Validity Generalization vs. Title VII: Can Employers Successfully Defend Tests without Conducting Local Validation Studies?

By Daniel A. Biddle, Ph.D. and Patrick M. Nooren, Ph.D.

Daniel Biddle is the CEO of Biddle Consulting Group, Inc. and Fire & Police Selection, Inc. He has been in the EEO/HR field for over 15 years and has been an expert witness and/or consultant in over 50 EEO-related cases, among other things. He is the author of* **Adverse Impact and Test Validation: A Practitioner's Guide to Valid and Defensible Employment Testing.** *His primary focus at BCG is test development and validation and working as an expert in EEO litigation matters.*

Patrick Nooren is the Executive Vice President of Biddle Consulting Group, Inc, with over 12 years experience in the EEO/HR field, working on the technical components of numerous small- and large-scale EEO cases. His primary focus at BCG is oversight of the EEO/AA division.

The 1991 Civil Rights Act requires employers to justify tests with disparate impact by demonstrating they are sufficiently "job related for the position in question and consistent with business necessity." This requirement is most often addressed by conducting validation studies to establish a clear connection between the abilities measured by the test and the requirements of the job in question.

Building a validation defense strategy in such situations requires employers to address the federal Uniform Guidelines on Employee Selection Procedures (1978), professional standards, and relevant court precedents. In recent years, some employers have attempted to "borrow" validation evidence obtained by other employers for similar positions rather than conduct their own local validation study. This strategy relies on a methodology known as "validity generalization" (VG). Despite the increase in popularity among test publishers and HR/hiring staff at corporations, relying entirely on VG to defend against Title VII disparate impact suits will likely lead to disappointing outcomes because the courts have generally required employers demonstrate *local and spe-*

cific validation evidence where there is *local and specific* evidence of disparate impact.

The goal of this article is to review Title VII requirements for establishing validity evidence, overview federal and professional requirements for validation strategies (specifically VG), outline how some courts have responded to VG strategies, and conclude by providing recommendations for validating tests that come under Title VII scrutiny.

OVERVIEW OF TITLE VII DISPARATE IMPACT DISCRIMINATION

The 1991 Civil Rights Act states disparate impact discrimination occurs when ". . . a complaining party demonstrates that a respondent uses a particular employment practice that causes a disparate 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."[1] Disparate impact occurs when two groups have substantially different passing rates on a test, and is normally evaluated using tests for both *statistical* (*i.e.*, whether the differences in passing rates are beyond what would be expected by chance) and *practical* significance (the practical impact or stability of the findings). When tests have such disparate impact, a finding of unlawful discrimination will likely be the judgment, absent an acceptable demonstration of the "job relatedness" of the test.

The basic necessity of providing "job relatedness" evidence for the test causing disparate impact has been set in stone since the famous U.S. Supreme Court *Griggs v. Duke Power*[2] case. However, during a two year period between 1989 and 1991, under the then-reigning U.S. Supreme Court *Wards Cove v. Atonio*[3] case, this standard was lowered. Under the *Wards Cove* standard, employers only needed to "produce a business justification." "Producing a justification" is a much less stringent requirement than "demonstrating job relatedness." Congress overturned this standard in 1991 with the passage of the 1991 Civil Rights Act, which reinstated the original *Griggs* standard (where it stands today).

Fundamental elements from the *Griggs* case were encapsulated into the federal treatise to enforce Title VII—the Uniform Guidelines on Employee Selection Procedures, a document jointly developed in 1978 by the U.S. Equal Employment Opportunity Commission, Department of Justice, Department of Labor, and the Civil Service Board, now the Office of Personnel Management (discussed in more detail below).

While the Uniform Guidelines have remained unchanged since 1978, the courts have continued to support one very important component: when an employer uses a specific test for a particular job, and such test has disparate impact, the employer must justify the use of the test by demonstrating that the test is job related. This is because Title VII requires a *specific justification for both the test itself as well as how it is being used* (*e.g.*, ranked, banded, used with a minimum cutoff, or weighted with other selection procedures) in *specific situations where disparate impact exists*.

TEST VALIDATION METHODS FOR DEMONSTRATING JOB RELATEDNESS

Challenges to an employer's testing practices can come from enforcement agencies (*e.g.*, the U.S. Equal Employment Opportunity Commission, Department of Justice, Department of Labor via the Office of Federal Contract Compliance Programs, state equal opportunity commissions) or from private plaintiffs' attorneys. In these situations, employers will generally need to defend their testing practices by demonstrating validity under the Uniform Guidelines and professional standards (the SIOP Principles and Joint Standards). Each set of standards is discussed briefly below.

Uniform Guidelines

The Uniform Guidelines are designed to enforce Title VII and were adopted by federal agencies to provide a uniform set of principles

governing the use of employee selection procedures.[4] The Uniform Guidelines define their "basic principle" as:

> A selection process which has a disparate impact on the employment opportunities of members of a race, color, religion, sex, or national origin group . . . and thus disproportionately screens them out is unlawfully discriminatory unless the process or its component procedures have been validated in accord with the Guidelines, or the user otherwise justifies them in accord with Federal law . . . This principle was adopted by the Supreme Court unanimously in *Griggs v. Duke Power Co.* (401 U.S. 424), and was ratified and endorsed by the Congress when it passed the Equal Employment Opportunity Act of 1972, which amended Title VII of the Civil Rights Act of 1964.[5]

Although they are not law, the Uniform Guidelines have been given great deference in federal litigation or enforcement settings where tests have exhibited disparate impact. This "great deference" endorsement was initially provided by the U.S. Supreme Court in *Albemarle Paper v. Moody*,[6] and has subsequently been similarly recognized in at least 20 additional federal cases.[7] The Uniform Guidelines have also been cited and used as the standard in hundreds of court cases at all levels.

Three primary types of validation evidence are presented in the Uniform Guidelines: content, criterion-related, and construct (listed below in the order most frequently used by employers):

- Content validity: Demonstrated by showing the content of a selection procedure is representative of important aspects of performance on the job. (See sections 5B and 14C)
- Criterion-related validity: Demonstrated empirically by showing the selection procedure is predictive of, or significantly correlated with, important elements of work behavior. (See sections 5B and 14B)
- Construct validity: Demonstrated by showing the selection procedure measures the degree to which candidates have identifiable characteristics which have been determined to be important for successful job performance. (See sections 5B and 14D)

The Uniform Guidelines also support a limited form of VG (called "transportability") to be used when "transporting" the use of a test from one situation or location to another (see Section 7B, discussed below). They also provide criteria for inferring validity evidence based on studies conducted elsewhere (see Section 15E, also discussed below).

Professional Standards: Joint Standards & SIOP Principles

The National Council on Measurement in Education (NCME), American Psychological Association (APA), and the American Educational Research Association (AERA) cooperatively released the Joint Standards in 1999. The purpose of the Joint Standards is to provide criteria for the evaluation of tests, testing practices, and test use for professional test developers, sponsors, publishers, and users that adopt the Standards.[8] One of the fifteen chapters (Chapter 14) is devoted exclusively to testing in the areas of employment and credentialing. The remaining chapters include recommended standards for developing, administering, and using tests of various sorts.

SIOP, the Society for Industrial and Organizational Psychology (Division 14 of the APA), is an association of about 3,000 I-O psychologists, some of whom specialize in developing and validating personnel tests. SIOP published an updated version of the SIOP Principles in 2003, a document offered as an official SIOP policy statement regarding personnel test development and validation practices. This document was also approved as policy by the APA Council of Representatives in August 2003.

