Hiring the Most Qualified Firefighters While Avoiding (and Defending) Lawsuits

Daniel A. Biddle, Ph.D.

CEO, Fire & Police Selection, Inc.

Fire & Police Selection, Inc. (FPSI)
193 Blue Ravine Rd., Suite 270 Folsom, California 95630

Office: 888-990-3473 | Fax: 916.294.4240

Email: info@FPSI.com

Websites: www.fpsi.com | www.NationalFireSelect.com

Contents

Overview: EEO and Testing in the Fire and EMS Service

This document includes the key strategies your fire department will need to stay EEO-litigation free and maintain a balanced workforce. These two goals—reducing litigation and achieving a diverse workforce—are also *related* goals because departments that have proactive plans to build a diverse workforce also tend to be involved in fewer lawsuits. While achieving one does not guarantee the other, the strategically-managed fire department will seek them both simultaneously.

The first goal (reducing EEO litigation) can only be achieved by first becoming *aware* of the basics behind EEO laws and regulations, and understanding how they work, including what sets the litigation into motion and how to avoid it. The second goal (achieving a diverse workforce) can be accomplished by understanding how to properly develop and use a balanced set of assessments for hiring that will assure that the most well-rounded, qualified, and diverse group of employees start in the career-climbing process in your department.

This document includes useful information for departments to forge into a strategic plan for accomplishing both of these goals. The first section outlines the basics surrounding civil rights, Title VII of the 1991 Civil Rights Act ("1991 CRA" hereafter), and *adverse impact*, which is the trigger for Title VII Litigation. This section also highlights the importance of developing well-rounded hiring selection systems and demonstrates the very real impact that your department's testing practices can have on your bottom-line diversity results. The next section reviews the essentials for validating hiring tests using two of the most proven validation strategies—content and criterion-related validation. This section reviews a national study that was conducted regarding the key competencies needed to be a well-rounded fire fighter—a study where over 130 fire executives weighed 24 skills and abilities according to how well each contributes to success as a firefighter. Determining how testing practices should be used—whether pass/fail, banded, ranked, or weighted, is also discussed in this section. The final section provides a *validation checklist* that can be used to evaluate how likely your department's testing practices might hold up if they were targeted in a Title VII lawsuit.

1991 Civil Rights Act (Title VII)

With the passage of the Civil Rights Act in 1964, it officially became illegal for employers to segregate their workforces based on race, or to otherwise consider race when making employment-related decisions. While the vast majority of employers readily adopted and welcomed this much-needed new piece of civil rights legislation, some employers looked for the loopholes. One such employer was the Duke Power Company, which (on the very date that the Civil Rights Act became effective—July 2, 1965), established a new policy of requiring applicants for jobs in the traditional white classifications (including transfers from other departments), to score sufficiently high on two aptitude tests in addition to possessing a high school diploma. This installed three potential barricades to future career advancement. After

years of deliberation in the courts regarding the validity of these tests relative to the requirements of the at-issue positions, the U.S. Supreme Court ruled (in Griggs v. Duke Power, 1971), by a vote of 8 to 0 that the tests constituted an illegal barricade to employment because they did not bear a *sufficient connection to the at-issue positions*. In other words, they were not *valid* for the positions. In essence, the Court ruled that invalid practices, however neutral in intent, that caused an "adverse effect" upon a group protected by the act, were illegal.

About one year after this decision, federal civil rights enforcement agencies (e.g., the U.S. EEOC and Department of Labor's enforcement arm, the OFCCP) began forging versions of the federal Uniform Guidelines that were designed to interpret the concepts laid down in the Griggs case. Finally, in 1978, four federal agencies solidified a set of Uniform Guidelines that serve to interpret the "job relatedness" and "validity" requirements from the Griggs case, as reported by the Questions & Answers accompanying the Guidelines:

> Question: What is the basic principle of the Guidelines? Answer: A selection process which has an adverse 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 (Uniform Guidelines, Q&A #2).

But the principles laid down in Griggs have not gone without challenge. In 1989, the U.S. Supreme Court handed down a decision that drastically changed the Griggs principles that had been incorporated into the Uniform Guidelines. This landmark case, Ward's Cove Packing Co. v. Atonio (1989), revised the legal burdens surrounding proof of discrimination from the Griggs-based employer's burden (where the employer had to prove that any disparate impact[1] caused by the testing practices was justified by job relatedness/validity) to the burden of proof *remaining with the plaintiff at all times*. Under Ward's Cove, all the employer was required to do was "produce a business justification" that would justify the disparate impact—a legal burden much easier to address than the "job relatedness/business necessity" standard laid down in Griggs.

In 1991, an Act of Congress *reversed* Wards Cove (and several other similar U.S. Supreme Court decisions it found disagreeable). This Act—the 1991 Civil Rights Act (or "1991 CRA")—shifted the validation burden back to the Griggs standard by specifically stating the two conditions in which an employer's testing practice will be deemed illegal:

> A(i) 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; OR, A(ii) the complaining party makes the demonstration described in subparagraph (C) with respect to an alternate employment practice, and the respondent refuses to adopt such alternative employment practice (Section 2000e-2[k][1][A][i]).

The above section from the 1991 CRA effectively summarizes the current "law of the land" with respect to disparate impact and test validation. Two key observations from this section should be pointed out. First, notice that the language is *job-specific*: "the challenged practice is job-related *for the position in question* and consistent with business necessity." Thus, by definition, evaluating the validity of a test is a test-by-test, job-by-job determination. And second, note that Section A(ii) is preceded by the word OR, indicating that there are clearly two routes available for proving adverse impact, with the first being the "classic" method (adverse impact with no validity), and second as the showing of a "alternative employment practice" that could be used with less adverse impact. Several cases have been tried using this alternative strategy, but this concept is not discussed further here because the classic type is the most common.

Legal Update: What's Happened Since the 1991 CRA?

Since the passage of the 1991 CRA, there has been a steady stream of disparate impact cases in the fire industry. As of the time of this writing, no cases have changed the foundational framework of the Griggs standard codified in the 1991 CRA. There have, however, been some changes on the legal front regarding the extent to which race-conscious actions can be taken *after* a testing process has been administered.

These changes were spurred by the Ricci v. DeStefano cased tried in the U.S. Supreme Court in 2009. The Ricci case arose from a lawsuit brought against the City of New Haven, Connecticut by 19 City firefighters who claimed that the City discriminated against them with regard to promotions. The firefighters (17 whites and 2 Hispanics), had all passed the promotional tests for the Captain or Lieutenant positions and just prior to finalizing their promotion eligibility, the City invalidated the test results because none of the black firefighters who passed the exam had scored high enough to be considered for the positions. The City claimed that their reason for cancelling the list was that they "feared a lawsuit over the test's disparate impact."

Ultimately, the Supreme Court found this decision to be discriminatory because the City lacked the *strong basis in evidence* that the City would have lost a disparate impact lawsuit (because the tests were not sufficiently valid). While there was no real contention regarding the existence of adverse impact, the evidence to weigh the validity of the test was never admitted into evidence. Rather, the City blocked the validation evidence (by requesting that the test vendor to not send them a validation report—one which was already scheduled for delivery) and

the plaintiffs (whites in this case) also did not want to contend the validity of the case. In this interesting set of circumstances, the actual validity of the tests was in a "pickle" situation—the City was motivated to prove that the tests *were not* valid because they were trying to claim that their actions of redacting the exam results was justified because the tests were "sufficiently invalid" and the plaintiffs also did not want the validity contested (because they wanted the exam results to stand and not be invalidated). Based on this very unique set of circumstances, a split (5-4) decision was rendered by the Court on June 29, 2009. The Court's ruling held that the City's decision to ignore the test results violated Title VII of the Civil Rights Act of 1964.

