HACKERRANK

Technical Skills Testing

Maximizing Pre-Employment Assessment Effectiveness and Legal Defensibility

www.hackerrank.com

Table of Contents

01 Executive Summary

Hiring is bit like playing poker. In poker, you look across the table at your opponent trying to determine the strength of his/her hand based on the information you have (e.g., the cards in your hand, the cards on the table) as well as his/her body language and other non-verbal cues. Ultimately, you decide whether to bet or fold.

In hiring, you sit across the table from the candidate trying to determine his/her capabilities to perform the job; you base this determination on the information you have in front of you (e.g., a résumé, what the candidate says in an interview) as well as the candidate's body language and other non-verbal cues. Ultimately, you decide whether to hire the candidate or move on to the next one.

What makes these two decisions drastically different, however, is that you have the ability to structure the hiring process in ways that yield a wealth of information about candidates' capabilities and likely performance on the job, if hired. You do not have this same luxury in poker.

Effective hiring processes have three basic phases: **Recruit, Screen, and Select.**

1. **Recruit:** The objective is to get a diverse, qualified pool of candidates to apply for the job
2. **Screen:** The objective is to gather structured information about candidates in an effort to narrow the funnel of candidates
3. **Select:** The objective is to make hiring decisions and offer jobs to those individuals who are most likely to be successful on the job

While all three phases of the hiring process are equally important, this paper focuses on the *screen* phase.

A. Problem

For technical hiring (e.g. Software Developers), companies today generally use a combination of resume screens, phone screens, and in-person whiteboard sessions. These interview sessions are often **unscripted, undocumented,** and lack consistency and objectivity. Simply stated, they are **unstructured and unstandardized.** Each candidate often has a different set of questions and experiences while interviewing for the same job. In addition, the initial screening is done by a non-technical recruiter who has difficulty ascertaining a candidate's true skill level. For those organizations that utilize pre-employment assessments, quite

often the appropriate validation research is not conducted, which places organizations at significant risk. As a result of these problems (i.e., lack of structure, standardization, validation), organizations often find themselves in the midst of costly litigation and EEOC discrimination suits.

B. Solution

Properly developed and validated pre-employment screening systems are the answer to these organizations' woes. Evidence for validity is based on demonstrating a strong linkage between the content of the selection procedure and important work behaviors, activities, worker requirements, or outcomes on the job. Through validation, organizations build a data-drive case illustrating the relevance and job-relatedness of their screening systems for their jobs.

The remainder of this paper is organized as follows. First, we begin with an overview of pre-employment assessments and examine the effectiveness of different strategies. We will see, for example, that the closest thing to a **silver bullet** in pre-employment screening for computer programming jobs is a work sample—some measure of an individual's ability to demonstrate actual job skills and knowledge, such as an assessment that requires computer programming job candidates to write actual code.

We then explore the concept of validation—what it means, why organizations should be concerned about it, the role that it plays in legal defensibility, how to establish it, and what happens when organizations do not have it.

Bottom line: Employers are at significant risk if they cannot provide sufficient validity evidence for their employment assessments and processes.

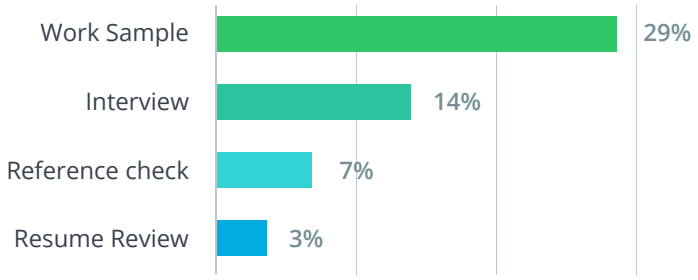
A case study describing HackerRank's assessment content and their multi-level approach to establishing validity is described next, followed by a series of tips and best practices for organizations that want to hire the best talent in a legally defensible manner.

02 Overview of Assessments

Organizations screen job candidates in a number of different ways—they review résumés, they interview candidates, they check candidates' references, and sometimes they require that candidates complete some sort of pre-employment assessment. Each of these screening methodologies is utilized in an effort to learn useful information about candidates, but they differ widely in their utility.

To illustrate, ***Figure 1*** displays the percentage of variance in job performance explained by different screening procedures across industries and job types[1]. Some people demonstrate excellent job performance, others demonstrate acceptable job performance, and others demonstrate sub-par job performance. The goal of any screening tool is to predict this variability in job performance—we measure candidates in different ways in an attempt to predict how they would perform on the job if hired.

