CoderPad

CONTENT VALIDATION FOR HIRING ASSESSMENTS

Executive Summary / Introduction

Hiring a new employee requires that staff make decisions that will have a critical impact on the overall success of an organization. Do I offer this person a job? How do I decide who is the best choice amongst all of the applicants? Is this person really as qualified as they appear to be on their resume? To help make these important decisions, organizations rely on a variety of instruments *(e.g., interviews, resumes, references)* to assess the qualifications of job applicants and narrow down the list of potential new hires. Whenever an organization uses an assessment instrument, they are essentially using a test to help them evaluate job applicants and as such, should ensure that the assessment instruments they administer are appropriate *(i.e., valid and reliable)* for helping them select the best people for the job in a legally defensible manner.

In this paper, we'll first explain what constitutes a test in a hiring situation and why it should matter to organizations. Then we'll cover the topic of adverse impact in testing and how it relates to the concept of test validation. Rather than seeing adverse impact at the proverbial "boogeyman" that scares organizations away from testing, it should serve as a reminder to organizations to ensure that the tests they use are valid. This ensures that

any potential adverse impact is job related, meaning that the adverse impact is related to the candidate's abilities to perform / not perform job-related duties rather than systematic bias against a certain group. Next, we'll review the concepts of test validity and reliability and how to determine if a test has evidence of both. Just because a test was created by a vendor does not automatically bestow it with validity. There are specific methodologies and guidelines that must be adhered to in order to properly claim that a test has evidence of reliability and validity. Lastly, we'll provide a summary of a validation study conducted on our CodinGame assessment product to provide evidence of the validity of the tests we offer so that our clients can rest assured that our tests have been rigorously vetted by industry experts in accordance with professional and legal guidelines.

What is a Test and Why Does it Matter?

Despite the critical importance that hiring decisions can have on the overall success of an organization *(e.g., culture, revenue/profit, morale, innovation)*, many companies do not dedicate sufficient attention to how they approach hiring. Their hiring procedures often lack structure or standardization, with an aura of "we'll figure it out as we go." A typical example of this is the unstructured interview where applicants for the same job are asked very different questions in their hiring interview. Essentially, each interview is unique, and the hiring manager can go in many different directions with the questions they ask, based on how a candidate responds. Imagine if this same approach was used to decide the winner of a marathon. Let's suppose that the organizers of the marathon allow individual runners to create and run their own course anywhere within the country. The only stipulation is that the course must be a total of 26.2 miles. In this silly example, some runners would be running at high altitude while others are at low altitude, other runners could create a course that is almost completely downhill running, or the course itself could be all paved for one runner and dirt trails for another. While this example takes an extreme approach to make a point, this is what organizations are doing when they take an unstructured approach to hiring. They are creating a unique "test" for each job applicant and then comparing everyone for the same job as if they had been provided the same opportunity to demonstrate their skills.

Some may argue that this doesn't apply to hiring procedures and that an unstructured approach to hiring allows managers to dive deeper into certain topics with an applicant so they can make a better decision. Unfortunately, current research does not support this opinion. If we look at the data when it comes to comparing unstructured versus structured approaches to hiring, a structured approach improves predictions in hiring *(i.e., quality of*

hire) by more than 50% when compared to an unstructured approach[1]. In summary, using a structured approach where all applicants are given the same test *(such as a standardized coding test)* improves the quality of employees that are eventually hired.

Aside from the organizational benefits, a well-developed structured approach to testing is also beneficial to organizations from a legal perspective. The Uniform Guidelines on Employee Selection Procedures *(UGESP)* provide guidance to employers within the United States when it comes to testing job applicants. Specifically, Section 2B of the UGESP states that the "guidelines apply to tests and other selection procedures which are used as a basis for any employment decision." Simply stated, anytime an organization eliminates people from consideration for a position, they are using a test. Some examples of tests include resume screens, minimum education/experience requirements, hiring interviews, and multiple-choice tests. In all of these examples some job applicants are eliminated from consideration and others are moved forward in the hiring process. Understanding how the UGESP defines a test is important because organizations within the United States that use tests[2] to evaluate job applicants may be required to provide evidence of their validity *(i.e., whether the test is job-related and consistent with business necessity)*. Providing evidence of validity is much easier to do with a structured test as opposed to an unstructured process.

The Relationship of Adverse Impact and Testing

Let's start with test validity and how it relates to adverse impact. According to the UGESP, test validity evidence is only required when there is an adverse impact against a member of a protected group *(e.g., race/ethnicity, gender)*[3]. For example, an organization can ask job applicants who their favorite superhero is and if it doesn't result in adverse impact, they don't have to defend the validity of their interview question. The superhero interview question would likely have no relation to predicting job performance, but legally it would be allowed. So, what is adverse impact and how do you know if your test has it?

