* (2nd ed.). Newbury Park, CA: Sage (p. 230).
- ³³ Estimate of criterion reliabilities range across various studies, with some reporting values of .52 (for overall performance) (Viswesvaran, C., Ones, D. S., & Schmidt, F. L. [1996]. Comparative analysis of the reliability of job performance ratings. *Journal of Applied Psychology*, 81[5], 557-574) and others adopting values as high as .80 (Hartigan, J. A., & Wigdor, A. K. [1989]. *Fairness in employment testing: Validity generalization, minority issues, and the General Aptitude Test Battery*. Washington, DC: National Academy Press [page 5]). Therefore, the estimates used in this example should be viewed as conservatively high, which underestimate the operational validity in this example.
- ³⁴ Uniform Guidelines (1978), Sections 14B6, 15B3, 15B5. Principles, pp. 15-16. Standards, pp. 11-12. Lanning v. SEPTA, 181 F.3d 478 (3rd Circuit 1999); Dickerson, supra.
- ³⁵ Uniform Guidelines (1978), Sections 14B6, 15B3, 15B5, 15B10. Principles (2003), pp. 46-47. Standards (1999), pp. 152-157.
- ³⁶ Uniform Guidelines (1978), Sections 14B2-3, 15B5-6. Principles (2003), pp. 16-17.
- ³⁷ Uniform Guidelines (1978), Sections 14B8, 15B8. Principles (2003), pp. 31-33.
- ³⁸ See the Uniform Guidelines, Section 7B and 15E1, the Joint Standards (1999), p. 15 and Standard 1.20 and 1.21, and the SIOP Principles (2003), pp. 9-10, 28-30.
- ³⁹ See the Uniform Guidelines, Section 7B and 15E1, Questions & Answers #43, the Joint Standards (1999), p. 15 and Standard 1.20 and 1.21, and the SIOP Principles (2003), pp. 9-10, 28-30.
- ⁴⁰ 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 (see p. 530).
- ⁴¹ 1991 Civil Rights Act, Section 2000e-2[k][1][A][ii].