Figure 1: % Variance Explained by Different Screening Procedures



Consider the process of screening résumés; most organizations look at applications and résumés submitted by candidates in an attempt to determine whether they have "what it takes" to perform the job. Typically, these reviews consider where someone went to school, what degree(s) he/she earned, where he/she has worked and for how long, etc. In reality, while such reviews might serve as a proxy for whether candidates meet minimum job qualifications, they only account for about 3% of the variance in actual job performance—this is depressingly low. It means that 97% of what determines whether someone will be an excellent performer vs. a sub-par performer on the job is not measured by simply reviewing résumés.

[1] Schmidt, F.L. & Hunter, J.E. (1998). The validity and utility of selection methods in personnel psychology: Practical and theoretical implications of 85 years of research findings. *Psychological Bulletin, 124 (2), 262-274.*

A similar story unfolds with typical reference checks, though, admittedly, they explain more than twice the variance as résumé reviews—but that is not saying much given that they still only explain about 7% of the variance in job performance.

Now consider the interview, which is the most frequently used screening tool across organizations, jobs, and job levels—most hiring managers would consider hiring a new employee without first conducting some sort of interview to be unthinkable[2]. The typical employment interview accounts for about 14% of the variance in actual job performance. This is a marked improvement (nearly five times) over résumé reviews.

The closest thing to a **silver bullet** in pre-employment screening, however, is a work sample—some measure of an individual's ability to **demonstrate** actual job skills and knowledge, such as an assessment that requires computer programming job candidates to write actual code. Such work samples, assuming they are job-related, account for about 29% of the variance in actual job performance on average.

A pragmatic interpretation of these numbers might leave one thinking "Even predicting 29% of the variance still means that the vast majority—71%—of the variance in job performance is left unmeasured…this cannot be good." True, but three **very important things.** First, hiring processes typically consist of multiple steps and while not completely additive in their utility in predicting job performance, there is incremental value in every selection procedure, assuming they are job-related. An organization would not simply administer a work sample and base a hiring decision on that tool alone. Instead, it would likely construct its hiring process to include multiple steps, each of which adds value and provides insight into candidates' capabilities.
Second, no hiring process or single screening tool is a perfect indicator of candidates' capabilities and behavior on the job if hired. Job candidates are humans and not computers, and we cannot predict with 100% accuracy what another human being will do—ever. The key to successful hiring is gathering consistent data across candidates on their respective abilities to perform the job, and then making informed decisions. Properly validated pre-employment assessments that are rooted in job requirements help us do just that.

Third, not all assessments are created equally, and some assessments do a better job of predicting success than other assessments of the same nature. Take two work samples, for example: one that asks candidates to program a fairly easy coding challenge that is much

[2] Huffcutt, A.I., & Culbertson, S.S. (2010). Interviews. In S. Zedeck (Ed.), *APA Handbook of Industrial and Organizational Psychology (volume 2), 185.*

easier than the job, and one that asks candidates to program a more complex coding challenge that mirrors the work more closely. The easy coding challenge will not be nearly as predictive of success as the more complex challenge—if everyone gets a coding challenge correct (i.e., the easy challenge) it cannot be expected to predict job success. Thus, the 29% noted above for work samples is an average of all work samples. Work samples that mirror the job in both content and complexity can be expected to exceed this level of prediction considerably.

03 Overview of Validity

An assessment's validity is the extent to which scores based on the assessment relate to individuals' ability to perform some job of interest. For example, we might expect that an individual's performance on an assessment of computer programming knowledge and skills would be predictive of—relate to—his/her ability to perform a computer programming job. Evidence of this assessment's validity would, therefore, hinge on the extent to which data exist to demonstrate this linkage; such evidence is critical to ensuring that the assessment is both effective and legally defensible.

Effectiveness

Continuing with the example of the computer programming knowledge and skills assessment, how do we know how effective the assessment actually is? The answer to this question relates to the **strength** of the relationship between scores on the assessment and job performance. Think back to the discussion about the percentage of variance in job performance explained by different screening procedures (i.e., ***Figure 1***). This percentage of variance explained is a measure of the strength of the relationship between scores on the assessment and job performance. Interviews and work samples may both be **valid** screening tools for a particular job, assuming that they measure attributes that relate to job success, but work samples are typically more **effective** measures of candidates' ability to perform the job[3].

