

# Best Practices and Key Considerations for Written Testing



2450 DEL PASO ROAD, SUITE 220  
SACRAMENTO, CALIFORNIA 95834



(916) 263-3600



# Contents

INTRODUCTION .....	2
DISPARATE IMPACT DISCRIMINATION .....	3
HOW A FINDING OF A DISPARATE IMPACT IS NORMALLY DETERMINED .....	4
THREE TYPES OF TEST VALIDATION METHODS DESCRIBED IN THE UNIFORM GUIDELINES .....	5
JOB ANALYSIS, TEST DEVELOPMENT AND TEST VALIDATION .....	6
SETTING PASS POINTS .....	8
TEST RELIABILITY AND COMPREHENSIVE MEASUREMENT .....	9
SUMMARY AND REFERENCES .....	10



# Introduction

Each year, the Equal Employment Opportunity Commission (EEOC), the U.S. Department of Justice, the Office of Federal Contract Compliance Programs (OFCCP) or plaintiffs represented by private counsel litigate cases under Title VII of the Civil Rights Act of 1964. Many of the cases involve practices, procedures or tests (PPT) used for hiring and promotion. Organizations stand to lose much, not only financially, but also legally and in harm to their reputations.

The litigation could require back pay and interest to the group discriminated against or having the court begin to carefully monitor the employer's practices, procedures or tests used in hiring or promotion. These costs and intrusions don't even include the lost time and costs incurred by the organization to develop the practice, procedure or test.

Written tests are one of the most frequently used and litigated types of PPTs. A written test can be an accurate way of evaluating many candidates with a relatively low time and cost investment on the part of the employer. There are various types of written tests, but, most frequently in the public sector, written tests are composed of multiple-choice items designed to assess candidates' job-specific knowledge or cognitive abilities such as reading comprehension or mathematical ability. Many ask candidates to indicate how they would respond to hypothetical scenarios on the job, known as situational judgement items. This article will examine the legal matters associated with using written tests and best practices that can be used in developing these tests to increase their legal defensibility and accuracy in selecting the best candidates.





# Disparate Impact Discrimination

In order to understand the legal side of side written tests, it's important to have a basic comprehension of Title VII and the two different types of employment discrimination. The goal of Title VII is "that the workplace be an environment free of discrimination, where race is not a barrier to opportunity" (Ricci v. DeStefano, 2009)<sup>1</sup>. The statute is designed to "promote hiring on the basis of job qualifications, rather than on the basis of race or color" (Griggs v. Duke Power, 1971)<sup>2</sup>. There are two types of employment discrimination under Title VII: disparate treatment and disparate impact. A disparate treatment claim accuses an employer of intentionally making employment decisions based on non-job-related reasons such as race. A disparate impact claim challenges an employment decision that appears not to have an intent of discrimination, but still results in less favorable outcomes for those in a protected class.

Written tests are most often litigated under disparate impact discrimination. Hiring and promotional disparate impact cases are evaluated against the federal Uniform Guidelines on Employee Selection Procedures (Uniform Guidelines, 1978)<sup>3</sup>. The Uniform Guidelines were developed in 1978 to set a standard for the development and validation of PPTs and guide the courts in how to try cases with disparate impact in selection, promotion or other employment decisions. The Uniform Guidelines define disparate impact as "a substantially different rate of selection in hiring, promotion or other employment decision which works to the disadvantage of members of a race, sex or ethnic group." If disparate impact has occurred, the next step is for the court to determine whether the PPT is job-related. While valid PPTs may have disparate impacts on a protected group, they may not be considered illegally discriminatory.

Under the Uniform Guidelines, an unlawful employment practice based on disparate impact is established only if the plaintiff (typically an applicant who wasn't hired) demonstrates that the defendant (typically the employer) used an employment practice based on race, color, religion, sex or national origin that caused a disparate impact on the plaintiff, and the defendant failed to demonstrate that the challenged practice, procedure or test was related to the job the plaintiff was applying for. The plaintiff must present initial evidence to prove a case of disparate impact discrimination by first identifying a particular employment practice. If the plaintiff can establish an initial case of discrimination, the defendant (the employer in this case) may either discredit the plaintiff's initial case of disparate discrimination, or the defendant may show that even if there was disparate impact discrimination, the employment practice is "job-related and consistent with business necessity." However, the plaintiff can still win the case if he or she can show that another employment practice was available that would have served the interest of the employer equally well and had less disparate impact.