Immediately after the case, legal blogs and presentations spread through the Internet—with some even claiming that the Griggs/1991 CRA legal foundation had been changed (which was clearly not the case). In fact, to bring clarity to the distinctions between the Ricci case—which was essentially a *disparate treatment case that involved a test*—some courts began clarifying that the Griggs/1991 CRA standard was still alive and well. One such case was Vulcan Society v. City of New York (2009) which was decided just weeks after the Ricci ruling. In Vulcan, the judge clarified the distinctions between the Ricci ruling (disparate *treatment*) and typical disparate *impact* cases:

> Before proceeding to the legal analysis, I offer a brief word about the Supreme Court's recent decision in Ricci … I reference Ricci not because the Supreme Court's ruling controls the outcome in this case; to the contrary, I mention Ricci precisely to point out that it does not. In Ricci, the City of New Haven had set aside the results of a promotional examination, and the Supreme Court confronted the narrow issue of whether New Haven could defend a violation of Title VII's disparate treatment provision by asserting that its challenged employment action was an attempt to comply with Title VII's disparate impact provision. The Court held that such a defense is only available when "the employer can demonstrate a strong basis in evidence that, had it not taken the action, it would have been liable under the disparate-impact statute." Id. at 2664. In contrast, this case presents the entirely separate question of whether Plaintiffs have shown that the City's use of [the Exams] has actually had a disparate impact upon black and Hispanic applicants for positions as entry-level firefighters. Ricci did not confront that issue…The relevant teaching of Ricci, in this regard, is that the process of designing employment examinations is complex, requiring consultation with experts and careful consideration of accepted testing standards. As discussed below, these requirements are reflected in federal regulations and existing Second Circuit precedent. This legal authority sets forth a simple principle: municipalities must take adequate measures to ensure that their civil service examinations reliably test the relevant knowledge, skills and abilities that will determine which applicants will best perform their specific public duties.

The gears of several disparate impact cases continued to turn during and after the Ricci case, with the Griggs standards in full effect. Should a case come along some day that does change the Griggs standard, a commensurate change to the Uniform Guidelines will also likely be required.

Adverse Impact: the Trigger for Title VII Litigation

Each situation where a Title VII claim is made has a common denominator: adverse impact. The way the federal law currently stands, plaintiffs cannot even bring a lawsuit unless and until a test causes adverse impact. Because this trigger underlies all of EEO-related litigation (at least of the disparate impact variety), some attention will now be turned to defining "adverse impact."

Rather than entering into a protracted academic definition of adverse impact, the topic will be simplified into just a few paragraphs.[2] When defining adverse impact, it should first be pointed out that there are some important, key terms relevant to the matter. These include the 80% test, statistical significance, and practical significance. Each is described briefly below.

80% Test

This test is calculated by dividing the focal group's (the focal group is typically minorities or women) passing rate on a test by the reference group's (typically whites or men) passing rate. Any resulting value less than 80% constitutes a "violation" of this test. This test was originally framed in 1972 (see Biddle, 2006), was codified in the Uniform Guidelines in 1978, and has been referenced in hundreds of court cases. Despite its widespread use, it should not be regarded as the final litmus test for determining adverse impact (this position is held exclusively by statistical significance tests, described next).

Statistical Significance Tests

A "statistically significant finding" is one that raises the eyebrows of the researcher and triggers the thought, "I think I've found something here, and it is not likely due to chance." So, in the realm of an adverse impact analysis, if a researcher conducts an adverse impact analysis and obtains a statistically significant result, they are capable of stating that a legitimate trend, and not a chance relationship, actually exists (with a reasonable level of certainty).

Statistical significance tests result in a p-value ("p" for probability), with p-values ranging from 0 to +1. A p-value of 0.01 means that the odds of the event occurring by chance is only 1%. A p-value of 1.0 means that there is essentially a 100% certainty that the event is "merely a chance occurrence," and cannot be considered as a "meaningful finding." P-values of .05 or less are said to be "statistically significant" in the realm of EEO analyses. This .05 level (or 5%) corresponds with the odds ratio of "1 chance in 20." This 5% chance level is the p-value threshold that has been endorsed in nearly every adverse impact case or federal enforcement setting.

Conducting a statistical significance adverse impact analysis is very straight-forward, provided that a statistical software program is used.[3] The process is completed by applying a statistical test to a 2 X 2 table, where the success rates (e.g., pass or failing a test) of two groups (e.g., Whites and Asians) are compared. See the example below.

2 X 2 Table Example

Group	Promoted	Not Promoted
Whites	30	20
Asians	20	30

There are over 20 possible statistical tests that can be used for computing the statistical significance of a 2 X 2 table, including "estimation" and "exact" methods, and various models that make certain assumptions regarding how the 2 X 2 table itself is constructed (see Biddle & Morris, 2011 for a complete discussion). For example, using an "estimation" technique (such as the Chi-Square computation) on the table above returns a p-value of .046 (below the .05 level needed for a "statistically significant" finding); whereas using a more exact method (the Fisher's Exact Test with Lancaster's Mid-P Correction)[4] returns a p-value of .06 (not significant).

Practical Significance

The concept of practical significance in the EEO analysis field was first introduced by Section 4D of the Uniform Guidelines ("Smaller differences in selection rate may nevertheless constitute adverse impact, where they are significant in *both statistical and practical terms* …"). Practical significance tests are applied to adverse impact analyses to evaluate the "practical impact" (typically reported as the shortfall pertaining to the group with the lower passing rate) or "stability" of the results (evaluating whether a statistical significance finding still exists after changing the passing/failing numbers of the disadvantaged group). While this concept enjoyed a run in the federal court system (Biddle, 2006), it has more recently been met with a considerable level of disagreement and condemnation in the courts.[5] For example, in the most recent circuit-level case dealing with practical significance, the court stated:

> Similarly, this Court has never established "practical significance" as an independent requirement for a plaintiff's prima facie disparate impact case, and we decline to do so here. The EEOC Guidelines themselves do not set out "practical" significance as an independent requirement, and we find that in a case in which the statistical significance of some set of results is clear, there is no need to probe for additional "practical" significance. Statistical significance is relevant because it allows a fact-finder to be confident that the relationship between some

> rule or policy and some set of disparate impact results was not the product of chance. This goes to the plaintiff's burden of introducing statistical evidence that is "sufficiently substantial" to raise "an inference of causation." Watson, 487 U.S. at 994-95. There is no additional requirement that the disparate impact caused be above some threshold level of practical significance. Accordingly, the District Court erred in ruling "in the alternative" that the absence of practical significance was fatal to Plaintiffs' case (Stagi v. National Railroad Passenger Corporation, 2010).

For these reasons, while the Uniform Guidelines are clear that practical significance evaluations (such as shortfalls and statistical significance "stability" tests) are *conceptually relevant* to adverse impact analyses, employers should "tread carefully" when evaluating the practical significance of adverse impact analysis results. While the concept is relevant, it will ultimately be left to a judge to decide whether (and to what extent) practical significance can be used in court. Certainly it would be a risky endeavor to adopt hard-and-fast practical significance rules to apply when analyzing adverse impact.

Validation: The Legal Requirement When Tests Exhibit Adverse Impact

Test validation is oftentimes a misunderstood topic. Yes, validation is a legal obligation whenever an employer's test exhibits adverse impact. But it's much more than that—validation is actually a scientific process for insuring that your department's testing process is focusing on the *key competencies that are needed for job success*. Without a properly-validated hiring process, hiring authorities might as well start pulling names out of a hat. So validation should be regarded as having two benefits: (1) Utility (the benefits enjoyed by employers by hiring highly-qualified candidates), and (2) legal defensibility.