Adverse impact is defined as a different selection rate *(i.e., test passing rate)* between groups. In other words, if a higher percentage of males pass your hiring test than females, you may have an adverse impact against females. The question then becomes, how large does the difference in passing rates need to be in order to be considered "different." The

[1] See Kuncel, Klieger, Connelly, & Ones, 2013. Mechanical versus clinical data combination in selection and admissions decisions: A meta-analysis. Journal of Applied Psychology, 98, 1060-1072.

[2] Unless an organization hires every person who applies for a job, they are using one or more tests

[3] Section 1B of the UGESP states, "these guidelines do not require a user to conduct validity studies of selection procedures where no adverse impact results."

federal government uses the 80% (4/5 Rule)[4] to determine if a test has a passing rate that is large enough to be considered different from other groups *(i.e., adversely impacts a protected group)*:

> **80% Test (4/5 Rule)** – A violation of the 80% test occurs when the selection rate of one group is less than 80% of the selection rate of another group. For example, if 100 males complete a test and 75 pass, the male passing rate is 75%. If 100 females complete the same test and 51 pass, the female passing rate is 51%. When we divide 51% by 75%, we get a value of 0.68. In this example, females pass the test at a rate that is 68% that of males and is a violation of the 80% test *(i.e., adverse impact)*. In this case, the test has an adverse impact against females.

It's important to note that adverse impact is the norm and not the exception. Most tests will not result in the same number of people passing by protected group status. It is quite common for organizations to give a test for a job and have more females than males pass, and six months later when they hire again for the same job, more males pass the test than females. In fact, tests that are most helpful in hiring the best employees often result in the most adverse impact against one or more groups. This is so common that we have come to refer to this as the "diversity-validity dilemma."[5] Some testing experts[6] have gone as far as calling the quest for elimination of adverse impact in testing a "Holy Grail" that is unobtainable. In other words, it is unrealistic to expect a test to never have any adverse impact against any group of people.

Given this, a logical conclusion for an organization could be to avoid using tests altogether to eliminate the requirement to validate their hiring tests. However there are a few problems with this line of thinking:

- First, organizations cannot avoid testing. Unless you hire every applicant who applies for a job, you are testing in one way or another. Whatever method you use to give one person a job and not another, that is, in essence, your test.
- Second, this approach fails to view hiring tests from a value-added perspective. Employment tests should assist organizations in hiring the best talent and not just be focused on avoiding adverse impact. Taking the time to ensure the validity of a hiring test adds value to the bottom line of a business through improved hiring decisions *(e.g., more productive employees, less turnover, improved organizational fit)*.

[4] Tests of statistical significance (e.g., Fisher's Exact Test) are also considered when evaluating adverse impact. If the statistical significance test reveals no statistically significant difference, then the disparity could have occurred by chance. If findings are statistically significant, then the disparity has less than a 5% possibility of occurring by chance (standard deviation of 1.96 or greater).

[5] See Ployhart & Holtz, 2008. The diversity-validity dilemma: Strategies for reducing racioethnic and sex subgroup differences and adverse impact in selection. Personnel Psychology, 61, 153-172.

[6] See Arthur, W. Jr., Doverspike, D., Barrett, G. V., & Miguel, R. (2013). Chasing the Title VII Holy Grail: The pitfalls of guaranteeing adverse impact elimination. Journal of Business Psychology, 28, 473-485.

It is important to note that the presence of adverse impact does not mean that a test is biased or not appropriate for use. There are many potential reasons that have nothing to do with the test itself as to why adverse impact occurs. The presence of adverse impact simply triggers the requirement[7] that an organization demonstrates that the passing rate differences by group status are job related and consistent with business necessity *(i.e., are valid)*. This means that an organization should be able to show that their test is a valid test that helps to differentiate the qualified from the unqualified or less qualified job applicants. Showing that a test is valid is much easier to do when the content of the test resembles the job *(i.e., a technical coding assessment)* and each person is given the same test in a structured, consistent manner.

How do you Provide Evidence of Validation for a Test?

There are essentially three approaches[8] for providing evidence of validity for tests used as part of the hiring process: **(1)** criterion-related validity studies, **(2)** construct validity studies, and **(3)** content validity studies. While each approach varies in the steps and evidence collected, all three of the approaches are designed to ensure that there is a link or connection between the test and its relevance to the job. Below are brief descriptions of each approach:

Criterion-Related Validity – This approach requires statistical data showing that performance on the test is correlated with job performance. This involves collecting test score and job performance information to determine if the employees that performed better on the test also perform better on the job *(i.e., does the test predict job performance)*. This type of validity generally requires test score and job performance information from 100+ individuals in order to determine if the relationship is statistically significant.