A photograph of four people (three women and one man) sitting around a table in a meeting, with a yellow overlay. The image is partially obscured by the title text.

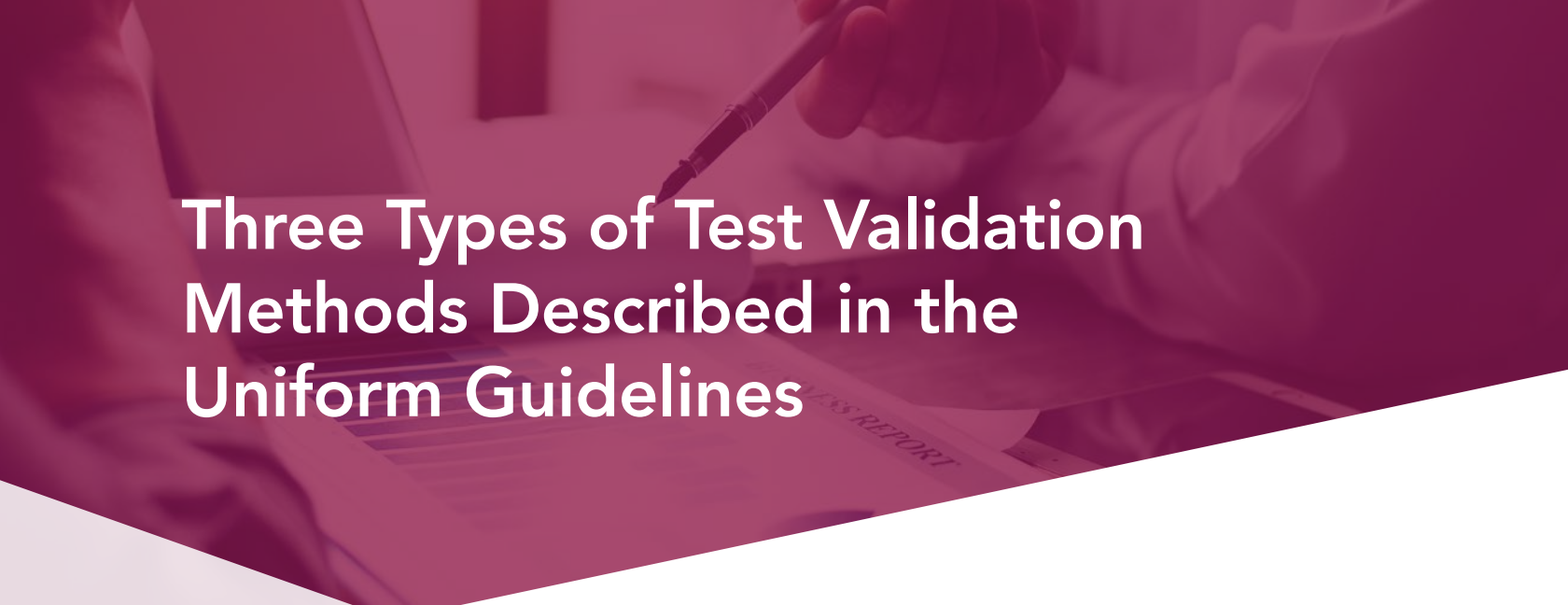
# How a Finding of a Disparate Impact is Normally Determined

The first determination by the courts that must be made is whether a PPT exhibits disparate impact. Two types of measurements can be used to demonstrate disparate impact: tests of statistical significance and the 4/5ths rule.

Statistical significance often refers to evidence that the chances of the minority candidates scoring lower on the selection procedure by chance alone was less than five percent. This is shown through a statistical test demonstrating that the p-value (probability value) is less than .05, meaning that the probability of the score occurring by chance is less than five in 100. The 4/5ths rule stipulates that if the selection rate for any racial group is less than 4/5ths (80 percent) of the selection rate of the group with the highest rate, there is evidence of disparate impact. Different circumstances and different courts vary in determining whether to rely more heavily on the 4/5ths rule vs. statistical significance.