Validation Techniques and Legal Requirements for Testing

Practically speaking, there are only two effective validation strategies: content validity and criterion-related validity. Both strategies are supported by the Uniform Guidelines, professional standards, and numerous court cases. While there are several possible angles to develop tests under either strategy, the most basic components of each are discussed below.

Content Validity

Content validity evidence is amassed by demonstrating a nexus (i.e., a connection) between the test and important job requirements. If conducted from the ground-up, a typical content validation study will include the following steps:

1. **Conducting Job Analysis Research.** Establishing content validity evidence requires having a clear understanding of what the job requires—especially the areas that are

targeted by the test. Generally speaking, content validity evidence is stronger in circumstances where a clear picture of the job has been developed through a thorough job analysis process.

2. **Developing a Clear Test Plan.** A Test Plan identifies the key knowledges, skills, abilities, and personal characteristics (KSAPCs) from the job analysis process that are necessary on the *first day* of employment. Ideally, the most important KSAPCs are included in the Test Plan.

3. **Connecting the Test to the Job.** A process needs to be completed that *establishes or demonstrates* a clear nexus between the test and the important job KSAPCs. This can be completed by using the expert opinions of either methodology (testing) experts or Job Experts. Ultimately, the KSAPCs measured by the test need to be linked to the important KSAPCs of the job, which are then linked to the important job duties. This three-way chaining process establishes content validity.

4. **Establishing How the Tests will be *Used*.** Adopting a content validity process requires *using* (e.g., ranking, pass/fail, banded) the test results in a way that accurately reflects how the important KSAPCs measured by the test are *actually applied on the job*. For example, possessing basic math skills is necessary for being a competent firefighter, but possessing increasingly higher levels of this skill does not necessarily translate into superior performance as a firefighter. Other skills, such as teamwork and interpersonal skills, are more likely to *differentiate performance* between firefighters when held at above-minimum levels. Following in this same spirit, tests measuring basic math should be used on a pass/fail (cutoff) basis, whereas tests measuring differentiating KSAPCs should be the primary basis for ranking decisions. A complete discussion of test use considerations is provided below.

Criterion-related Validity

Criterion-related validity is statistical in nature, and is established by demonstrating a significant correlation between the test and some important aspect of job performance. For example, a department might have the supervisory staff assign job performance ratings to the firefighters, run the firefighters through a physical ability test, then conduct a statistical correlation between the test and job performance ratings to assess whether they are significantly correlated.

Criterion-related validity studies can be conducted in one of two ways: using a predictive model or a concurrent model. A predictive model is conducted when applicant test scores are correlated to subsequent measures of job performance (e.g., six months after the tested applicants are hired). A concurrent model is conducted by giving a test to incumbents who are currently on the job and then correlating these scores to current measures of job performance (e.g.,

performance review scores, supervisor ratings, etc.). The following steps can be completed to conduct a predictive criterion-related validity study:

1. Be sure that your department has at least 150 subjects to include in the study (these are applicants who will take the pre-employment tests and will be subsequently hired). This is helpful for detecting whether a significant correlation will exist (smaller samples won't be as reliable).

2. Conduct a job analysis (see previous section) or a "review of job information" to determine the important aspects of the job that should be included in the study (both from the testing side and the rating side).

3. Develop one or more criterion measures by developing subjective (e.g., job performance rating scales) or objective measures (e.g., absenteeism, work output levels) of critical areas from the job analysis or job information review. A subjectively-rated criterion should only consist of performance on a job duty (or group of duties). In most cases it should not consist of a supervisor's or peer's rating on the incumbent's level of KSAPCs (a requirement based on Section 15B5 of the Uniform Guidelines) unless the KSAPCs are clearly linked to observable work behaviors, or they are sufficiently operationally defined. It is important that these measures have sufficiently high reliability (at least .60 or higher is preferred).

4. Work with Job Experts and supervisors, trainers, other management staff, and the job analysis data to form solid speculations ("hypotheses") regarding which KSAPCs "really make a difference" in the high/low scores of such job performance measures (above).

5. Develop tests that are reliable measures of those KSAPCs. Choosing tests that have reliability of .70 or higher is preferred.[6]

6. After a period of time has passed and criterion data has been gathered (e.g., typically between 3 and 12 months), correlate each of the tests to the criterion measures using the =PEARSON command in Microsoft Excel and evaluate the results.

To complete a concurrent criterion-related validation study, complete steps 2–5 above and replace step 6 by administering the test to the current incumbent population and correlate the test scores to current measures of job performance.

In either study design, the resulting correlation coefficient must, at a minimum, be statistically significant at the .05 level (before making any corrections). Ideally, it should also be sufficiently strong to result in practical usefulness in the hiring process. The U.S. Department of Labor (2000, pp. 3-10) has provided reasonable guidelines for interpreting correlation

coefficients, with coefficients between .21 and .35 classified as “likely to be useful” and coefficients higher than .35 as “very beneficial.”

Building a Balanced Hiring Program by Sorting the Key Competencies for Success

Sometimes fire executives ask: “Which entry-level firefighter test is the best?” Answering this question requires re-framing it to: “What is the best *set* of tests for selecting the most qualified, well-rounded group of firefighters?” Until the testing field advances to using ultra-precision testing instruments that can conduct “deep scans” of each applicant’s true potential as a firefighter in just 20 minutes, the testing process for screening entry-level applicants will need to consist of a *wide range* of tests that measure a wide range of abilities. It will also continue to be an arduous process—by the time each applicant’s resume, background, written, and interview screens have been reviewed, tallied, and scored, the time investment typically exceeds several hours per candidate.

While the process will always be tedious, carefully choosing which tests to use and which competency areas to measure are the most important strategic factors to consider. To explore options for these key decision factors, this section provides the results of a national survey that was completed by over 130 fire executives (fire chiefs of all levels) that asked them to identify the most important competency areas needed for the entry-level firefighter position. Specifically, they were asked to assign 100 points to three major competency areas[7]:

- Cognitive/academic (such as reading, math, writing);
- Personal characteristics (such as working under stress, allegiance, integrity); and
- Physical abilities (such as upper body strength, stamina, speed).

After the survey participants assigned 100 points to these three major competency areas, they were asked to assign 100 points *within* each. The results are shown in Table 1.

Table 1. Entry-Level Firefighter Competency Weights

Cognitive/Academic (32% of Total)	**% Importance**
Math	10%
Reading	14%
Verbal Communication	15%
Writing	12%
Map Reading	8%
Problem Solving	15%
Strategic Decision-Making	13%
Mechanical Ability	12%
Personal Characteristics (40% of Total)	**% Importance**
Teamwork	12%
Working Under Stress	10%
Allegiance/Loyalty	9%
Truthfulness/Integrity	13%
Public Relations	8%
Emotional Stability	10%
Sensitivity	8%
Proactive/Goal-Oriented	8%
Thoroughness/Attention to Detail	9%
Following Orders	10%
Physical Abilities (28% of Total)	**% Importance**
Wrist/Forearm Strength	13%
Upper Body Strength	17%
Lower Torso and Leg Strength	17%
Speed	12%
Dexterity, Balance, and Coordination	16%
Endurance	21%

The survey revealed that supervisory fire personnel valued the cognitive/academic domain as 32% of the firefighter position, personal characteristics as 40%, and physical as 28%. It should be noted that most professionals who have seasoned their careers in the fire service would be the first to admit that the typical entry-level testing process does not reflect these ratios. In fact, most testing processes focus mostly on the cognitive/academic areas (typically through a pass/fail written test), use a basic physical ability test (again, pass/fail), and only measure a very limited degree of the "soft" personal characteristics through an interview process (but only for the final few candidates who are competing for a small number of open positions).