Construct Validity[9] – This approach requires data showing that scores on one test are similar to scores on another test that measure the same construct. For example, we would expect that if a person completes two different tests that measure basic math, their scores would be similar. That is, our two math tests appear to measure the same knowledge area as indicated by the correlation between the scores, and therefore have construct validity.

[7] Section 3A of the UGESP states that "the use of any selection procedure which has an adverse impact on the hiring, promotion, or other employment or membership opportunities of members of any race, sex, or ethnic group will be considered to be discriminatory and inconsistent with these guidelines, *unless the procedure has been validated in accordance with these guidelines*."

[8] See UGESP Sections 5A and 5B for a brief overview on the types of validity. There are additional approaches to providing evidence of validity that are not mentioned in the UGESP and/or in this report (e.g., validity generalization, synthetic validity, transportability).

[9] Construct validity is not often utilized in practice because the UGESP Section 14D(3) requires that criterion-related evidence of a validity be provided as well for this approach. However, criterion-related validity does not require evidence of construct validity. Hence, construct validity is a more onerous process.

Content Validity – This approach focuses on the degree to which the content of the test resembles what is done on the actual job. In other words, does the test measure things that a person will be expected to use or demonstrate when working in the job? This approach requires people who work in and/or supervise the job to provide ratings of the similarity of the test to what is performed on the job. This typically involves an analysis of the job(s) a test will be used for, where ratings are collected about the important work behaviors and knowledge, skill, and ability *(KSA)* areas. For example, if a test of basic math is used to hire cashiers at a store, a job analysis would determine how often cashiers use basic math on the job and how important math is to successful job performance for cashiers. We can then establish a connection or linkage between the math questions and the important KSAs to establish the content validity of the test. This type of validity evidence can be completed with as little as seven employees and is the most common approach used to provide evidence of validity.

Reliability and Test Validity

Test reliability is a necessary component to establishing test validity. Reliability is about how consistent a test measures certain attributes and validity is about how accurate the test measures the same attributes. The interaction between the two can be illustrated by the following example involving a scale to weigh fruits and vegetables. You pick up an apple at the grocery store and place it on the scale to weigh it. You take it off the scale and then place it back on the scale to weigh it again. You do this several times. A reliable scale will provide the same or very similar weight for the apple each time you place it on the scale. However, just because the scale is reliable does not mean that the weight it provides you is valid. If the real weight of the apple is two ounces and the scale consistently indicates that it weighs eight ounces, the scale is reliable but not valid. A valid scale will be both consistent and accurate.

The same is true for an employment test. If we were to give the same or very similar version of the test to a job applicant a month apart, a reliable test will result in a very similar overall score for the applicant both times. A valid test will provide a score that is an accurate measure of the job applicant's knowledge or ability being assessed by the test. It is critical that tests used for hiring are both reliable *(i.e., consistent)* and valid *(i.e., accurate)*.

Validation of the Tests in CoderPad's CodinGame Assessments Product

At CoderPad, our goal is to provide our clients with high-quality, structured tests that provide reliable and valid insights about the technical expertise of job applicants through the CodinGame assessment product. In adherence to the UGESP, we relied on a content validation strategy to provide reliability and validity evidence for our tests. Subsequently, we will provide a brief overview of the content validation process we followed for the tests we offer; however, those interested may request the full technical report which includes all of the study details in compliance with the content validity requirements of Section 15C of the UGESP. This analysis was conducted by ioPredict, an independent third-party.

As the foundation for content validity, we conducted job analyses of 15 of the most common job titles within the Software Engineering field. For each job title, we had a group of experts who work in that job provide us ratings on the important work behaviors and knowledge, skill, and ability areas for the job. Below is a list of the 15 job titles for which job analyses were conducted.

- Back-end developer
- BI / Data analyst
- Data engineer
- Data scientist
- Database administrator
- Database engineer
- Dev Ops
- Embedded system engineer
- Front end developer
- Full stack developer
- Mobile developer
- Network Admin
- QA engineer
- Security engineer
- System admin

For the assessments related to these fifteen jobs, the job analysis evidence showed the content to be valid.