The following is an example of the 4/5ths rule calculation. If there are 100 white applicants and 100 African American applicants and 60 white applicants are offered and accept the job and 40 African American applicants are offered and accept, the selection ratio for whites is 60 percent (60/100), and the selection ratio for African Americans is 40 percent. When the 40 percent selection ratio for African Americans is divided by the 60 percent selection ratio for whites, the result is 66.66 percent, or 2/3rds. Therefore, since the selection ratio for African Americans is less than 4/5ths of the selection ratio for whites, the ratio of whites hired to African Americans hired is in violation of the 4/5ths rule.


The 4/5ths rule stipulates that if the selection rate for any racial group is less than 4/5ths (80 percent) of the selection rate of the group with the highest rate, there is evidence of disparate impact.



# Three Types of Test Validation Methods Described in the Uniform Guidelines

If the court rules that disparate impact has occurred, the next step is to evaluate if the test is both job-related for the position in question and consistent with business necessity as required by the Uniform Guidelines. Job-related and “consistent with business necessity” means that the test was related to the job and was necessary to help the business function effectively.

Three types of test validation methods are described in the Uniform Guidelines for use in determining whether practices, procedures or tests (PPT) are job-related and consistent with business necessity: criterion validation, content validation and construct validation. The Uniform Guidelines provide a set of minimally accepted requirements to follow when conducting validation studies. It does not, however, include an overly specific methodology that needs to be followed for a validation study to be considered potentially legally defensible. The first validation method, criterion validity, provides statistical evidence that those who perform better on the PPT are more likely to be successful on the job, showing that the PPT is job-related. The second validation method, content validity, provides inferential evidence that a PPT is job-related. This is accomplished through an in-depth study of the job (a job analysis) and a series of subject matter expert (SME) opinion surveys. The third validation method that evaluates whether the PPT is job-related is construct validity. Construct validity is demonstrated by identifying the relationships among three things: a specific job-related characteristic, a PPT measuring that characteristic and measures of job performance. Due to the difficulty and complexity of demonstrating construct validity, this method is rarely used in PPT evaluations. Many written tests use a content validation approach, which will be described here. A fundamental first step for demonstrating content validity is conducting a job analysis.



The Uniform Guidelines provide a set of minimally accepted requirements to follow when conducting validation studies.



# Job Analysis, Test Development and Test Validation

When using a content validation method, the job analysis is the foundation that the test ultimately sits on. When examining the test validity, courts will evaluate the extent to which the content of the test is related to the job analysis. A job analysis process involves developing a list of tasks performed on the job and corresponding KSAOs (knowledge, skills, abilities and other characteristics) necessary to perform those tasks. The list of duties is developed from job descriptions and interviewing and observing incumbents that perform the job. A job analysis questionnaire is then created to obtain task and KSA ratings from incumbents and their supervisors. Data from the job analysis questionnaire is then analyzed to help identify the tasks and KSAs that can be used to create a PPT. A technical report is prepared, documenting the methodology and the results of the job analysis.

Since many agencies use written tests, the rest of the paper will focus on written tests. Often, the job analysis is finalized before the multiple-choice test is created. The test developer then frequently converts the job analysis into a test plan document that outlines which KSAs will be assessed by the exam. During the data analysis phase of the job analysis, KSAs can be identified that meet the following criteria:

- Related to the job
- Learned before assignment to the job
- Requiring more training than a brief orientation period
- Having the capability of distinguishing performance to a high or moderate degree
- Required or desirable to perform the job effectively
- Linked to one or more important or critical job duties

This pool of KSAs can then be used as the basis to create test items designed to measure those KSAs. After the test items are created, SMEs then review the test items, identify which KSAs matched the questions and evaluate the questions for difficulty, readability and recommendation for use. SME opinion is critical at this stage in the test development process because it provides validation that the test items are related to the job.