Not only does this disconnect result in a balance mismatch between the competencies that are required for the job and those that are included in the screening process, it results in hiring processes that sometimes unnecessarily exasperate adverse impact against minorities and/or women. For example, a hiring process that focuses exclusively on cognitive/academic skills will

magnify adverse impact against minorities. The other cost for using such an unbalanced hiring process is that it leaves a massive vacuum (40%, to be precise) in the personal characteristics area—leaving these important competencies completely untapped. A hiring process that over-emphasizes the importance of physical abilities will amplify adverse impact against women.

Beyond the adverse impact issues, other substantial problems occur when a fire department adopts an unbalanced hiring process. A testing process that leaves out (or under-measures) important cognitive/academic skills will likely result in washing out a significant number of cadets in the academy, and can also lead to poor performance on the job. On the other hand, over-measuring this area while under-measuring personal characteristics, could lead to a group of book-smart firefighters who have no idea how to work cooperatively in the close living conditions required by firefighters. Under-measuring physical abilities in the hiring process typically leads to a group of firefighters who cannot perform the strenuous physical requirements of the job—especially as they age in the fire service.

Challenges to Building a Balanced Testing Program and Recommendations for Success

The most significant challenge to building an effective testing program for entry-level firefighters lies with testing the "soft skills" (personal characteristics). This is because these skills—such as teamwork and interpersonal skills—are crucial ingredients for success but they are the most difficult to measure in a typical testing format. For example, developing a math test is easy; developing a test for measuring teamwork skills is not—but the latter was rated as more important for overall job success!

The reason for this is that many skills and abilities are "concrete" (opposed to theoretical and abstract). An applicant's math skills can readily be tapped using questions that measure numerical skills at the same level these skills are required on the job. Abstract or soft skills like teamwork are more difficult to measure during a two-hour testing session. Fortunately, there are some effective and innovative testing solutions available. These are discussed in the tables below.

Table 2. Proposed Solutions for Testing Cognitive/Academic Competencies for Entry-Level Firefighters

Cognitive/Academic (32% Overall Importance)	Weight	Proposed Testing Solution	Typical Validation Method
Math	10%	Use written or "work sample" format; measure using a limited number of multiple-choice items. Balance various types of math skills (add/subtract/multiple/divide, etc.).	CV
Reading	14%	Measure using either (1) "Test Preparation Manual" approach (where the applicants are given a manual and asked to study it for a few weeks prior to taking the test based on the Manual); or (2) a short reading passage containing material at a similar difficulty/context to the job that applicants are allowed to study during the testing session and answer related test items.	CV
Verbal Communication	15%	While a Structured Interview is the best tool for measuring this skill (because the skill includes verbal and non-verbal aspects), some level of this skill can be measured using word recognition lists or sentence clarity items.	CV
Writing	12%	Measure using writing passages or word recognition lists, sentence clarity, and/or grammar evaluation items.	CV
Map Reading	8%	Measure using maps and related questions asking applicants how they would maneuver to certain locations. Include directional awareness.	CV
Problem-Solving	15%	Measure using word problems measuring reasoning skills in job-rich contexts.	CRV
Strategic Decision-Making	13%	While a Structured Interview is the best tool for measuring this skill (because the applicant can be asked to apply this skill in firefighter-specific scenarios), some level of this skill can be measured using word problems or other contexts supplied in written format where applicants can consider cause/effect of certain actions.	CV
Mechanical Ability	12%	Using CV, measure mechanical comprehension skills such as leverage, force, and mechanical/physics contexts regarding weights, shapes, and distances. Also can measure spatial reasoning (when using a CRV validation strategy).	CV/CRV

Notes: CV: Content Validity; CRV: Criterion-related validity.

Table 3. Proposed Solutions for Personal Characteristics for Entry-Level Firefighters

Personal Characteristics (40% Overall)	Weight	Proposed Testing Solution	Typical Validation Method
Teamwork	12%	Under a CV strategy, a Situational Judgment Test (SJT) can be used for measuring these skills. Alternatively, a custom personality test can be developed using CRV. While these types of assessments can measure *if an applicant knows* the most appropriate response (using an SJT) or the best attitude or disposition (personality test), they are limited in that they cannot measure whether an applicant *would actually respond* in such a way. For these reasons, measuring the *underlying traits* that tend to generate these positive behaviors is typically the most effective strategy. Structured Interviews can also provide useful insight into these types of competencies, as well as background and reference evaluations. However, these tools are time consuming and expensive, so measuring these areas in the testing stage is an effective strategy.	CV/CRV
Working Under Stress	10%		
Allegiance/Loyalty	9%		
Truthfulness/Integrity	13%		
Public Relations	8%		
Emotional Stability	10%		
Sensitivity	8%		
Proactive/Goal-Oriented	8%	These competencies can be effectively measured using either an SJT (using a CV or CRV strategy) or a Conscientiousness (CS) scale (using a CRV strategy). A CS test can be developed using just 20-30 items (using likert-type responses). Such tests are typically successful in predicting job performance in fire settings.	CV/CRV
Thoroughness/ Attention to Detail	9%		
Following Orders	10%		

Notes: CV: Content Validity; CRV: Criterion-related validity.

Table 4. Proposed Solutions for Testing Physical Abilities for Entry-Level Firefighters

Physical Abilities (28% Overall)	Weight	Proposed Testing Solution	Typical Validation Method
Wrist/Forearm Strength	13%	It is the opinion of the authors that these unique physical competencies should be *collectively and representatively* measured in a work-sample style Physical Ability Test (PAT) (using a content validity strategy). While other types of test (such as clinical strength tests) that do not directly mirror the requirements of the job can be used (if they are based on CRV), the benefits of using a high-fidelity work sample has greater benefits.	CV
Upper Body Strength	17%		CV
Lower Torso/Leg Strength	17%		CV
Dexterity, Balance, Coord.	16%		CV
Speed	12%	Strenuous work-sample PATs can measure some level of endurance (and speed) if they are continuously-timed and exceed at least five (5) minutes in length. Actual cardiovascular endurance levels can only be measured using a post-job-offer V02 maximum test (which would require a using a CRV strategy).	CV/CRV
Endurance	21%		

Notes: CV: Content Validity; CRV: Criterion-related validity.

The importance weights displayed in the tables above may or may not be representative of individual fire departments. This is because some departments serve communities that have more multiple structure or high-rise fires than others do, some have a higher occurrence rate of EMS incidents, etc. Therefore, we recommend that each fire department investigate the relative

importance of these various competencies and the tests used to measure them (discussed further in the next section).

Test Use: Setting Pass/Fail Cutoffs, Banding, Weighting, and Ranking

Because validation has to do with the *interpretation of scores*, a perfectly valid test can be *invalidated* through improper use of the scores. Conducting a search through professional testing guidelines (e.g., the Principles, 2003, and Standards, 1999), the Uniform Guidelines (1978), and the courts, one can find an abundance of instruction surrounding how test scores should be used. The safe way to address this complex maze of guidelines is to be sure that tests are *used in the manner that the validation evidence supports*.

If classifying applicants into two groups—qualified and unqualified—is the end goal, the test should be used on a pass/fail basis (i.e., an absolute classification based on achieving a certain level on the test). If the objective is to make relative distinctions between substantially equally qualified applicants, then banding is the approach that should be used. Ranking should be used if the goal is making decisions on an applicant-by-applicant basis (making sure that the requirements for ranking discussed herein are addressed). If an overall picture of each applicant's combined mix of competencies is desired, then a weighted and combined selection process should be used.

For each of these different procedures, different types of validation evidence should be gathered to justify the corresponding manner in which the scores will be used and interpreted. This section explains the steps that can be taken to develop and justify each.