Both the Joint Standards and the SIOP Principles are in agreement on the essential definition of validity, stating that validity is a "unitary

concept" with ". . . different sources of evidence contributing to an understanding of the inferences that can be drawn from a selection procedure" (Standards, p. 4). The Joint Standards and SIOP Principles collectively allow five different sources of evidence to generate validity evidence under this "unitary concept" framework:

- Relationships between predictor scores and other variables, such as selection procedure–criterion relationships;
- Content (meaning the questions, tasks, format, and wording of questions, response formats, and guidelines regarding administration and scoring of the selection procedure. Evidence based on selection procedure content may include logical or empirical analyses that compare the adequacy of the match between selection procedure content and work content, worker requirements, or outcomes of the job);
- Internal structure of the selection procedure (*e.g.*, how well items on a test cluster together);
- Response processes (examples given in the Principles include (a) questioning test takers about their response strategies, (b) analyzing examinee response times on computerized assessments, or (c) conducting experimental studies where the response set is manipulated); and
- Consequences of testing (Principles, 2003, p. 5).

The SIOP Principles explain that these five "sources of evidence" are not distinct types of validity, but rather "… each provides information that may be highly relevant to some proposed interpretations of scores, and less relevant, or even irrelevant to others" (p. 5).

There is a great deal of overlap between the Uniform Guidelines and the two professional standards in this area. For example, all three "types" of validation described in the Uniform Guidelines are also contained in the Joint Standards:

- Content validity is similar to the "validation evidence" required in sources 2 and 5 (to a limited degree) of the professional standards,
- Criterion-related validity is similar to the "relationship" evidence required in sources 1 and 5 of the professional standards, and
- Construct validity is similar to the general requirements of sources 1, 3, and 5 of the professional standards.

All three of these documents agree on the importance and relevance of the basic tenets of validation research, including job analysis, test reliability, statistical significance testing, and several other fundamental elements of test validation.

There is, however, a very important distinction that should be noted between the Uniform Guidelines and both sets of professional standards. The very purpose of the Uniform Guidelines is to establish the criteria for weighing "job relatedness and business necessity" evidence in a situation where an employer's testing practice exhibits disparate impact and has come under Title VII scrutiny. The Joint Standards and SIOP Principles are not designed with this sole purpose in mind; nor do they have the statutory or governmental backing to achieve such status. The SIOP Principles have been cited fewer than 20 times, and sometimes with less than favorable results when they are found to be at odds with the Title VII *Griggs* standard that has been adopted by the Uniform Guidelines.

A specific example of this can be seen in *Lanning v. Southeastern Pennsylvania Transportation Authority*[9] where the court stated: "The District Court seems to have derived this standard from the Principles for the Validation and Use of Personnel Selection Procedures ("SIOP Principles") … To the extent that the SIOP Principles are inconsistent with the mission of *Griggs* and the business necessity standard adopted by the Act, they are not instructive." However, in *U.S. v. City of Erie*,[10] the court placed a caveat to this criticism stating that the *Lanning* decision did not "throw out" or otherwise invalidate the SIOP Principles in their entirety when making this statement.

In contrast to the Uniform Guidelines, the Joint Standards and SIOP Principles are designed as widely applicable advisory sources

with a far more exhaustive set of guidelines, whereas the narrowly-tailored Uniform Guidelines are designed to enforce the mission of *Griggs*. Further, the Joint Standards and SIOP Principles cover a much broader scope of testing issues than the Uniform Guidelines. By way of comparison, the Uniform Guidelines are only 27 pages; whereas the Joint Standards and SIOP Principles are 194 and 73 pages respectively, and the terms "disparate impact," "Uniform Guidelines," and "Title VII" are not mentioned a single time in either treatise. Also, while the Joint Standards and SIOP Principles do discuss subgroup differences in testing, they do not discuss the technical determination of disparate impact because it is a legal term of art. This is because the professional standards *were not developed primarily as guidelines for evaluating testing practices in light of Title VII*. The Uniform Guidelines were, however, designed for this express purpose. This is a marked distinction between the Uniform Guidelines and the professional standards and is especially critical when it comes to applying VG as currently framed by the professional standards.

OVERVIEW OF VALIDITY GENERALIZATION

Meta-analysis is a statistical technique used to combine the results of several related research studies to form general theories about relationships between variables (*e.g.*, tests, job performance) across different situations. When meta-analysis is applied to tests and job performance in the personnel testing field, it is referred to as VG. While the specific procedures involved in conducting a VG study may vary, the primary reason for conducting VG studies in an employment setting is to evaluate the effectiveness (*i.e.*, validity) of a specific personnel test or type of test (*e.g.*, cognitive ability, personality) and to describe what the findings mean in a broader sense. To accomplish this, a series of validation studies are combined and then various corrections are made to determine the overall operational validity of the test or type of test, with the intent to ascribe universal effectiveness of the test in different situations and/or locations.

To understand VG, some basic statistical concepts need to be introduced. The most integral element to a VG study is a *validity coefficient*, which is a statistical measure that indicates the strength of a correlation between a certain test and a given job performance criteria (*e.g.*, supervisory ratings). Statistical correlations occur between two variables when high values on one variable are associated with high values on the other variable (and low with low, etc), and range in value between 0 (no correlation) to 1.0 (perfect correlation). In the personnel testing field, correlations that are .35 and higher can be labeled "very beneficial," correlations ranging from .21 to .35 are "likely to be useful," those ranging from .11 - .20 are labeled as "depends on circumstances," and those less than .11 are branded "unlikely to be useful."[11]

Regardless of the size of the validity coefficient (*e.g.*, .15 or .35), it needs to be "statistically significant" beyond a 5% level of chance to be "valid" in a Title VII situation (a requirement also adopted by federal and professional standards), and this determination depends on the sample size involved in the study (with higher validity coefficients required for smaller studies). For example, a coefficient of .20 with a sample of 69 has a corresponding statistical significance probability value (referred to a "p-value") of .0496 (using a one-tail test for significance), which could be argued as defensible under Title VII. However, the same coefficient of .20 with a sample of only 68 has a resulting probability value of .051, which is not statistically significant (because it exceeds the .05 threshold needed for labeling the finding as a "beyond chance occurrence").

Another statistical concept that is important for understanding VG is *statistical power*. In a practical sense, statistical power refers to the *ability of the study to find a statistically significant finding if it exists to be found*. Validity studies that have large sample sizes (*e.g.*, 500 subjects) have high statistical power, and those with small samples have low statistical power. For example, assume that a personnel 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, they could be about 73% confident (*i.e.*, have 73% *power*) of finding such a coefficient (if it was there to be found). With odds of about 3 to 4, the researcher has a "decent shot" at finding validity. With twice the sample size (160 subjects), power would increase to about 94%, which would provide the researcher a near certain ability to find out whether the test was valid at that particular location. And, if the researcher conducts such a study and finds no validity (by obtaining a coefficient that was not statistically significant), they would be comfortable in concluding that validity did not exist at that location, or was sufficiently suppressed by statistical artifacts.

The issue of statistical power frames a problem with personnel researchers that VG attempts to address. By rolling up and combining several independent studies, VG attempts to cast a vision of the "big picture" of what validity for that test might look like over various situations (with some including small samples). Consider the sample VG data in Table 1.