Legal Defensibility

Filing a lawsuit against an employer in the United States is easy—very easy. Thankfully, the courts are fairly good about dismissing frivolous claims. Employment discrimination claims revolving around Equal Employment Opportunity (EEO) violations, however, are taken very seriously.

Title VII of the *Civil Rights Act of 1964* prohibits employment discrimination based on race, color, religion, sex, or national origin. Further, the Act prohibits both **disparate treatment** and **disparate impact.**

[3] The reader should note that there are a number of things organizations can do to increase the effectiveness of their employment interviews; for a detailed review, see Chambers, B.A. & Arnold, J.D. (2015). Using technology to improve the interview as a section tool. *Personnel Assessment and Decisions, 1(1), Article 7.*

1. **Disparate Treatment:** Treating one group of candidates differently than another (e.g., adjusting scores or using different cutoff scores for different groups of candidates)
2. **Disparate Impact** (aka Adverse Impact): Using procedures that (1) affect—impact—one group of candidates differently than another, and (2) are not job-related and consistent with business necessity

Title I of the American with Disabilities Act and the Age Discrimination in Employment Act extend these same principles to the prohibition of employment discrimination based on disabilities and age (40 and over), respectively.

In the event of an employment discrimination suit surrounding an alleged EEO violation, the courts will request adverse impact statistics from the employer. If these statistics reveal evidence of adverse impact, then the burden of proof shifts to the employer to demonstrate that the assessment is job-related and consistent with business necessity—i.e., valid.

A common and straightforward (though by no means only) measure of adverse impact is the four-fifths rule of thumb: If the pass rate for the minority group is less than four-fifths (.80) the pass rate for the majority group, evidence of adverse impact exists. For example, if 40% of minority group members pass an assessment and 80% of majority group members pass the same assessment, the ratio is 40% divided by 80%, or .50. In this example, evidence of adverse impact would exist.

Adverse impact is common and to be expected on cognitive assessments such as our hypothetical work sample that we have used as example throughout this paper. However, adverse impact is not illegal so long as the employer can demonstrate that the assessment is job-related and consistent with business necessity—the assessment measures candidate attributes that are critical to performing the job for which the assessment is used. If the employer cannot demonstrate that the assessment is job-related and consistent with business necessity, well, that's a different story.

Evaluating Risk

Countless court cases underscore the importance of proper assessment validation, one of the most notable being Griggs v. Duke Power Company[4] in the early 1970s. At the time *the Civil Rights Act of 1964* was enacted, Duke Power Company had 95 employees in a North Carolina facility, of whom 14 were African American. As part of its hiring process, the company required candidates to pass a general aptitude (i.e., cognitive ability) assessment—passed

[4] Griggs v. Duke Power Co., 401 U.S. 424 (1971).

by approximately 6% of African American Candidates and 58% of Caucasian candidates (ratio = .10, which is less than the four-fifths rule of thumb described previously). A group of African American job candidates sued Duke Power Company claiming that its use of the assessment violated the *Civil Rights Act of 1964*. Unfortunately for Duke Power Company, the organization could not supply adequate proof of the assessment's job-relatedness—validity—and Duke Power Company lost the case.

More recently, Target Corporation paid $2.8 million in 2015 to resolve an EEO discrimination finding that resulted from its use of three pre-employment assessments for exempt-level professional positions[5]. In this case, the assessments were found to yield adverse impact against females and ethnic minorities. Just like Duke Power Company, Target Corporation could not provide adequate evidence of the assessments' validity, and they lost the case.

Hitting even closer to home is the 2017 U.S. Department of Labor case against Palantir Technologies Inc. for alleged systemic hiring discrimination against Asian candidates for engineering positions at the company's Palo Alto facility[6]. In this case, the company was unable to provide sufficient evidence that it was **not** discriminating during the résumé review and initial telephone interview processes, and agreed to pay $1.6 million in back wages and other monetary relief to class members.

Bottom line: Employers are at significant risk if they cannot provide sufficient validity evidence for their employment assessments and processes.