So far in this paper, we have been treating job analysis, test development and test validation as if the organization were starting from scratch. Oftentimes however, organizations are using tests developed by another test developer because developing and validating tests from the ground up can require significant effort and specialized expertise. It's a common misconception that, if a test is developed by a professional test developer, the test is automatically "valid." Just because a test has been created by a test developer and has a foundational job analysis and other validity documentation associated with it, doesn't necessarily mean that it can be considered valid for use in your organization. For instance, while a test developer may have a test created for accountants, not all the areas on that test may be applicable to accounting positions at your organization. As an example, let's say that you are looking to use a written test as part of a selection process to hire an accountant at your organization and you're trying to see if the CPS HR stock [Accountant test](#) would be appropriate.








# Job Analysis, Test Development and Test Validation (cont.)

Our off-the-shelf Accountant test measures the following content areas:

- Accounting Principles and Practices
- Financial Analysis and Practices
- Auditing Principles and Practices
- Mathematical Ability
- Analytical Ability
- Written Communication

This test was created to test for a variety of KSAs that are typically associated with accountants at public sector agencies. However, it might be the case that after having accountants at your organization review the test, you determine that the financial analysis and practices and auditing principles and practices areas are not relevant to the job. Perhaps you work at large organization where other jobs in the finance department perform financial analysis and auditing, and accountants perform a narrower scope of duties. In this case, it may make more sense to use one of our semi-stock exams that allows you to mix and match test questions from our stock exams. Using the semi-stock exam process, you may end up adding items from an interpersonal relations content area from another exam and deleting the financial analysis and auditing content areas. Generally speaking, the degree of test content specialization needed usually becomes greater as the KSAs become more specialized. For instance, if you were administering a test to fire captain candidates that was based on your organization's specific policies and procedures, it might be worthwhile to create a customized test using your organization's specific policies and procedures.

When using a test developed by a vendor, a local content validation study can increase your legal defensibility. A local content validation study uses a specific job and a specific test at your organization. Even if a test and job analysis were created by an independent test developer, conducting a local validation study provides validity evidence for the tests used at your organization. A local content validation study typically involves collecting job analysis ratings from SMEs using the same KSA and duty statements that the test developer used as well as having SMEs review the test items for difficulty and applicability to the job. In addition, a local content validation study can also assist with determining the appropriate pass point to set. CPS HR can further assist you in conducting local content validation with our written tests.



Even if a test and job analysis were created by an independent test developer, conducting a local validation study provides validity evidence for the tests used at your organization.



A photograph of four professionals (three men and one woman) in a meeting, overlaid with a semi-transparent purple gradient. The woman in the center is holding a tablet. The title 'Setting Pass Points' is written in large white font across the middle of the image.

# Setting Pass Points

Another important consideration in terms of evaluating the validity of a written test is how it is used to make a selection decision. There are three primary ways a written test can be used to make selection decisions. If determining how to separate qualified from unqualified applicants is the goal, that test should be used on a pass/fail basis with a minimum passing score set. If the goal is to make distinctions between candidates that are equally qualified but may have slightly different raw scores on a written test, then banding is the approach that should be used.

Banding is a statistical procedure that puts similarly scoring applicants into groups, and each group can be considered as having the same score. Ranking should be used if the goal is to make decisions based on test score on an applicant-by-applicant basis. In other words, the test is used in a rank order fashion to either hire or move applicants forward in the selection process starting from the top of the list moving down. If you're looking to make decisions on applicants based on several selection procedures across many KSAOs that differ in levels of importance, then a weighted or combined selection process can be used. The level of validity and reliability the courts require increase going from pass/fail to banding to ranking (Biddle, 2011<sup>4</sup>). Since how the test is used is so important in determining the overall validity of the test, courts use a high level of scrutiny when examining how the test is used.

Many organizations use a minimum passing score on written tests in deciding which candidates to move forward in a selection process; however, they often set an arbitrary passing score. Section 5(H) of The Uniform Guidelines states "where cutoff scores [minimum passing scores] are used, they should normally be set so as to be reasonable and consistent with normal expectations of acceptable proficiency within the work force." Suppose a multiple-choice written test was created for a position where all the items were completely job-relevant, but an arbitrary 70 percent minimum passing score was set. What evidence is there that 70 percent is the correct cutoff score to accurately identify minimally qualified candidates? Without job expert input regarding what a minimally qualified candidate would score on the test, a 70 percent cutoff can't be justified.