In these sample data, the average sample size was about 134 subjects, yielding about 90% statistical power (on average) for each study to detect a validity coefficient of about .25 in each respective local situation. Notice that 12 of the 22 studies (over half) showed no validity (*i.e.,* had corresponding probability of less than .05 in local settings). Eight (8) studies had correlations that would be considered too low (< .11) to be acceptable in litigation settings. The average validity coefficient across the 22 studies is about .15, which is just barely above the level needed to be statistically significant at the .05 level. However, when these studies are combined into a VG analysis and various corrections are applied, this average validity coefficient increases to between .24 and .48 (based on the type of corrections applied assuming typical reliability estimates and range restriction values). Due to these upward corrections, VG analyses estimate the level of validity that might be found absent the suppressive factors that negatively impact validity studies (*see* Tables 2-4 for some of these factors).

Unfortunately, while these "corrected" VG studies can often offer researchers useful insights into the strength of the relationship between the test and job performance in the studies included in the VG analysis, 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 situationally-specific factors exist in each and every new situation that may drastically impact the validity of a test (see discussion and

TABLE 1
SAMPLE VALIDITY GENERALIZATION RESULTS

Study #	Validity Coefficient	Sample Size	Power (1-tail)	p-value	Valid?
1	0.030	120	87%	0.37	No
2	0.135	130	89%	0.06	No
3	0.180	140	91%	0.02	Yes
4	0.290	150	93%	0.00	Yes
5	0.340	120	87%	0.00	Yes
6	0.180	130	89%	0.02	Yes
7	0.150	140	91%	0.04	Yes
8	0.110	150	93%	0.09	No
9	0.090	120	87%	0.16	No
10	0.126	130	89%	0.08	No
11	0.210	140	91%	0.01	Yes
12	0.390	150	93%	0.00	Yes
13	0.198	120	87%	0.02	Yes
14	0.164	130	89%	0.03	Yes
15	0.109	140	91%	0.10	No
16	0.094	150	93%	0.13	No
17	0.020	120	87%	0.41	No
18	0.114	130	89%	0.10	No
19	0.164	140	91%	0.03	Yes
20	0.070	150	93%	0.20	No
21	0.010	120	87%	0.46	No
22	0.010	130	89%	0.46	No

tables below). In addition, there are a number of issues with typical VG studies that may further limit their relevance and reliability when ascribing test validity into new situations (also see discussion below).

VALIDITY GENERALIZATION, UNIFORM GUIDELINES, JOINT STANDARDS, AND SIOP PRINCIPLES

Validity Generalization and the Uniform Guidelines

The Uniform Guidelines include several provisions for *transporting* validity evidence from either a VG study or a single validity study conducted elsewhere. Validity transportability is based on the notion that acceptable validity evidence for a particular test may exist if that test is "imported" into another situation. This application is based on criterion-related validity identified in one or more situations that is transported to the present situation, coupled with the fact that current conditions parallel past conditions on which acceptable validity evidence for the test exists to properly allow the link to be made. The Uniform Guidelines further require that, when this transportability connection is made between previous studies and the present situation, evidence of test fairness also be provided.

The Uniform Guideline's transportability requirements are not overwhelming and can be easily addressed in practice. First, a criterion-related validity study must be completed to support the relationship between the test and the at-issue criterion. This will typically involve one or more employers and positions that sufficiently address Section 14B (most of the criteria in this section are very basic and overlap with the Joint Standards and SIOP Principles). Second, the "borrowing" employer needs to make a comparison (*e.g.*, using surveys completed by job experts) between the job duties of the positions involved in the original study and the new local location. Strong similarity between the originating positions and the new target position indicates successful transportability. It should be noted that the seminal article on VG in the I-O field agrees that conducting a job analysis in the new local situation is necessary for transporting validity evidence.[12] Third, the transporting user needs to obtain evidence of test fairness. If the originating study included a sufficiently large sample with adequate minority representation, this type of study is fairly routine (in fact, highly detailed recommendations are provided in the SIOP Principles). If such a study is not available from the originating user, the transporting user can rely on the test until such study becomes available.

Section 7 of the Uniform Guidelines also includes the caveat that when transporting validity evidence from other studies, specific attention should be given 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 (*see* Sections 7C and 7D). Fortunately, the Joint Standards, SIOP Principles, and recent VG research have elaborated on just what variables are, in fact, likely to affect (or moderate) validity significantly between the original studies and new local situations (further discussion on this topic is provided below).

Section 15E of the Uniform Guidelines provides additional guidance regarding transporting validity evidence from existing studies into new situations. Like Section 7B, this section includes elements that are likely to be concerns shared by HR and testing professionals 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.

Making sure that the test adopted by the employer is a good "fit" for the target position and insuring that the job performance criteria predicted by the test in the original setting is also relevant in the new setting makes practical business sense (Section 15E1[b]). As a result, insuring that extraneous variables are not operating in a way that negatively impacts test validity (Section 15E1[c]) is often a key component evaluated in VG analyses. Finally, considering how the test is used (*e.g.*,

ranked, banded, or used with a cutoff) also has significant impact on the utility and diversity outcomes of the employer (Section 15E1[d]).

Rather than being an action taken solely to justify disparate impact, addressing the requirements of the Uniform Guidelines when conducting validation research can actually help employers insure their testing practices screen in high-quality applicants. In fact, all four sections of 15E1(a-d) are *employer-relevant* objectives—they are not just "government requirements" surrounding EEO compliance.

Validity Generalization and the Joint Standards

The Joint Standards include a one-page preamble (p. 15) and two standards (along with comments) surrounding VG. While the complex issue of VG is given only a 2-page treatment in the entire 194-page book, the discussion is compact and to the point. The two standards dealing with the subject (Standard 1.20 and 1.21) advise test users and test publishers regarding the conditions under which validity evidence can be inferred into a new situation based on evidence from other studies. Note that these two standards are specifically tailored around the use of modern VG and meta-analysis techniques (whereas the Uniform Guidelines cover some of these same issues, but more generally).

> *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 SIOP Principles provide a more extensive discussion on VG than the Joint Standards. The entire discussion relevant to VG provided by the SIOP Principles is contained within pages 8-10 and 27-30. Under the section headed "Generalizing Validity Evidence," the SIOP Principles outline three strategies that are neither mutually exclusive nor exhaustive: (a) transportability, (b) synthetic validity/job component validity, and (c) meta-analytic validity generalization (p. 27).

Compared to the previous two types of generalized validity evidence, the SIOP Principles provide the most detailed requirements regarding the use of meta-analysis for generalizing validity evidence. Some of the essential elements are listed below:

- The importance of applying professional judgment in interpreting and applying the results of meta-analytic research.
- Consideration of the meta-analytic methods used, the underlying assumptions, and statistical artifacts that may influence the results.
- Concern and evaluation of potential moderators (situational factors that affect validity findings in specific settings).
- Consulting the relevant literature to ensure that the meta-analytic strategies used are sound and have been properly applied.
- Consideration of study characteristics that may possibly impact the study results.
- Awareness of continuing research and critiques that may provide further refinement

of the techniques as well as a broader range of test-criterion relationships to which meta-analysis has been applied.

- Evaluating the similarity of the constructs measured by the tests included in the meta-analysis and those in the local situation.
- Evaluating 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.

The various requirements presented by the Uniform Guidelines, Joint Standards, and SIOP Principles can be mapped back to two major areas. The first is the *internal quality* of the VG study itself. This includes factors such as study design features, similarity of the tests, jobs, and job performance criteria used in the study, and the number of studies included. The second pertains to factors regarding the *comparability between the VG study and the new local situation*.

Beyond these two primary areas, there are additional considerations necessary that pertain to the *assumptions* that must be made when "importing" validity evidence into a new local situation without conducting a local study. This is important because when courts evaluate the validity of a test that is potentially discriminating against a certain group (which occurs with disparate impact absent validity), *they typically do not like to rely on assumptions*.[13] Rather, they have consistently required *situational-specific evidence regarding the job relatedness of a particular test and its relationship with accurately and specifically defined job requirements*. Such situationally-specific evidence does not need to take the form of a local criterion-related validity study, but can be accomplished by using a link-up study as described in Section 7B and/or 15E of the Uniform Guidelines, or other source of validity evidence (*e.g.*, content validity).