Establishing Validity

In 1978, the Civil Service Commission, the Department of Labor, the Department of Justice, and the Equal Opportunity Commission jointly adopted the Uniform Guidelines on Employee Selection Procedures[7] (aka Uniform Guidelines) to establish a common set of standards regarding the use of employment assessments and other selection procedures in the United States. These standards cover a range of topics such as adverse impact, assessment validation, and record-keeping guidelines, and they document a uniform federal position on the prohibition of discriminatory employment practices and procedures. Although the

[5] U.S. Equal Employment Opportunity Commission (2015, August 24). Target Corporation to Pay $2.8 Million to Resolve EEOC Discrimination Finding. Retrieved from https://www.eeoc.gov/eeoc/newsroom/release/8-24-15.cfm.

[6] United States Department of Labor (2017, April 25). US Department of Labor Settles Charges of Hiring Discrimination with Silicon Valley Company. Retrieved from https://www.dol.gov/newsroom/releases/ofccp/ofccp20170425.

[7] Equal Employment Opportunity Commission, Civil Service Commission, Department of Labor, & Department of Justice (1978). Uniform guidelines on employee selection procedures. Federal Register, 43(166), 38290-38315.

Uniform Guidelines are not legislation or law, they are repeatedly referenced by the courts as a source of technical information and are often given deference in litigation concerning employment issues.

The *Uniform Guidelines* outline three strategies for establishing the validity of an assessment tool or procedure: **Content Validity, Criterion Validity**, and **Construct Validity.**

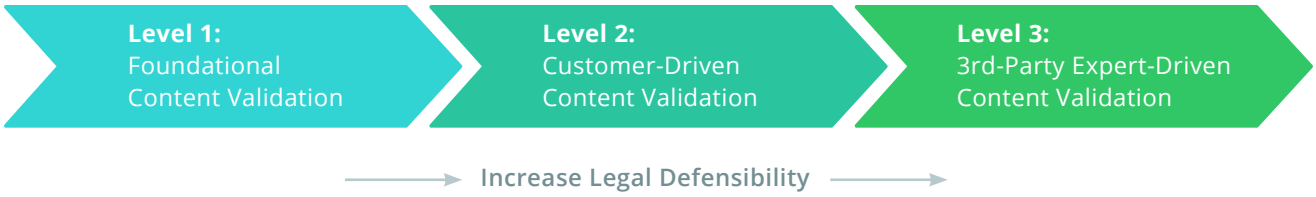
1. **Content Validity:** Involves a demonstration, through structured judgments and ratings, that the content of the assessment tool or procedure is representative of importance aspects of the job; in other words, the assessment measures characteristics or attributes (e.g., knowledge, skills) that are important for job success
2. **Criterion Validity:** Involves a statistical demonstration of the relationship (i.e., correlation) between the assessment tool or procedure and some outcome measure of interest, such as job performance; criterion validity studies can be conducted utilizing both concurrent designs in which case assessment and job performance data is collected at a single point in time, and predictive designs in which case assessment data are collected at one point in time and job performance data are collected at a later point in time
3. **Construct Validity:** Involves a statistical demonstration that the assessment tool or procedure relates strongly with other assessment tools or procedures that are designed to assess the same attributes or characteristics

No single validation strategy is universally better than the others. Instead, the specific situation (e.g., assessment tool, job, number of employees and/or candidates) often drives decisions regarding the specific strategy utilized. Assessment tools measuring technical knowledge and skills, for example, are often validated using a content-oriented strategy whereas measures of personality are often validated using a criterion-oriented strategy.

04 HackerRank's Efforts

HackerRank is a skills-based tech hiring platform that helps companies evaluate technical skills more effectively through its innovative assessment platform. The HackerRank assessment enables companies to presents candidates with coding challenges that they must solve by writing code that is then automatically scored by HackerRank's proprietary algorithms. Companies set a cutoff score to determine the baseline of technical skills required for their jobs, and candidates who meet or exceed these thresholds move on to subsequent stages of the companies' hiring processes Thus, through HackerRank's assessment platform, companies are able to improve the quality of candidates in their candidate funnel who ultimately make it to an interview, thereby making better use of interviewers' limited time.

HackerRank has conducted significant research and due diligence to ensure that its assessment content is both job-related and highly impactful. In partnership with Polaris Assessment Systems, Inc., a firm of industrial-organizational psychologists specializing in developing and validating effective and legally defensible employment screening systems, HackerRank takes a multi-level approach to validating its assessments—**Foundational Content Validation, Customer-Driven Content Validation, and Third-Party Expert-Driven Validation**—each of which is described in more detail below.