Developing Valid Cutoff Scores

Few things can be as frustrating as being the applicant who scored 69.9% on a test with an 70% cutoff! Actually, there *is* one thing worse: finding out that the employer elected to use a 70% as a cutoff for *no good reason whatsoever*. Sometimes this arbitrary cutoff is chosen because it just *seems* like a "good, fair place to set the cutoff" or because 70% represents a C grade in school. Arbitrary cutoffs simply do not make sense, neither academically nor practically. Further, they can incense applicants who might come to realize that a meaningless standard in the selection process has been used to make very *meaningful decisions* about their lives and careers.

For these reasons, and because the federal courts have so frequently rejected arbitrary cutoffs that have adverse impact, it is essential that practitioners use *best practices* when developing cutoffs. And, when it comes to best practices for developing cutoffs, there is perhaps none better than the *modified Angoff method*.[8] The Angoff method is solid because it makes good practical sense, Job Experts can readily understand it, applicants can be convinced of its validity, the courts have regularly endorsed it,[9] and it stands up to academic scrutiny.

Developing a cutoff score using this method is relatively simple: Job Experts review each item on a written test and provide their "best estimate" on the percentage of minimally qualified

applicants they believe would answer the item correctly (i.e., each item is assigned a percentage value). These ratings are averaged[10] and a valid cutoff for the test can be developed. The *modified* Angoff method adds a slight variation: After the test has been administered, the cutoff level set using the method above is lowered by one, two, or three Conditional Standard Errors of Measurement (C-SEMs)[11] to adjust for the unreliability of the test.

The Uniform Guidelines require that pass/fail cutoffs should be ". . . set so as to be reasonable and consistent with the normal expectations of acceptable proficiency in the workforce" (Section 5H). The modified Angoff method addresses this requirement on an item-by-item basis.

Setting Cutoffs that are Higher than the "Minimum Cutoff Level"

It is not uncommon for large fire departments to be faced with situations where thousands of applicants apply for only a handful of open positions. What should be done if the department cannot feasibly process all applicants who pass the validated cutoff score? Theoretically speaking, all applicants who pass the modified Angoff cutoff are qualified; however, if the department simply cannot process the number of applicants who pass the given cutoff, two options are available.

The first option is to use a cutoff that is *higher* than the minimum level set by the modified Angoff process. If this option is used, the Uniform Guidelines are clear that the degree of adverse impact should be considered (see Section 3B and 5H). One method for setting a higher cutoff is to subtract one Standard Error of Difference (SED)[12] from the highest score in the distribution, and passing all applicants in this score band. Using the SED in this process helps ensure that all applicants within the band are *substantially equally qualified*. Additional bands can be created by subtracting one SED from the score immediately below the band for the next group, and repeating this process until the first cutoff score option is reached (i.e., one Conditional SEM below the cutoff score). This represents the distinguishing line between the qualified and unqualified applicants.

While this option may be useful for obtaining a smaller group of applicants who pass the cutoff score and are substantially equally qualified, a second option is strict rank ordering. Strict rank ordering is not typically advised on written tests because of the high levels of adverse impact that are likely to result and because written tests typically only include a narrow measurement of the wide competency set that is needed for job success. To hire or promote applicants in strict rank order on a score list, the employer should be careful to ensure that the criteria in the Ranking section below are sufficiently addressed.

Banding

In some circumstances applicants are rank-ordered on a test and hiring decisions between applicants are based upon score differences at the one-hundredth or one-thousandth decimal place (e.g., applicant A who scored 89.189 is hired before applicant B who scored 89.188, etc.). The troubling issue with this practice is that, if the test was administered a second time,

applicants A and B could very likely change places! In fact, if the reliability of the test was low and the standard deviation was large, these two applicants could be separated by several whole points.

Banding addresses this issue by using the Standard Error of Difference (SED) to group applicants into "substantially equally qualified" score bands. The SED is a tool that can be used by practitioners for setting a confidence interval around scores that are substantially equal. Viewed another way, it can be used for determining scores in a distribution that represent *meaningfully different* levels of the competencies measured by the test.

Banding has been a hotly debated issue in the personnel field, especially over the last 20 years.[13] Proponents of strict rank ordering argue that making hiring decisions in rank order preserves meritocracy and ultimately ensures a slightly more qualified workforce. Supporters of banding argue that, because tests cannot adequately distinguish between small score differences, practitioners should remain blind to miniscule score differences between applicants who are within the same band. They also argue that the practice of banding will almost always produce less adverse impact than strict rank ordering.[14] While these two perspectives may differ, various types of score banding procedures have been successfully litigated and supported in court,[15] with the one exception being the decision to band *after* a test has been administered, if the only reason for banding was to reduce adverse impact (Ricci, 2009). Thus, banding remains as an effective tool that can be used in most personnel situations.

Ranking

The idea of hiring applicants in strict order from the top of the list to the last applicant above the cutoff score is a practice that has roots back to the origins of the merit-based civil service system. The limitation with ranking, as discussed above, is that the practice treats applicants who have almost tied scores as if their scores are meaningfully different *when we know that they are not*. The C-SEM shows the degree to which scores would likely shuffle if the test was hypothetically administered a second time.

Because of these limitations, the Uniform Guidelines and the courts have presented rather strict requirements surrounding the practice of rank ordering. These requirements are provided below, along with some specific recommendations on the criteria to consider before using a test to rank order applicants.

Section 14C9 of the Uniform Guidelines states:

> If a user can show, by a job analysis or otherwise, that a higher score on a content valid test is likely to result in better job performance, the results may be used to rank persons who score above minimum levels. Where a test supported solely or primarily by content validity is used to rank job candidates, the test should *measure those aspects of performance which differentiate among levels of job performance*.

Performance differentiating KSAPCs distinguish between acceptable and above-acceptable performance on the job. Differentiating KSAPCs can be identified either *absolutely* (each KSAPC irrespective of the others) or *relatively* (each KSAPC relative to the others) using a "Best Worker" likert-type rating scale to rate KSAPCs regarding the extent to which it distinguishes the "minimal" from the "best" worker. KSAPCs that are rated high on the Best Worker rating are those that, when performed above the "bare minimum," distinguish the "best" performers from the "minimal." As discussed above, possessing basic math skills is a necessity for being a competent firefighter, but possessing increasingly higher levels of this skill does not necessarily translate into superior performance as a firefighter. Other skills, such as teamwork and interpersonal skills, are more likely to differentiate performance when held at above-minimum levels.

A strict rank ordering process should not be used on a test that measures KSAPCs that are only needed at *minimum levels* on the job and do not distinguish between acceptable and above-acceptable job performance (see the Uniform Guidelines Questions & Answers #62). Content validity evidence to support ranking can be established by linking the parts of a test to KSAPCs that are performance differentiating.[16] So, if a test is linked to a KSAPC that is "performance differentiating" either *absolutely* or *relatively* (e.g., with an average differentiating rating that is 1.0 standard deviation above the average rating compared to all other KSAPCs), some support is provided for using the test as a ranking device.

While the Best Worker rating provides some support for using a test as a ranking device, some additional factors should be considered before making a decision to use a test in a strict rank-ordered fashion:

1. Is there adequate score dispersion in the distribution (or a "wide variance of scores")? Rank ordering is usually not preferred if the applicant scores are "tightly bunched together"[17] because such scores are "tied" to even a greater extent than if they were more evenly distributed. One way to evaluate the dispersion of scores is to use the C-SEM to evaluate if the score dispersion is adequately spread out within the relevant range of scores when compared to other parts of the score distribution. For example, if the C-SEM is very small (e.g., 2.0) in the range of scores where the strict rank ordering will occur (e.g., 95 – 100), but is very broad throughout the other parts of the score distribution (e.g., double or triple the size), the score dispersion in the relevant range of interest (e.g., 95-100) may not be sufficiently high to justify rank ordering.