As discussed below, the U.S. Supreme Court and appellate courts have indicated a clear and consistent disfavor towards employers using tests to hire "smart and educated" employees in the abstract. Rather, following the requirements of federal civil rights law, they are required to *demonstrate* (the term used in the 1991 Civil Rights Act) how the specific test has a *manifest* (the term used in *Griggs*) relationship to the clearly-defined (and researched) requirements of the job.

VG AND THE COURTS

VG and Griggs v. Duke Power

In the *Griggs* case, the Duke Power Company required all employees who desired employment in any division outside the general labor department to obtain satisfactory scores on a test which purported to measure general intelligence, a mechanical comprehension test, and possess a high school education. None of these requirements were designed to measure the ability to directly perform the job duties of a *particular job or category of jobs*. The court ruled that the requirements 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." Rather, a vice president of the company testified that the requirements were instituted on the company's judgment that they ". . . generally would improve the overall quality of the workforce." 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."

Knowing that these tests had high levels of disparate impact against minorities, Duke Power 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. The Supreme Court, however, in its 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.*

VG and EEOC v. Atlas Paper

The Sixth Circuit, in *EEOC v. Atlas Paper,*[14] re-affirmed specific versus generic validity requirements when the court ruled that VG, as a matter of law, could not be used to justify testing practices that had disparate impact. In *Atlas*, the Sixth Circuit completely rejected the use of VG to justify a test purporting to measure general intelligence (the Wonderlic), which had disparate impact when used for screening clerical employees. Without conducting a local validity study, an 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.

Note that the requirement mandated above is exactly what is currently required by the Uniform Guidelines for transporting validity evidence into a new situation (Section 7B). Both simply require that a job comparability study be done between the job in the original validation study and the new local situation.

The Sixth Circuit decision in *Atlas* offered even a more direct 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.

The judge issued a factual conclusion based upon the applicability of the *Albemarle* findings regarding the *situational specific* validity requirements and concluded:

> 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.

The *Atlas* case demonstrates the likely outcome of what will happen to employers if they take unnecessary risks by relying solely on VG evidence when their testing practices

exhibit disparate impact. In fact, some authors have stated that even if the Uniform Guidelines were changed to adopt a more open stance toward VG that 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."[15] Conducting Uniform Guidelines-style "transportability" studies (to address Section 7B) offers much higher levels of defensibility (conducting a local validation study perhaps offers even higher levels of defensibility).

The Fifth Circuit has accepted such validation evidence (based on job comparability evidence, as required by the Uniform Guidelines) in at least two cases: *Cormier v. PPG Industries* (1983) and *Bernard v. Gulf Oil Corporation* (1989).[16] However, because these cases predate the 1991 Civil Rights Act and the latter was tried under the less-stringent *Wards Cove* standards (which only required employers to provide a "business justification" for the test causing disparate impact), they likely have little applicability because the 1991 CRA reinstated the more stringent *Griggs* standard, which requires employers to demonstrate that the test is job related *for the position in question* and consistent with business necessity.[17] Rather than allowing a "generalized inference" of validity for a test, the 1991 CRA requires a demonstration of job relatedness for the specific position in question, not for an entire, sweeping category of employment tests (*e.g.*, those measuring cognitive abilities).

TABLE 2
SITUATIONAL FACTORS PERTAINING TO THE APPLICANT POOL THAT CAN INFLUENCE THE RESULTS OF A LOCAL VALIDATION STUDY

Factor #	Factor	Possible Impact on a Local Validation Study
1	Sample size	Sample size is perhaps one of the most influential factors in a statistical study (larger samples have a higher likelihood of finding a significant test-criterion relationship if it in fact exists).
2	Percentage of applicants who are qualified	The qualification level of the applicant pool as a whole can expand or restrict the effective utility of the test.
3	Competitive environment	The competitive nature of the position can impact the range and distribution of applicant scores.
4	Other tests used in the hiring process	This affects the statistical power of the test because tests that are used before and after the target test will restrict the qualification levels of the applicants.

TABLE 3
SITUATIONAL FACTORS PERTAINING TO THE TEST THAT CAN INFLUENCE THE RESULTS OF A LOCAL VALIDATION STUDY

Factor #	Factor	Possible Impact on a Local Validation Study
1	Test content	While different tests may measure similar constructs, their underlying content and statistical qualities may differ substantially.
2	Test administration conditions (proctoring, time limits, etc.)	Test results are highly susceptible to external influences during the testing situation.
3	Test administration modality (*e.g.*, written vs. online)	The mode (or method) in which a test is given can have an impact on applicant scores.
4	Test use (ranking, banding, cutoffs)	The way in which test scores are used defines the test distribution characteristics, and the extent to which test scores can relate to other variables (*e.g.*, job performance criteria).
5	Test reliability (internal consistency)	The reliability of a test sets the maximum validity threshold of a test-job performance relationship.
6	Test bias (*e.g.*, culturally-loaded content)	Test bias can impact the level of validity obtained in a local study by introducing error into the process.

TABLE 4
SITUATIONAL FACTORS PERTAINING TO THE JOB THAT CAN INFLUENCE THE RESULTS OF A LOCAL VALIDATION STUDY

Factor #	Factor	Possible Impact on a Local Validation Study
1	Job content comparability	Even jobs that share similar titles may, in actuality, perform vastly different job duties (Uniform Guidelines, 7B-1). Even jobs that have identical duties can spend different amounts of time on each and can vary in the importance level of similar duties (Uniform Guidelines, Q&A #51).
2	Job performance criteria	The comparability between the job performance criteria used in the original study and those used in the local study (*e.g.*, objective, subjective, production, sales, turnover, etc.) can have a substantial impact on validity level that would likely be found in the new local situation.
3	Reliability of job performance criteria	The reliability level of job performance criteria sets a maximum validity threshold. A wide variety of factors can impact rater reliability.[18]
4	Rating bias on job performance criteria	Rating bias can have a substantial impact on test validity in a new local situation.[19]
5	Range restriction on job performance criteria	Range restriction on the criteria occurs when less than 100% of the applicants tested and hired are available to receive job performance ratings (this has a suppressive effect on test validity).
6	Level of supervision/autonomy	The level of supervision or autonomy in the jobs in the original validation study and the new situation can possibly have an impact on test validity.[20]
7	Level/quality of training/coaching provided	Employees can often "rise and fall" in organizations based on the level of training and coaching they receive, which obviously can have an impact on job performance ratings.[21]
8	Organizational- and unit-level demands and constraints	These factors can have a wide degree of impact on both individual-level job performance and job performance ratings.
9	Job Satisfaction	Job satisfaction can have a significant influence on job performance.[22]
10	Management and leadership styles and role clarity	Interactions between leaders and members are strongly related to supervisory ratings of performance.[23]
11	Reward structures and processes	Employee incentive systems can vary greatly between various organizations/positions and can impact job performance.[24]
12	Organizational citizenship, morale, and commitment of the general workforce	The effects of organizational citizenship behaviors, morale, and perceived organizational support have a significant impact on individual- and organizational-level performance.[25]
13	Organizational culture, norms, beliefs, values, expectations surrounding loyalty and conformity	These factors can have a wide impact on both individual- and team-level performance.[26]
14	Organizational socialization strategies for new employees	How new employees are introduced and acculturated into the workforce can have an impact on both employee performance and job ratings.[27]
15	Formal and informal communication (style, levels, and networks)	Communication between supervisors, employees, and work units can have a significant impact on employee performance and job ratings.[28]
16	Centralization and formalization of decision-making	An organization's decision-making characteristics and structure can play a major role in employee performance and job ratings.[29]
17	Organization size	Organization size can impact a wide array of factors that can have an impact on test validity.
18	Physical environment (lighting, heating, privacy)	These factors have been studied (sometimes controversially) for decades in I-O psychology, and have mixed results. Nonetheless, these are some factors that can obviously have an impact on employee performance and job ratings.[30]

The basic requirement that tests must be proven job related and consistent with business necessity 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). Reaffirmation by the passage of the 1991 Civil Rights Act—which overturned the U.S. Supreme Court on this specific issue—indicates that the requirement is likely to endure subsequent challenges.