# Test Reliability and Comprehensive Measurement

One of the other elements that courts will examine is test reliability. Section 14 (C)5 of the Uniform Guidelines states: “whenever feasible, appropriate statistical estimates should be made of the reliability of the selection procedure.” The reliability of a test refers to measurement consistency or the degree to which repeated measurement of the same candidate would (hypothetically) tend to produce the same result. If a test has high reliability, a candidate that performs well on the test on Monday should perform equally well if the test is retaken on Tuesday (with no memory of the first test or without practice effects). This is referred to as ‘test-retest reliability.’ A test must be both valid (i.e., accurate) and reliable (i.e., consistent) for the test scores to be meaningful.

## *Testing for a Representative Sample of Job Skills*

While written tests can be an important piece of your selection process, you should also look at your selection process holistically to ensure it is measuring a representative sample of the job skills. The Uniform Guidelines state in 14(C)(1) that “A selection procedure can be supported by a content validity strategy to the extent that it is a representative sample of the content of the job.” It is important to make sure when deciding what selection procedures to use for a given job that you ensure you are testing applicants on enough of the KSAs and not only using a single selection device that tests for a narrow number of KSAs. For instance, let’s say you were conducting promotional testing for police lieutenants and had identified more than 100 KSAs that were critical to performing the job, but you were making promotional decisions based on a written multiple-choice test that only examined knowledge of agency policies and procedures. This test would likely just measure a very small number of KSAs. Many other KSAs might have been related to other critical abilities such as problem-solving and decision-making and “soft skill” KSAs such as oral communication and interpersonal skills. These types of KSAs may constitute a large portion of the job and may be better assessed through assessment center exercises that may involve candidates providing verbal responses to hypothetical scenarios, in-basket exercises or simply including a structured interview in the selection process. Failure to assess for these other KSAs in the selection process would make it more difficult to defend in court if it were ever challenged. In fact, this was one of the primary reasons that a selection process was ruled as not valid in a recent court case. You can see a series of articles published in the OFCCP Digest that examine a disparate impact litigation case using some of the principles discussed in this article here: [Part 1](#) and [Part 2](#).





# Summary

The potential cost of litigation is high and there is value in having a valid selection process that identifies the best candidates for the job. Having a basic understanding of disparate impact and test validation is fundamental to comprehending the legal aspects of testing. There are a variety of topics that are important to consider when conducting job analysis and test validation. It's important to consider the KSAs that should be tested and the level of test customization that would be appropriate given your organization's unique needs. It's also necessary to examine your selection process holistically and not completely rely on written tests to inform selection decisions.

## About CPS HR Consulting

*CPS HR Consulting* is a self-supporting public agency providing a full range of integrated HR solutions to government and nonprofit clients across the country. Our strategic approach to increasing the effectiveness of human resources results in improved organizational performance for our clients. We have a deep expertise and unmatched perspective in guiding our clients in the areas of organizational strategy, recruitment and selection, classification and compensation and training and development.

To discuss how our services and expertise can benefit your organization, reach out to us at:

*CPS HR Consulting*  
2450 Del Paso Road, Suite 220  
Sacramento, CA 95834  
1-800-822-4277

## References

1. Ricci v. DeStefano, 557 U.S. 557, 580 (2009).
2. Griggs v. Duke Power Co., 401 U.S. 424, 434 (1971).
3. Uniform Guidelines – Equal Employment Opportunity Commission, Civil Service Commission, Department of Labor, and Department of Justice (August 25, 1978), Adoption of Four Agencies of Uniform Guidelines on Employee Selection Procedures, 43 Federal Register, 38,290-38,315.
4. Biddle, D. A. (2011). Adverse Impact and Test Validation: A Practitioner's Handbook (3rd ed.). Scottsdale, AZ: Infinity Publishing.







# CPS HR CONSULTING



2450 DEL PASO ROAD, SUITE 220  
SACRAMENTO, CALIFORNIA 95834



(916) 263-3600