2. Does the test have high reliability? Typically, reliability coefficients should be .85 to .90 or higher for using the results in strict rank order.[18] If a test is not reliable (or "consistent") enough to "split apart" candidates based upon very small score differences, it should not be used in such a way that considers small differences between candidates as meaningful.

While the guidelines above should be considered when choosing a rank ordering or pass/fail strategy for a test, the extent to which the test measures KSAPCs[19] that are performance differentiating should be the *primary consideration*.

Employers using a test that is based on criterion-related validity evidence have more flexibility to use ranking than with tests based on content validity. This is because criterion-related validity demonstrates scientifically what content validity can only speculate what is occurring between the test and job performance. Criterion-related validity provides a correlation coefficient that represents the strength or degree of correlation relationship between some aspects of job performance and the test.

While the courts have regularly endorsed criterion-related validity studies, they have placed some minimum thresholds for the correlation value necessary (typically .30 or higher) for strict rank ordering on a firefighter tests based on criterion-related validity:

- Brunet v. City of Columbus (1993). This case involved an entry-level firefighter Physical Capacities Test (PCT) that had adverse 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). This case involved an entry-level firefighter written test. 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). This case involved an entry-level firefighter written test. 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 adverse impact, the higher the correlation which will be required."

- Zamlen v. City of Cleveland (1988). 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.

Weighting Tests into Combined Scores

Tests can be weighted and combined into a composite score for each applicant. Typically, each test that is used to make the combined score is also used as a screening device (i.e., with a pass/fail cutoff) before including scores from applicants into the composite score. Before using a test as a pass/fail device and as part of a weighted composite, the developer should evaluate the extent to which the KSAPCs measured by the tests are performance differentiating—especially if the weighted composite will be used for ranking applicants.

There are two critical factors to consider when weighting tests into composite scores: (1) determining the weights and (2) standardizing the scores. Developing a reliability coefficient[20] for the final list of composite scores is also a critical final step if the final scores will be banded into groups of substantially equally qualified applicants. These steps are discussed below.

Determining a set of job-related weights to use when combining tests can be a sophisticated and socially sensitive issue. Not only are the statistical mechanics often complicated, choosing one set of weights versus another can sometimes have very significant impact on the gender and ethnic composition of those who are hired from the final list. For these reasons, this topic should be approached with caution and practitioners should make decisions using informed judgment.

Generally speaking, weighting the tests that will be combined into composite scores for each applicant can be accomplished using one of three methods: *unit weighting*, weighting based on *criterion-related validity* studies, and using *content validity* weighting methods.

Unit weighting is accomplished by simply allowing each test to share an equal weight in the combined score list. Surprisingly, sometimes unit weighting produces highly effective and valid results (see the Principles, 2003, p. 20). This is probably because each test is equally allowed to contribute into making the composite score, and no test is hampered by only contributing a small part to the final score. Using unit weighting, if there are two tests, they are each weighted 50%. If there are five, each is allowed 20% weight.

If the tests that are based on one or more criterion-related validity studies are being used, the data from these studies can be used to calculate the weights for each. The steps for this method are outside the scope of this text and will not be discussed here.[21]

Using content validity methods to weight tests is probably the most common practice. Sometimes practitioners get caught up in developing complicated and computationally-intensive methods for weighting tests using job analysis data. Sometimes these procedures involve using complicated formulas that consider frequency and importance ratings for job duties and/or KSAPCs, and/or the linkages between these. While this helps some practitioners feel at ease, these methods can produce misleading results. Not only that, there are easier methods available (proposed below).

For example, consider two KSAPCs that are equally important to the job. Now assume that one is more complex than the other, so it is divided into two KSAPCs on the job analysis and the other (equally important) KSAPC remains in a single slot on the Job Analysis. When it comes time to use multiplication formulas to determine weights for the tests that are linked to these two KSAPCs, the first is likely to receive more weight *just because it was written twice on the Job Analysis*. The same problem exists if tests are mechanically linked using job duties that have this issue.

What about just providing the list of KSAPCs to a panel of Job Experts and having them distribute 100 points to indicate the relative to the importance of each? This method is fine, but can also present some limitations. Assume there are 20 KSAPCs and Job Experts assign

importance points to each. Now assume that only 12 of these KSAPCs are actually tested by the set of tests chosen for the weighted composite. Would the weight values turn out differently if the Job Experts were allowed to review the 12 remaining KSAPCs and were asked re-assign their weighting values? Most likely, yes.

Another limitation with weighting tests by evaluating their relative weight from job analysis data is that sometimes different tests are linked to the same KSAPC (this can cause the weights for each test are no longer unique and become convoluted with other tests). One final limitation is that sometimes tests are linked to a KSAPC for collecting the weight determination, but they are weak measures of the KSAPC (while others are strong, relevant linkages). For these reasons, there is a "better way" (discussed below).

Steps for Weighting Tests Using Content Validity Methods

The following steps can be taken to develop content valid weights for tests that are combined into single composite scores for each applicant:

1. Select a panel of 4 to 12 Job Experts who are truly experts in the content area and are diverse in terms of ethnicity, gender, geography, seniority (use a minimum of one year experience), and "functional areas" of the target position.

2. Provide a copy of the Job Analysis for each Job Expert. Be sure that the Job Analysis itemizes the various job duties and KSAPCs that are important or critical to the job.

3. Provide each Job Expert with a copy of each test (or a highly detailed description of the content of the test if confidentiality issues prohibit Job Experts from viewing the actual test).

4. Explain the confidential nature of the workshop, the overall goals and outcomes, and ask the Job Experts to sign confidentiality agreements.

5. Discuss and review with Job Experts the content of each test and the KSAPCs measured by each. Also discuss the extent to which certain tests may be better measures of certain KSAPCs than others. Factors such as the vulnerability of certain tests to fraud, reliability issues, and others should be discussed.

6. Provide a survey to Job Experts that asks them to distribute 100 points among the tests that will be combined. Be sure that they consider the importance levels of the KSAPCs measured by the tests, and the job duties to which they are linked, when completing this step.

7. Detect and remove outlier Job Experts from the data set.

8. Calculate the average weight for each test. These averages are the weights to use when combining the test into a composite score.

Standardizing Scores

Before individual tests can be weighted and combined, they should be *standard scored.* Standard scoring is a statistical process of *normalizing* scores and is a necessary step to place different tests on a level playing field.

Assume a developer has two tests: one with a score range of 0 – 10 and the other with a range of 0 – 50. What happens when these two tests are combined? The one with a high score range will greatly overshadow the one with the smaller range. Even if two tests have the same score range, they should still be standard scored. This is because if the tests have different means and standard deviations they will produce inaccurate results when combined unless they are first standard scored.

Standard scoring tests is a relatively simple practice. Converting raw scores into *Z scores* (a widely used form of standard scoring) can be done by simply subtracting each applicant's score from the average (mean) score of all applicants and dividing this value by the standard deviation (of all applicant total scores). After the scores for each test are standard scored, they can be multiplied by their respective weights and a final score for each applicant calculated. After this final score list has been compiled, the reliability of the new combined list can be calculated.[22]

Would Your Tests Survive a Challenge? Checklists for Evaluating Your Test's Validity

Most tests in the fire service are supported under a content validation model. Some tests, such as personality tests and some types of cognitive ability tests, are supported using criterion-related validity. There are fundamental requirements under the Uniform Guidelines that should be addressed when a department claims either type in an enforcement or litigation setting. These are provided in the tables below.