ASSUMPTIONS MADE WHEN USING VG TO IMPORT VALIDITY EVIDENCE INTO A NEW LOCAL SITUATION

Consider a hypothetical VG study that combines 22 unique validation studies for a test used for similar positions at different employers. Then assume that the test exhibited an average validity coefficient of .25 across these 22 studies. Now consider a new local situation (called "Site 23") where the same test is being used for a position that is similar to those included in the other 22 studies but is challenged in a federal Title VII case because the test exhibits disparate impact. What factors could possibly influence the level of validity that would be found at Site 23 if a local study was conducted? What assurances do we have that a validity coefficient of .25 would be found at Site 23? Should validity be automatically assumed at Site 23 if the test was found valid overall at the previous 22 sites involving similar jobs?

Answering this question absent a local validation study is certainly of utmost interest to a federal court, since prior to using the VG study to infer that a *sufficient level of validity* (adequate to justify the disparate impact) exists at Site 23, numerous assumptions must be made. Some of the obvious assumptions include the similarity between jobs, consistent and reliable administration of the test, the percentage of applicants who possess the qualifications for the job, and the cutoff score used for the test. Each of these factors would have a major impact on the level of validity that could be found at Site 23.

Additional factors can also impact the validity of the test at Site 23: different time limits, an inconsistent proctor, or different content than the test used in the original 22 studies. If Site 23 requires certain educational qualifications of applicants, or tests applicants using other measures before they take the at-issue test, this could also possibly impact the level of validity found. VG methods can attempt to isolate and control for statistical nuisances that can act to suppress or lower validity; however, these are sometimes difficult to sell to the judge, especially when high levels of disparate impact have already been observed in the local setting. The judge is forced to rely on "estimated and generalized" validity to justify empirically-demonstrated disparate impact.

The situational factors that can influence the strength of the test-job performance relationship in the local situation can be broken down into three major categories: those related to the applicant pool, the test itself, and characteristics of the job. Without conducting a local study and relying wholly on VG results to import validity into a new local situation, employers are placed in the uncomfortable position of assuming that these factors would not have impacted the validity from transferring over to the new situation. Tables 2-4 below outline 28 of the possible hindrances to finding validity in new local situations.

Theoretically, each of the 28 factors listed in Tables 2-4 can impact the validity of a test in a new situation. However, this article does not propose that each of these factors *will* have an influence on the outcome of validity studies, just that they *can*…and if one relies wholly on VG studies of validity evidence in a Title VII situation (without conducting some type of a local study), one will never know the impact these factors will have on the validity in the new local situation. In fact, these factors are so critical that it is entirely possible to take a test from a job/employer that had high validity based on some job performance criteria, such as supervisory ratings, and get *completely different validity results* in a new local setting through the impact of *any one of these factors*.

There is overlap between the factors that impact the internal quality of a VG study and the factors pertaining to the similarity between the VG study and the local situation (outlined in the professional standards), and the complete list of factors that will ultimately determine the level of validity found in a new local situation (Tables 2-4). Because some of these factors are more widely known to affect validity studies than others, they have been incorporated into the federal and professional standards (*e.g.*, factors 1 and 4 in Table 2; factors 1, 4, and 5 in Table 3; and factors 1-3 and 5 in Table 4). Further, several factors listed in Table 4 could be classified as "moderators," which are consistently mentioned in both sets of professional standards (*e.g.*, Standards 1.20 and 1.21 in the Joint Standards and pages 9, 28, and 29 of the SIOP Principles).

Although modern VG analyses utilize statistical assessments to evaluate whether (and the extent to which) situational factors may have been present that worked to limit validity generalization inferences into new local situations, there is no way to determine *if* and *to what extent* the factors in Tables 2-4 would inhibit validity from being found if an actual study was conducted in the *new local situation*. No matter how compelling and clear the evidence may be from a VG study, it is possible for any one of the factors discussed above to completely undermine the generalization of validation evidence from other studies to a new local situation.

Given this fact, how safe can an employer be when relying solely on VG to refute a class claim of disparate impact discrimination? It is because of this potential shortfall that the Uniform Guidelines include Q&A #43 to clarify that tests can in fact be valid predictors of performance on a job in a certain location and yet be invalid for predicting success on a different job or the same job in a different location. This is also why the Uniform Guidelines require that specific standards be satisfied before a user may rely upon findings of validity generated in another situation (*i.e.*, through a 7B transportability study). It should also be noted that while the above factors can work to suppress validity in a new local situation, sometimes the reason that validity evidence is not found in a local study is simply because the correlation is just not there to be found because the test has no relationship to job performance.

DESPITE THE INCREASE IN POPULARITY AMONG TEST PUBLISHERS AND HR/HIRING STAFF AT CORPORATIONS, RELYING ENTIRELY ON VG TO DEFEND AGAINST TITLE VII DISPARATE IMPACT SUITS WILL LIKELY LEAD TO DISAPPOINTING OUTCOMES BECAUSE THE COURTS HAVE GENERALLY REQUIRED EMPLOYERS DEMONSTRATE *LOCAL AND SPECIFIC* VALIDATION EVIDENCE WHERE THERE IS *LOCAL AND SPECIFIC* EVIDENCE OF DISPARATE IMPACT.

In addition to these limitations, it will be especially difficult to argue in a litigation setting exactly what the *actual correlation value* "would have been" had a local study been conducted. VG techniques include statistical tools that are designed to provide confidence boundaries around what the "true, population" validity coefficient might in fact be. Ultimately however, it is still an *assumed value* rather than an *actual value* that one can take to the witness chair, and anything that appears as a "guess" is less likely to be accepted by a conservative court (opposed to "actual proven scientific validation evidence" that can be derived from a local study). This may be especially true of judges who have less than

adequate statistical training and are sometimes speculative about the statistical sciences (which is arguably most judges).

Employers that elect to employ a VG-only defense in court will be in for an uphill climb to effectively argue before a skeptical court that none of these factors *did* (during the past use of the test), *would* (in the present), or *will* (to justify use of the test in the future) come into play in an employer's local situation, and that the validity results observed in the VG study will just "fit right into" the employer's unique situation—complete with a validity coefficient large enough to justify the degree of disparate impact exhibited by the test at the local situation. In Title VII litigation, it is the employer's burden to prove these moderating factors did not (or would not) hinder the test from being valid in the local situation (and hence it will likely be the plaintiff's point of attack to show how they did or would).