Level 1: Bias and Sensitivity Review

Alpine is committed to conducting a Cultural Sensitivity review of HackerRank's assessment items in order to minimize bias in the performance of these items. The goal of fairness in testing should be approached by making sure that test properties are as barrier-free as possible and fair for all groups of test takers, despite differences in characteristics including,

but not limited to, disability status, ethnic group, gender, regional background, native language, race, religion, sexual orientation, and socioeconomic status.

In accordance with the guidelines of the Standards for Educational and Psychological Testing, specifically Standard 3.2: "Test developers are responsible for developing tests that measure the intended construct and for minimizing the potential for tests' being affected by construct-irrelevant characteristics, such as linguistic, communicative, cognitive, cultural, physical, or other characteristics."

In compliance with this standard and to help ensure the fairness of the exam, Alpine staff with specific training and experience in this area conducts careful review of all items. This includes the following:

1. Each item is considered with regard to culturally sensitive content that could
2. An established checklist is used in evaluating each item. The checklist contains questions such as "Does the item contain vocabulary or language that has different meaning for various subgroups of candidates?"
3. If and when any issues are identified through this review, Alpine staff offers suggested revisions to items based on industry standards and best practices in exam development.
4. A summary report is provided in addition to suggested item revisions, providing documentation of results of this review.

Level 2: Customer-Driven Content Validation

While the **Foundational Content Validation** evidence collected by HackerRank certainly documents the relevance of its assessment content for a number of common computer programming jobs across organizations, it does not demonstrate a linkage between assessment content and an organization's **specific** jobs, which may be desired by an organization and is certainly a best practice. Level 2 validation equips client organizations with the tools needed to conduct its own content validation study on HackerRank's assessment content (*i.e., Customer-Driven Content Validation*). Thus, organizations that desire additional validation evidence for their jobs specifically and feel comfortable conducting their own content validation studies can do so without engaging a third-party to lead the validation process.

Level 3: Third-Party Expert-Driven Validation

For client organizations that would like a third-party's involvement in the validation process, HackerRank's Level 3 approach to validation involves Polaris Assessment Systems, Inc. as a third-party expert. With this approach (*i.e., Third-Party Expert-Driven Validation*), rather than the client driving the content validation process as described in Level 2, Polaris leads all data collection, data analysis, and report-writing activities.

In addition to content validation studies, Polaris also assists HackerRank's clients with criterion validation studies which provide additional confirmation that the assessment results in hiring the right individuals for the job. With these studies, job performance information (e.g., supervisory ratings of effectiveness, productivity metrics) for individuals who have completed the assessment are gathered and analyzed against HackerRank assessment scores. Such studies would be expected to reveal that the better individuals perform on the HackerRank assessment, the better their job performance. Thus, such results not only speak to the assessment's utility as a selection tool, but they also serve as another level of validation support.

05 Tips & Guidance for Implementing Assessments

So, what should an organization do if it wants to hire the best talent and do so in a legally defensible manner? Certainly, properly validated pre-employment assessments can add significant value to organizations in this regard. This section provides tips and guidance for implementing pre-employment assessments in your hiring process.

Determining What to Measure/Assess

Pre-employment assessments differ widely in what they measure. Before choosing a pre-employment assessment, the buyer must first ask him/herself, "What do I want this assessment to tell me?" At the most fundamental level, the answer to this question is "how well candidates will likely fit into the job and be able to perform the job effectively." But a more specific answer comes from a detailed analysis of the job and a determination of the knowledge, skills, and other personal characteristics required to perform the job effectively. Pre-employment assessments are not one-size-fits-all. They differ widely in what they measure, and therefore, what you should expect them to tell you about candidates.

Evaluating Assessment Options

After determining what the assessment should measure, you then face the task of choosing an assessment. Certainly, the assessment that you choose should measure the characteristics identified as important for effective job performance. But how do you know how well the assessment measures these characteristics? The answer to this question is two-fold.

1. **Reliability**
   In the world of pre-employment assessment, **reliability** refers to the consistency, or precision, with which an assessment measures some desired construct or characteristic. Imagine that you cut a rubber band in one spot, thus allowing you to stretch it flat. Now imagine that you have a standard, 12-inch ruler. Which tool, the rubber band or the ruler, would be more reliable in measuring the length of an object? Answer: the ruler. If you try to measure the length of the object with the rubber band, you will most likely get a different answer every time based simply on how hard you stretch the rubber band. The ruler, however, will give you a consistent answer each time.