Table 5. Content Validation Checklist for Written Tests

Req. #	Uniform Guidelines Requirement	Uniform Guidelines Reference
1	Does the test have **sufficiently high reliability?** (Generally, written tests should have reliability values that **exceed .70**[23] **for each section of the test that applicants are required to pass**).	14C(5)
2	Does the test measure Knowledges, Skills, or Abilities (KSAs) that are **important/critical** (essential for the performance of the job)?	14C(4)
3	Does the test measure Knowledges, Skills, or Abilities (KSAs) that are **necessary on the first day of the job** (Or, will they be trained on the job or could they be possibly "learned in a brief orientation"?).	14C(1), 4F, 5I(3)
4	Does the test **measure Knowledges, Skills, or Abilities (KSAs) that are concrete and not theoretical?** (Under content validity, tests cannot measure abstract "traits" such as intelligence, aptitude, personality, common sense, judgment, leadership, or spatial ability, if they are not defined in concrete, observable ways).	14C(1,4)
5	**Is sufficient time allowed for nearly all applicants to complete the test?**[24] (Unless the test was specifically validated with a time limit, sufficient time should be allowed for nearly all applicants to finish).	15C(5)
6	For tests measuring job knowledge only: Does the test measure job knowledge areas that need to be **committed to memory?** (Or, could the job knowledge areas be easily looked up without hindering job performance?).	15C(3), Q&A 79
7	Were "substantially equally valid" test alternatives (with less adverse impact) investigated?	3B, 15B(9)

Table 6. Criterion-related Validation Checklist for Written Tests

Req. #	Uniform Guidelines Requirement	Uniform Guidelines Reference
1	Is there a description of the test? Look for title, description, purpose, target population, administration, scoring and interpretation of scores.	15B(4)
2	If the test is a combination of other tests or if the final score is derived by weighting different parts of the test or different tests, is there a description of the rationale and justification for such combination or weighting?	15B(10)
3	Does the test have sufficiently high reliability (e.g., > .70[25]).	15C(7)
4	Is there a description of the criterion measure, including the basis for its selection or development and method of collection? For ratings, look for information related to the rating form and instructions to raters.	15B(5)
5	Does the criterion (*i.e.*, performance) measure reflect either: (a) important or critical work behaviors or outcomes as identified through a job analysis or review or (b) an important business need, such as absenteeism, productivity, tardiness or other?	14B(2)(3), 15B(3)
6	Is the sample size adequate **for each position that validity is being claimed**? Look for evidence that the correlations between the predictor and criterion measures are sufficient for each position included in the study.	14B(1)
7	Is the study sample representative of all possible test-takers? Look for evidence that the sample was chosen to include individuals of different races and gender. For concurrent validity studies, look for evidence that the sample included individuals with different amounts of experience. Where a number of jobs are studied together (e.g., a job group), look for evidence that the sample included individuals from all jobs included in the study.	14B(4)
8	Are the methods of analysis and results described? Look for a description of the method of analysis, measures of central tendency such as average scores, measures of the relationship between the predictor and criterion measures and breakouts results by race and gender.	15B(8)
9	Is the correlation between scores on the test and the criterion statistically significant **before** applying any statistical corrections?	14B(5)
10	Is the test being used for the same jobs (and test-takers) for which it was validated?	14B(6), 15B(10)
11	Have steps been taken to correct for overstatement and understatement of validity findings, such as corrections for range restriction, use of large sample sizes or cross-validation? If corrections are made, are the raw and corrected values reported?	14B(7)
12	Has the fairness of the test been examined or, is there a plan to conduct such a study?	14B(8)
13	Has a validation study been conducted in the last 5 years or, if not, is there evidence that the job has not changed since the last validity study?	5K
14	Were "substantially equally valid" test alternatives (with less adverse impact) investigated?	3B, 15B(9)
15	**If criterion-related validity for the test is being "transported"** from another employer/position, were the following requirements addressed? (a) the original validation study addressed Section 14B of the Guidelines, (b) the jobs perform substantially the same major work behaviors (as shown by job analyses in both locations), (c) a fairness study was conducted (if technically feasible).	7B, 14B

Table 7. Content Validation Checklist for Interviews

Req. #	UGESP Requirement	Uniform Guidelines Reference
1	If multiple raters are involved in the Interview Administration/Scoring, does the interview have **sufficiently high inter-rater reliability?** (Generally, interviews should have reliability values that **exceed .60** for each section of the interview that applicants are required to pass).	14C(5)
2	Does the interview measure Knowledges, Skills, or Abilities (KSAs) that are **important/critical** (essential for the performance of the job)?	14C(4)
3	Does the interview measure Knowledges, Skills, or Abilities (KSAs) that are **necessary on the first day of the job** (or, will they be trained on the job or could they be possibly "learned in a brief orientation"?).	14C(1), 4F, 5I(3)
4	Does the interview **measure Knowledges, Skills, or Abilities (KSAs) that are concrete and not theoretical?** (Under content validity, tests cannot measure abstract "traits" such as intelligence, aptitude, personality, common sense, judgment, leadership, or spatial ability, if they are not defined in concrete, observable ways).	14C(1,4)
5	For Interviews measuring job knowledge only: Does the interview measure job knowledge areas that need to be **committed to memory?** (Or, can the job knowledge areas be easily looked up without hindering job performance?).	15C(3), Q&A 79
6	Were "substantially equally valid" test alternatives (with less adverse impact) investigated?	3B, 15B(9)

Table 8. Content Validation Checklist for Work Sample (WS) or Physical Ability Tests (PATs)

Req. #	UGESP Requirement	Uniform Guidelines Reference
1	Does the test have **sufficiently high reliability**? (Typically, WS/WS/PATs need to be supported using test-retest reliability, unless they have a sufficient number of scored components to be evaluated using internal consistency. Generally, WS/PATs should have reliability values that exceed **.70** for each section of the test that applicants are required to pass).	14C(5)
2	Does the WS/PAT measure Knowledges, Skills, or Abilities (KSAs) that are **important/critical** (essential for the performance of the job)?	14C(4)
3	Does the WS/PAT measure Knowledges, Skills, or Abilities (KSAs) that are **necessary on the first day of the job** (or, will they be trained on the job or could they be possibly "learned in a brief orientation"?).	14C(1), 4F, 5I(3)
4	Does the WS/PAT measure Knowledges, Skills, or Abilities (KSAs) that are **concrete and not theoretical**? (Under content validity, tests cannot measure abstract "traits" such as intelligence, aptitude, personality, common sense, judgment, leadership, or spatial ability, if they are not defined in concrete, observable ways). Measuring "general strength," "fitness," or "stamina" cannot be supported under content validity unless they are *operationally defined in terms of observable aspects of work behavior* (job duties).	14C(1,4), 15C(5)
5	If the WS/PAT is designed to replicate/simulate actual work behaviors, are the **manner, setting, and level of complexity** highly similar to the job?	14C(4)
6	**If the WS/PAT has multiple events and is scored using a time limit** (e.g., all events must be completed in 5 minutes or faster), are the events in the WS/PAT typically performed on the job **with other physically-demanding duties performed immediately prior to and after each event**?	15C(5)
7	**If the WS/PAT has multiple events and is scored using a time limit** (e.g., all events must be completed in 5 minutes or faster), is **speed** typically important when these duties are performed on the job?	15C(5)
8	**If the WS/PAT includes weight handling requirements** (e.g., lifting, carrying certain objects or equipment), do they represent the **weights, distances, and duration** that objects/equipment are typically carried by a **single person** on the job?	15C(5)
9	If there are any **special techniques** that are learned on the job that allow current job incumbents to perform the events in the test better than an applicant could, are they demonstrated to the applicants before the test?	14C(1), 4F, 5I(3)
10	Does the WS/PAT **require the same or less exertion** of the applicant than the job requires?	5H, 15C(5)
11	Were "substantially equally valid" test alternatives (with less adverse impact) investigated?	3B, 15B(9)