Some of the factors listed in Tables 2-4 are statistical characteristics that serve to suppress the true validity of the test (*e.g.*, range restriction). Other factors, such as the sample size involved in a statistical study, lower the statistical power of a correlational comparison and decrease the study's ability to uncover the actual correlation value that may be present. Some factors can be corrected for using statistical formulas. Other factors, however, are not of this nature and constitute real characteristics that may lower the actual validity in the study (*e.g.*, test reliability, the percentage of applicants who are qualified, etc.). To a certain extent, however, all of these factors are relevant in Title VII matters because ultimately a court will want to know the level of validity in a local situation, with all of these factors taken into consideration (at least in cases where a criterion-related validity strategy is used). Experts can argue about which statistical corrections should be used given the particular circumstances at hand, but there is no argument quite as strong as a statistically significant validity coefficient that is based on the employer's specific local situation. In Title VII situations, a court will likely desire to know the bottom-line validity irrespective of *if and how* each one of these factors came into play.

Experts can also argue about the extent to which the factors discussed above can or will actually play into whether a test would actually be found valid in a new local situation. This will not change, however, the reality that for a test to show validity in a new situation requires that *each one* of these factors not substantially hinder the relationship between test and job performance criteria. However, the fact that each of these factors could have an impact in the local setting is not even the substantive issue in a Title VII situation. The substantive issue is that they *could* have an impact and, despite this, the employer chose to rely solely on outside validity evidence. With so much at stake, prudent employers may not want to make such a leap of faith.

It may be possible to find validation studies where each of the above moderating factors (*see* Tables 2-4) manifested in a way that inhibited test validity. But beyond the research, experienced employees can likely identify with each of these factors and how they have impacted test and/or work performance (on individual and organizational levels). Along these lines, there are two observations about the moderating factors outlined in Tables 2-4.

First, if researchers were hypothetically allowed to arrange the conditions of a validation study to maximize their chances of finding the highest correlation, it is extremely likely that every single one of the factors outlined in Tables 2-4 would be manipulated and controlled. For example, assume a researcher wanted to conduct a validation study on a personality test that measured conscientiousness (a construct that has been shown to predict job performance in a wide variety of different positions and organizational types and levels). Assume for this discussion that VG studies conducted on the specific test of interest resulted in a validity coefficient of .30 for positions similar to the target situation. If a researcher was hypothetically allowed to go through every factor on Tables 2-4 and dictate the conditions for each (*e.g.*, select a large sample, a low percentage of qualified applicants, and a low test cutoff to allow wide variance on the criterion measure, test content perfectly like the original in the study, perfect test administration

conditions, job criteria that was both job and test construct relevant, no rating bias, supervisory ratings offered by at least two trained raters, supervision levels that allowed individual abilities wide variance in job performance), rather than leaving each one to chance, experienced practitioners would manipulate each to maximize the chances of finding high validity. This is because each of these factors can operate to maximize or suppress validity. In fact, experts in personality testing might even want to design a local situation differently for different types of tests.

Second, the Uniform Guidelines, Joint Standards, and SIOP Principles have already established that many of these factors *do in fact constitute key considerations* both within a VG-style meta-analysis and when seeking to externalize from them (*i.e.*, infer validity into a new situation). Some practitioners argue that the Uniform Guidelines are extremely outdated because they are based on the situational specificity doctrine (*i.e.*, that test validity varies specifically from situation to situation).[31] The Uniform Guidelines, however, offer less restrictive guidelines than the professional standards for transporting validity from other validation studies into new local situations (*i.e.*, the three requirements in Section 7B, versus the two standards in the Joint Standards and several pages in the SIOP Principles).

The Uniform Guidelines allow for validity to be imported into a new local situation if the employer can simply demonstrate that the at-issue position is highly similar to the position in the original study. By following the modest and non-burdensome requirements in the Uniform Guidelines (Section 7B) for transporting validity evidence, an employer has at least some level of assurance that many of these key factors (*e.g.*, job similarity) have been adequately addressed.

THE ELEMENTS OF A CRITERION-RELATED VALIDITY STUDY THAT ARE TYPICALLY EVALUATED IN TITLE VII SITUATIONS

When the courts evaluate criterion-related validity evidence, which is the type of evidence of validity that can be included in statistical VG studies, four basic elements are typically brought under inspection: (1) statistical significance, (2) practical significance, (3) the type and relevance of the job criteria, and (4) evidence available to support the specific use of the testing practice. If any of these elements are missing or do not meet certain standards, courts often infer discrimination because disparate impact was not justified by validity evidence. Each of these elements is discussed in more detail below.

Statistical significance. The courts, Uniform Guidelines, and professional standards are in agreement when it comes to the issue of statistical significance thresholds and criterion-related validity. Indeed, the .05 threshold is used on both sides of disparate impact litigation: for determining statistically significant disparate impact (using hypergeometric probability distributions for testing cases) as well as determining the statistical significance of the correlation coefficient obtained in the validation study.

Practical significance. Just like statistical significance, the concept of practical significance has also been applied to both the disparate impact and validity aspects of Title VII cases. As it relates to disparate impact, the courts have sometimes evaluated the practical significance or "stability" and effect size of the disparate impact.[32] This is typically accomplished by evaluating the statistically significant findings when just a couple of applicants are hypothetically changed from failing to passing on the selection procedure that exhibited the disparate impact. If this hypothetical process changes the statistically significant finding from "significant" ($<.05$) to "non-significant" ($>.05$), the finding is not practically significant.

In the realm of criterion-related validity studies, practical significance relates to the strength of the validity coefficient (*i.e.*, its raw value and actual 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 .15 explains only 2.3% of the criterion variance, whereas coefficients of .25 and

.35 explain 6.3% and 12.3% respectively. Some cases have included lengthy deliberations about these "squared coefficient" values to argue the extent to which the test validity is practically significant. A few examples are provided below.

- *Dickerson v. U. S. Steel Corporation*[33]: A validity study was inadequate where the correlation level was less than .30, the disparate impact on minorities from the use of the selection procedure was severe, and the employer did not present any evidence regarding its evaluation of alternative selection procedures. 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*[34]: Judge Pointer rejected statistically significant correlations of .21, because they were too small to be meaningful.
- *EEOC v. Atlas Paper*[35]: 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*, Id; *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*[36]: The court debated the level of the validity coefficients extensively: "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) are in concert regarding the importance of taking the strength of the validity coefficient into practical consideration.

Type and relevance of the job criteria. There are many cases that have deliberated the type and relevance of the job criteria included as part of a validity study, including the cases cited above. The Uniform Guidelines (15B6) and SIOP Principles (p. 16) also include discussion on this topic.

Considering the validity coefficient and the specific use of the testing practice. Some cases have set minimum thresholds for validity coefficients that are necessary to justify the particular *use* of a test (*e.g.*, ranking versus using a pass/fail cutoff). Conceptually speaking, tests that have high levels of reliability (*i.e.*, accuracy in defining true ability levels of applicants) and have high validity can be used at a higher degree of specificity than tests that do not have such characteristics.[37] When employers have used tests as ranking devices, they are typically subject to a more stringent validity standard than when pass/fail cutoffs are used. The cases below placed minimum thresholds

on the validity coefficient necessary for strict rank ordering on a test:

- *Brunet v. City of Columbus*[38]: 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*[39]: 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*[40]: 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*[41]: 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 the level of validity into consideration when considering how to *use* a test in a selection process. The reason that test usage is such a critical consideration is because, ultimately, validity has to do with the *interpretation of individual scores*. Tests, per se, are not necessarily valid; rather, specific scores may or may not be valid given how closely they are aligned with the true needs of the job, and the level to which they are aligned. For example, an English proficiency test may be valid for both the positions of an office manager and a proofreader at a newspaper agency; however, the test will likely have higher relevancy (and require a higher passing score) for the proofreader position.