The expert's ratings on the job analysis survey provided a blueprint of the most critical areas to be measured for each job title as part of the hiring process for the position. Their input into what is done on the job formed the foundation and rationale for the types of questions included in the test for that particular job title. From the list of 15 common job titles, we collected content validity evidence for the three job titles most commonly used for testing: **(1)** Full Stack Developer, **(2)** Front End Developer, and **(3)** Back End Developer. For each of these job titles we invited 30 experts to complete the test as if they were a

job applicant and provide content validation evidence/feedback on the test to ensure it is an accurate measure of the knowledge, skills, and abilities required for the job. The table below provides the validation questions asked of each of the 30 experts for each test.

Validation Survey Questions

1) Do you feel this test is relevant to the job?
2) Does the test cover technical knowledge or skill areas that a job applicant should know prior to being hired?
3) Would you be more likely to offer a person a job with a higher score on this test compared to a person with a lower score on this test?
4) Were the test instructions and the test itself clear and understandable?
5) Does the test require you to demonstrate technical capabilities that are similar to those you would need to demonstrate on the job?
6) Is the test fair to all groups of people in terms of race/ethnicity, gender, or age?
7) If you answered no to any of the questions above, please provide an explanation below. Also, if you have any general comments about the test, you may enter them here.

The results from the subject matter experts indicated strong agreement of the content validity or job relatedness of the tests. Experts who work in these job titles confirmed that the questions asked in the test are relevant to the job and require individuals to demonstrate technical capabilities that are similar to those required on the job.

In addition to confirmation from the experts that the tests measure important job-related knowledge, ability, and skill areas, the tests were also found to be reliable *(consistent)*. Specifically, the statistical reliability coefficient *(Cronbach's Alpha*[10]*)* for each of the tests was above 0.70 *(Back End = 0.76, Front End = 0.72, Full Stack = 0.78)*. The U.S. Department of Labor interprets reliability coefficients of greater than rxx = 0.70 as being "adequate." In summary, based on the feedback from 90 experts *(30 for each job title)* who work as Full Stack, Front End, or Back End Developers, the three tests were found to be both reliable and valid.

[10] This is a measure of internal consistency that is commonly used to report the reliability of a test

CodinGame Assessments and Adverse Impact

In addition to collecting evidence of content validity, we evaluated a CodinGame Software Engineer Intern test for adverse impact against females and males. A current customer using the CodinGame assessments product provided test scores for 105 job applicants *(46 males and 59 females)*; the results are presented in the table below. Overall, average test score differences by gender were moderate in size[11]. Adverse impact was evaluated at 10-point comparative score intervals, ranging from 90 on the high end to 30 on the low end. Adverse impact results are presented for both the 80% *(or 4/5th)* rule of thumb and in terms of standard deviation differences according to statistical significance *(Fisher's Exact Test[12])*. There are some instances of adverse impact (red shading) at different cutoff options *(i.e., passing scores)*; however, only three of the seven potential cutoff scores violate both the 80% rule and are statistically significant. As mentioned previously regarding most tests, adverse impact is the norm and not the exception. The levels of adverse impact found for this test are not severe and, given the amount of content validity evidence in support of the test, a client can confidently stand behind and defend their use of the test in the event of a legal challenge as job related and consistent with business necessity.

Test Descriptive Statistics by Gender

	Raw Score (%)		Comparative Score	
	Average	**Std Dev**	**Average**	**Std Dev**
Male (n = 46)	55.35	22.58	58.00	31.59
Female (n = 59)	46.76	23.44	46.14	31.93

Adverse Impact by Gender at Different Cutoff Scores

Cutoff Score (Comparative Score)	**Percentage Passing at Cutoff**		**Adverse Impact Results**	
	Male	**Female**	**80% Rule**	**Std Dev**
90	21.74%	13.56%	62.37%	1.15
80	26.09%	18.64%	71.44%	0.81
70	54.35%	27.12%	49.90%	2.89
60	58.70%	35.59%	60.63%	2.25
50	63.04%	40.68%	64.53%	2.24
40	67.39%	57.63%	85.52%	1.09
30	73.91%	66.10%	89.43%	0.94

[11] The d-value or average mean scores difference between males and females is -0.37.

[12] This is a statistical significance test that was explained previously.

Conclusion

Hiring decisions matter, and the tests that organizations rely on to help make those decisions matter as well. Organizations benefit immensely from using structured tests that are both reliable and valid, as valid tests have a demonstrable impact on the bottom line and overall quality of new hires. At CoderPad, we are committed to providing our clients with fair, reliable, and valid tests to assist them in hiring the most qualified applicants. We have invested a significant amount of time and resources to ensure that our tests are valid. We have partnered with experts in the industry to develop and review our tests to confirm their relevance to the work done on the job and to ensure their accuracy in assessing job-related technical knowledge. We are confident in the quality of our tests and look forward to supporting your organization in hiring the most qualified job applicants.