Table 9. Validation Checklist for Using Test Results Appropriately

Req. #	UGESP Requirement	Uniform Guidelines Reference
1	If a **pass/fail cutoff** is used, is the cutoff "set so as to be reasonable and consistent with normal expectations of acceptable proficiency within the work force?"	5G, 5H, 15C(7)
2	If the test is **ranked or banded above a minimum cutoff level**, **and is based on content validity**, can it be shown that either **(a)** applicants scoring below a certain level have little or no chance of being selected for employment, or **(b)** the test measures KSAs / job duties that are "performance differentiating?"	3B, 5G, 5H, 14C(9)
3	If the test is **ranked or banded above a minimum cutoff level**, **and is based on criterion-related validity**, can it be shown that either **(a)** applicants scoring below a certain level have little or no chance of being selected for employment, or **(b)** the degree of statistical correlation and the importance and number of aspects of job performance covered by the criteria clearly justify ranking rather than using the test in a way that would lower adverse impact (e.g., banding or using a cutoff)? (Tests that have adverse impact and are used to rank that are only related to one of many job duties or aspects of job performance should be subjected to close review.)	3B, 5G, 5H, 14B(6)
4	Is the test **used in a way that minimizes adverse impact**? (Options include different cutoff points, banding, or weighting the results in ways that are still "substantially equally valid" but reduce or eliminate adverse impact)?	3B, 5G

References

Biddle, D. A. (2006). Adverse impact and test validation: A practitioner's guide to valid and defensible employment testing (2nd ed.). Ashgate Publishing Company: Burlington, VT.

Biddle, D.A. & Morris, S. B. (2010). Using Lancaster Mid-P correction to the Fisher Exact Test for adverse impact analyses. Unpublished Manuscript.

Boston Chapter, NAACP, Inc. v. Beecher, 504 F.2d 1017, 1026-27 (1st Cir. 1974).

Brunet v. City of Columbus, 1 F.3d 390, C.A.6 (Ohio, 1993).

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

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

Ricci v. DeStefano, 129 S. Ct. 2658, 2671, 174 L. Ed. 2d 490 (2009).

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.

Stagi v. National Railroad Passenger Corporation, No. 09-3512 (3d Cir. Aug. 16, 2010).

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 Tests, 43 Federal Register, 38,290-38,315; Adoption of Questions and Answers to Clarify and Provide a Common Interpretation of the Uniform Guidelines on Employee Tests, 44 Federal Register 11,996-12,009.

US 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.

Vulcan Society v. City of New York, 07-cv-2067 (NGG)(RLM) (July 22, 2009).

Ward's Cove Packing Co. v. Atonio, 490 U.S. 642 (1989).

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

Endnotes

[1] In this text, disparate impact and adverse impact mean the same.

[2] Readers interested in the historical and theoretical background of adverse impact are encouraged to read Adverse Impact and Test Validation, A Practitioner's Guide to Valid and Defensible Employment Testing (Biddle, 2006).

[3] See http://www.disparateimpact.com for an online tool for computing adverse impact.

[4] The Fisher Exact Test should not be used without this correction—see Biddle & Morris, 2010.

[5] See, for example: OFCCP v. TNT Crust (US DOL, Case No. 2004-OFC-3); Dixon v. Margolis (765 F. Supp. 454, N.D.Ill., 1991), Washington v. Electrical Joint Apprenticeship & Training Committee of Northern Indiana, 845 F.2d 710, 713 (7th Cir.), cert. denied, 488 U.S. 944, 109 S.Ct. 371, 102 L.Ed.2d 360 (1988). Stagi v. National Railroad Passenger Corporation, No. 09-3512 (3d Cir. Aug. 16, 2010).

[6] While these guidelines are suitable for most tests that have either a single or a few highly-related abilities being measured, sometimes wider guidelines should be adopted for multi-faceted tests that measure a wider range of competency areas (e.g., situational judgment, personality, behavior, bio-data tests).

[7] The survey was limited to competency areas that can be possibly measured in a testing process.

[8] When tests are based on criterion-related validity studies, cutoffs can be calibrated and set based on empirical data and statistical projections that can also be very effective.

[9] For example US v. South Carolina (434 US 1026, 1978) and Bouman v. Block (940 F.2d 1211, C.A.9 Cal., 1991) and related consent decree.

[10] Be careful to first remove unreliable and outlier raters before averaging item ratings into a cutoff score—see Biddle, 2006 for a set of recommended procedures.

[11] See the Appendix for recommended strategies for computing the C-SEM, SED, and creating score bands.

[12] The SED should also be set using a conditional process wherever feasible (see the Appendix).

[13] For example, Schmidt, F. L. (1991), 'Why all banding procedures in personnel selection are logically flawed', *Human Performance*, 4, 265-278; Zedeck, S., Outtz, J., Cascio, W. F., and Goldstein, I. L. (1991), 'Why do "testing experts" have such limited vision?', *Human Performance*, 4, 297-308.

[14] One clear support for using banding as a means of reducing adverse impact can be found in Section 3B of the Uniform Guidelines, which states: "Where two or more tests are available which serve the user's legitimate interest in efficient and trustworthy workmanship, and which are *substantially equally valid* for a given purpose, the user should use the procedure which has been demonstrated to have the lesser adverse impact." Banding is one way of evaluating an alternate use of a test (i.e., one band over another) that is "substantially equally valid."

[15] Officers for Justice v. Civil Service Commission (CA9, 1992, 979 F.2d 721, cert. denied, 61 U.S.L.W. 3667, 113 S. Ct. 1645, March 29th, 1993).

[16] See Section 14C4 of the Uniform Guidelines.

[17] Guardians v. CSC of New York (630 F.2d 79). One of the court's reasons for scrutinizing the use of rank ordering on a test was because 8,928 candidates (two-thirds of the entire testing population) was bunched between scores of 94 and 97 on the written test.

[18] Gatewood, R. D. & Feild, H.S. (1994), *Human Resource Selection* (3rd ed.) Fort Worth, TX: The Dryden Press (p. 184); Aiken, L.R. (1988). *Psychological Testing and Assessment* (2nd ed.). Boston: Allyn & Bacon (p. 100); Weiner, E. A. & Stewart, B. J. (1984). *Assessing Individuals.* Boston: Little, Brown. (p. 69).

[19] For tests that are designed to directly mirror job duties (called "work sample tests"), only test-duty (and not test-KSAPC) linkages are required for a content validity study (see Section 14C4 of the Guidelines). In this case, the Best Worker ratings on the duties linked to the work sample test should be the primary consideration for evaluating its use (i.e., ranking or pass/fail). For tests measuring KSAPCs (and not claiming to be direct "work sample tests"), the extent to which the test measures KSAPCs that are differentiating should be the primary consideration.

[20] See, for example, Mosier, C.I. (1943). On the reliability of a weighted composite. *Psychometrika, 8*, 161-168. (6,11).

[21] See Uniform Guidelines Questions & Answers #47, the Principles (2003, p. 20, 47), and Cascio, W. (1998), *Applied Psychology in Human Resource Management*, Upper Saddle River, NJ: Prentice-Hall, Inc. for more information on this approach.

[22] See Feldt, L.S., & Brennan, R.L. (1989), *Reliability.* In R.L. Linn (Ed.), Educational Measurement (3rd ed.). New York, Macmillan. (pp. 105-146).

[23] U.S. Department of Labor, Employment and Training Administration (2000). *Testing and Assessment: An Employer's Guide to Good Practices.*

[24] Crocker, L. & Algina, J. (1986). *Introduction to classical and modern test theory* (p. 145). Fort Worth, TX: Harcourt Brace Jovanovich.

[25] U.S. Department of Labor, Employment and Training Administration (2000). *Testing and Assessment: An Employer's Guide to Good Practices.*