In the event of a Title VII lawsuit, employers who have relied *solely* on a VG study to infer evidence of validity would not have the information necessary to show the court that these four critical factors have been properly supported. In fact, employers electing to rely solely on VG evidence in Title VII situations 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). As a result, 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.

By relying solely on a VG study, there is no way for the employer to determine whether a validity coefficient would be statistically significant in its local situation because no local validity coefficients were ever calculated. While VG studies can generate estimated population validity coefficients (with various types of corrections), it is not possible to determine if such validity coefficient would be obtained in the local situation, and (more importantly), whether it would exceed the court-required level needed for statistical significance ($<.05$). Even if one considers the population validity coefficient calculated from a VG study at face value (*e.g.*, $r = .25$), calculating the statistical significance level requires also knowing the sample size in-

cluded in the study, which is another unknown unless a local study is conducted.

Without knowing what the actual validity coefficient would be in a local situation, it is also not possible to evaluate its practical significance in the local job context. While contemporary VG techniques estimation techniques for speculating the levels of validity that might be attained in studies outside those included in the VG study, this is also not helpful in litigation settings because one still must "guess" at what the actual validity *would have been* in the actual situation. In many circumstances, judges will be more likely to make determinations whether the test would "survive scrutiny" in light of the situational factors of the case (*e.g.*, the level of disparate impact, relevancy of the criterion, etc.) only after he/she is in possession of the *actual validity coefficient*.

Irrespective of not knowing the *level of correlation* in a particular situation, judges are likely to be further reluctant to support a test when they don't know the *type* and *relevance* of the job criteria. Oftentimes VG studies include a wide mix of various job criteria predicted by a test and, without conducting a local study, there is no way to know if the test would in fact be correlated to criteria sufficiently important to the local job. The Uniform Guidelines deliver a specific caution about this very issue: "Sole reliance upon a single selection instrument which is related to only one of many job duties or aspects of job performance will also be subject to close review" (Section 14B6). If employers already have an uphill battle proving the relevance of job criteria that are significantly correlated with a test in their local situation, relying on one step "further removed" from the actual situation (by using VG) may leave employers even more challenged.

Lastly regarding the "specific use" factor, judges will be hard-pressed to support the specific use an employer has chosen for the test being challenged. Again, the Uniform Guidelines advise that the use of a test should be evaluated to insure its appropriateness for operational use, including the establishment of cutoff scores or rank ordering (Section 15B6). The court cases outlined above represent a small portion of the litigation over exactly how tests should be used in a particular situation. Absent validity results from a local study, this is again one less factor a judge will be able to use in the employer's favor.

It may be unlikely that judges will justify *disproportionate passing rates* (*i.e.*, disparate impact) on a test based on *speculated validity* by assuming that these factors would not play a part in lowering the level of validity that would be found in the new local situation. In light of the high stakes coupled with significant (and avoidable) risks, employers would be much better insulated in Title VII lawsuits where at least some local validity evidence is amassed.

RECOMMENDATIONS

Over the past 30 years, VG and its related tools, techniques, and research results have contributed greatly to the overall effectiveness and utility of a wide range of selection procedures. It has also spawned years of debate[42] that have led to great progress in many research areas. Perhaps the most effective and least controversial application of VG is to identify the types of tests that have been previously shown to be the most effective for particular job classifications (and for specific types of criteria). After such tests have been identified, they can be adopted and used either under a transportability model (under 7B of the Uniform Guidelines) or a local study can be conducted (if technically feasible). These steps will especially be important when the tests have disparate impact. And, when they do have disparate impact, the use of VG—just like any other source of validity evidence (*e.g.*, content validity)—should follow some conservative guidelines when being used to prove job relatedness in Title VII situations (*e.g.*, government enforcement activities, private plaintiff litigation, civil service hearings, etc.). Guidelines for this are suggested below:

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. 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. 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).

IN FACT, ONE MAJOR STUDY... COMPARED THE VALIDITY DIFFERENCES BETWEEN WRITTEN TESTS BASED UPON *JOB SPECIFICITY*. THE RESULTS SHOWED THAT TESTS HIGHLY SPECIFIC TO JOB REQUIREMENTS DEMONSTRATED MUCH HIGHER VALIDITY (ABOUT DOUBLE THAT OF "GENERIC" TESTS), AND THE RESULTS WERE CONSISTENT WITH BOTH ON-THE-JOB AND TRAINING PERFORMANCE.

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 (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 disparate 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 Act Section 2000e-2[k][1][A][ii] and Section 3B of the Uniform Guidelines).

CONCLUSION

When choosing between relying on VG evidence to import validity of generic ability tests or conducting local validation studies and/or developing job- and employer-specific tests based on *researched* job requirements (job analyses, test plans, etc.), the latter option enjoys several major benefits. First, using customized tests is more likely to result in higher validity. In fact, one major study[43] (including 363,528 persons and 502 validation studies) compared the validity differences between written tests based upon *job specificity*. The results showed that tests highly specific to job requirements demonstrated much higher validity (about double that of "generic" tests), and the results were consistent with both on-the-job and training performance.

Another benefit is that custom tests provide a stronger defense if the employer is challenged. Judges and juries (who are almost always novices in testing and statistics) prefer to see, touch, taste, and feel how the job is *rationally and empirically* related to the test. As pointed out above, only local validation studies can provide local and specific evidence regarding the statistical and practical significance of the test, the type and relevance of the job criteria, and evidence to support the specific use of the testing practice. When employers elect to rely solely on VG studies, they cannot *really know* that the test is valid for their job or setting.

Tests validated at local situations provide higher assurance of utility. Local validity coefficients provide assurance that the test is *actually working* for the specific criteria of interest rather than borrowing validities from studies conducted for similar jobs, tests, and criteria. Likewise, local validation studies provide specific information on how to *weight* (combine) and *use* (rank, band, pass/fail cutoffs) various selection procedures. This is because local validation studies utilizing either content or criterion validity strategies result in narrowly-defining the job requirements and typically evaluate the relative importance of various qualifications necessary for the job. ▲

ENDNOTES

* Biddle Consulting Group is an EEO litigation support firm that specializes in assisting employers in Title VII compliance and litigation, including areas such as affirmative action planning, EEO compliance, and test development and validation. BCG also develops and distributes pre-employment tests for the administrative and 911 dispatcher industries. FPSI develops and validates written, verbal and physical ability tests for the police and fire industries.

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

2 Griggs v. Duke Power Co, 401 U.S. 424 (1971).

3 Wards Cove Packing Co., Inc. v. Atonio, 109 S.Ct. 2115 (1989).

4 Uniform Guidelines, Questions & Answers #1.

5 Uniform Guidelines, Questions & Answers #2.

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

7 Based on search conducted on Westlaw in May, 2006.

8 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 (p. 2).

9 Lanning v. Southeastern Pennsylvania Transportation Authority, 181 F.3d 478, fn 20 (3rd Cir. 1999), 80 Fair Emp. Prac. Cas. (BNA) 221, 76 Emp. Prac. Dec. (CCH) ¶46,160.

10 U.S. v. City of Erie, 411 F.Supp.2d 524, fn 18 (W.D. Pa. 2005).

11 U.S. Department of Labor: Employment and Training Administration (2000), *Testing and Assessment: An Employer's Guide to Good Practices*. Washington DC: Department of Labor Employment and Training Administration (p. 3-10).

12 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.

13 Berger, M.A. (2000). *The Supreme Court's trilogy on the admissibility of expert testimony*. In D. Rubenfeld (Ed.), Reference Manual on Scientific Evidence (2nd ed., pp. 10-38). Federal Judicial Center.