In pre-employment assessment, **reliability** is most often measured in one of two ways, both of which can be acceptable from the perspectives of legal guidelines and professional standards assuming that the assessment developer has done things properly. The first way to quantify an assessment's reliability is to give the same assessment to the same people on two different occasions and evaluate whether they score about the same on both occasions. This form of reliability, known as test-retest reliability, certainly has its problems which is why most assessment developers stay away from it. Depending on the characteristic(s) being measured, we might expect an individual's score to be higher the second time simply because he/she has seen the assessment before (i.e., a practice effect).

The more common method to quantify an assessment's reliability is known as internal consistency reliability, communicated on a scale from 0.00 to 1.00. In essence, this form of reliability measures how consistently people answer questions that are designed to measure the same characteristic. Assume that we have an assessment with 15 items that are all designed to measure candidates' knowledge of java. High levels of internal consistency reliability (i.e., professional standards call for 0.70 and above) tell us that how candidates respond to any one of these 15 items is highly consistent with how they respond to the other 14 items. This, in turn, tells us that our measurement of java knowledge is consistent and precise. That said, having high reliability alone does not give us proof that we are actually measuring java knowledge. We just know that these 15 items all measure the same characteristic, whatever that characteristic may be. Proof that we are actually measuring java knowledge is the assessment's validity.

2. **Validity**
   A previous section of this paper described multiple ways to validate assessments. Regardless of what an assessment provider may publicize about its assessments' validity, you should exercise due diligence and review the provider's technical documentation. If you feel that you do not have the level of expertise required to review such documentation, seek the assistance of an employment attorney and/or industrial-organizational psychologist to assist in this review.

Measuring an Assessment's Utility/Effectiveness Over Time

After implementing an assessment, it is important to measure its utility and effectiveness over time—if the assessment does not help you make better hiring decisions, why are you using it? One method of examining an assessment's utility and effectiveness is to conduct a predictive criterion validation study. For example, administer the assessment to candi-

dates before they are hired, and track their job performance. At some later point in time, such as 12 months after hire, collect information that speaks to these same individuals' job performance—e.g., supervisory ratings, lines of code produced, compiling errors—and correlate this job performance information with assessment scores. Statistically significant positive correlations between assessment scores and future job performance would provide evidence that the assessment can actually predict job performance.

Another strategy to measure an assessment's utility and effectiveness over time is through return on investment (ROI) analyses. Simply stated, ROI can be calculated using the following formula:

$$\frac{\text{Benefit - Cost}}{\text{Cost}} \text{ x 100= \% Increase}$$

For example, assume that an assessment costs $1,000 annually and that it produces $5,000 in increased results (e.g., increased productivity). ROI would be 400%. Do the benefits of the assessment or process outweigh its costs? Could the assessment or process be tweaked in some way to provide better ROI? These are important questions to consider.

Monitoring Adverse Impact Statistics

Finally, on some regular basis, which may mean quarterly or annually depending on candidate volume[8], organizations should measure adverse impact. As noted above, there are multiple methods for measuring adverse impact, one of which is the four-fifths rule of thumb: If the pass rate for the minority group is less than four-fifths (.80) the pass rate for the majority group, evidence of adverse impact exists.

It is important to note that the four-fifths rule of thumb is not the only way to measure adverse impact; the other formulas, however, are much more complicated and beyond the scope of this paper. If you feel that you do not have the level of expertise required to conduct a thorough review of your assessments' adverse impact, seek the assistance of an industrial-organizational psychologist to assist in this review.

[8] At low candidate volumes, adverse impact statistics can be unreliable and meaningless; organizations should calculate adverse impact when they have at least 50 candidates per sub-group being compared (e.g., males vs. females).

06 Conclusion

Properly validated pre-employment assessments can add significant value to organizations' hiring processes. They serve as a standardized method of evaluating candidates on characteristics that have been deemed important for job success. That said, no hiring process or screening tool is perfect, and no one—no matter how rigorous the hiring process or tool—can make perfect hiring decisions every single time. Sometimes you will hire candidates who turn out to be a poor fit for the job, and other times you will pass on candidates who would have been great employees if hired. The key to hiring, however, is stacking the deck in your favor so that you minimize these hiring errors. Properly validated pre-employment assessments that are rooted in job requirements help us do just that.