14 EEOC v. Atlas Paper, 868 F.2d 1487 (6th Cir. 1989), *cert. denied*, 493 U.S. 814.

15 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.

16 Cormier v. PPG Industries, 519 F.Supp. 211 (W.D. La. 1981), *aff'd* 702 F.2d 567 (5th Cir. 1983); Bernard v. Gulf Oil Corporation, 890 F.2d 735 (5th Cir. 1989).

17 Gutman, A. (2005). *Disparate impact: judicial, regulatory, and statutory authority*. In F.J. Landy (Ed.), Employment Discrimination Litigation: Behavioral, Quantitative, and Legal Perspectives (pp. 20-46). San Francisco: Jossey Bass.

18 Landy, F. J. & Farr, J. L. (1983). The Measurement of Work Performance: Method, Theory, and Applications. San Diego, CA: Academic Press (pp. 120-122).

19 Stauffer, J. M. & Buckley, M. R. (May 2005). *The existence and nature of racial bias in supervisory ratings*. Journal of Applied Psychology, *90* (3), 586-591; Landy & Farr, *supra* n. 18, (pp. 120-122).

20 Morgeson, F. P., Delaney-Klinger, K., Hemingway, M. A. (March, 2005). *The importance of job autonomy, cognitive ability, and job-related skill for predicting role breadth and job performance*. Journal of Applied Psychology, *90* (2), 399-406; Barrick, M. R. & Mount, M.K. (1993). *Autonomy as a moderator of the relationships between the Big Five*

personality dimensions and job performance. Journal of Applied Psychology, *78* (1), 111-118.

[21] For example, types and amounts of feedback have been shown to increase performance by 8% - 26% (Landy & Farr, *supra*, n. 18, 264-265). See also Ellinger, A. D., Ellinger, A.E., Keller, S.B. (2003). *Supervisory coaching behavior, employee satisfaction, and warehouse employee performance: A dyadic perspective in the distribution industry.* Human Resource Development Quarterly, *14* (4), 435-458.

[22] Judge, T. A. & Ferris, G. R (1993). *Social context of performance evaluation decisions.* Academy of Management Journal, *36*, 80-105; Schleicher, D. J., Watt, J. D. Greguras, G.J. (2004). *Reexamining the job satisfaction-performance relationship: The complexity of attitudes.* Journal of Applied Psychology, *89* (1), 165-177.

[23] Gerstner, C.R., Day, D.V. (1997). *Meta-Analytic review of leader-member exchange theory: Correlates and construct issues.* Journal of Applied Psychology, 82, 827-844; Podsakoff, P. MacKenzie, S., Bommer, W. (1996). *Meta-analysis of the relationships between Kerr and Jermier's substitutes for leadership and employee job attitudes, role perceptions, and performance.* Journal of Applied Psychology, 81 (4), 380-399.

[24] Jenkins, D.G., Mitra, A., Gupta, N., & Shaw, J.D. (1998) *Are financial incentives related to performance? A meta-analytic review of empirical research.* Journal of Applied Psychology, 83, *777-787*; Pritchard, R., Jones, S., Roth, P, Stuebing, K, & Ekeberg, S. (1988). *Effects of group feedback, goal setting, and incentives on organizational productivity.* Journal of Applied Psychology, *73* (2), 337-358.

[25] Allen, T.D. & Rush, M.C. (1998). *The effects of organizational citizenship behavior on performance judgments: a field study and a laboratory experiment.* Journal of Applied Psychology, 83, 247-260; Rhoades, L. & Eisenberger, R. (2002*). Perceived organizational support: A review of the literature.* Journal of Applied Psychology, *87*, 698-714.

[26] Meglino, B. M., Ravlin, E. C. Adkins, C. L. (1989). *A field test of the value congruence process and its relationship to individual outcomes.* Journal of Applied Psychology, *74* (3), 424-432; Brown, S.P. & Leigh T.W. (1996). *A new look at psychological climate and its relationship to job involvement, effort, and performance.* Journal of Applied Psychology, 81, 358-368; Heck, R.H. (1995). *Organizational and professional socialization: Its impact on the performance of new administrators.* Urban Review, *27*, 31-49.

[27] Steel, R.P., Shane, G.S. & Kennedy, K.A. (1990). *Effects of social-system factors on absenteeism, turnover, and job performance.* Journal of Business and Psychology, 4, 423-430; Chao, G., O'Leary-Kelly, A., Wolf, S., Klein, H., Gardner, P. (1994). *Organizational socialization: its content and consequences.* Journal of Applied Psychology, *79* (5), *730-743*.

[28] Abraham, S. (1996). *Effects of leader's communication style and participative goal setting on performance and attitudes.* Human Performance, 9 (1), 51; Alexander, E.R., Helms, M.M. & Wilkins, R.D. (1989). *The relationship between supervisory communication and subordinate performance and satisfaction among professionals.* Public Personnel Management, 18, 415-429; Kacmar, K.M., Witt, L.A., Zivnuska, S. & Gully, S.M. (2003). *The interactive effect of leader-member exchange and communication frequency on performance ratings.* Journal of Applied Psychology, 88, *764-772*.

[29] Judge, T. A. Thoresen, C.J., Bono, J. E., Patton, G.K. (2001). *The job satisfaction-job performance relationship: A qualitative and quantitative review.* Psychological Bulletin, *127* (3), 376-407; Steel, R.P., Shane, G.S. & Kennedy, K.A. (1990). *Effects of social-system factors on absenteeism, turnover, and job performance.* Journal of Business and Psychology, 4, 423-430.

[30] Peters, L. H., O'Connor, E. J., & Rudolph, C. J. (1980). *The behavioral and affective consequences of performance-relevant situational variables.* Organizational Behavior and Human Performance, 25, *79-96*.

[31] Copus, D. (2006). *Open Letter to OFCCP Re: Validation of Cognitive Ability Tests.* Morristown, NJ: Author; Weiner, J.A., Schmidt, F.L., Sharf, J.C., Pyburn, K.M., & Copus, D. (2006, May). *Validity generalization at work: Is it legal to be scientific?* Presentation at the 2006 SIOP Conference in Gaylord, TX.

[32] *See* Contreras v. City of Los Angeles, 656 F.2d 1267 (9th Cir. 1981), US v. Commonwealth of Virginia, 569 F2d 1300 (4th Cir. 1978), Waisome v. Port Authority, 948 F.2d 1370, 1376 (2d Cir. 1991).

[33] Dickerson v. U. S. Steel Corp., 472 F. Supp. 1304 (E.D. Pa. 1978).

[34] NAACP Ensley Branch v. Seibels, 13 Emp. Prac. Dec. (CCH) ¶ 11,504 (N.D. Ala. 1977), at pp. 6793, 6803, 6806, *aff'd in relevant part, rev'd in other part,* 616 F.2d 812, 818 and note 15 (5th Cir.), *cert. den.* Personnel Board of Jefferson County v. U.S., 449 U.S. 1061 (1980).

[35] *Atlas*, supra, note 14.

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

[37] 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 (1981).

[38] Brunet v. City of Columbus, 1 F.3d 390 (6th Cir. 1993).

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

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

[41] Zamlen v. City of Cleveland, 686 F.Supp. 631 (N.D. Ohio 1988).

[42] Sackett, P. R., Schmitt, N., Tenopyr, M. L., Kehoe, J., & Zedeck, S. (1985). *Commentary on forty questions about validity generalization and meta-analysis.* Personnel Psychology, 38, *697-798*.

[43] Dye, D. A., Reck, M., & McDaniel, M. A. (July, 1993). *The validity of job knowledge measures.* International Journal of Selection and Assessment, 1 (3), 153-157.