IBM® **Smarter Workforce Institute**

Evaluating Assessments in the Age of Big Data and AI

Nigel Guenole, Ph.D., Sheri Feinzig, Ph.D.

*With contributions from: Sean Keeley, Amanda Klabzuba, Ph.D.,
Kevin Impelman, Ph.D., Jeff Labrador, Ph.D.*

> *“Innovation in assessment that enhances the hiring experience is more critical than ever in talent acquisition, but it needs to come with sound theory, proper design and unbiased prediction.”*
>
> — Robert Gibby, Chief Talent Scientist, IBM

Executive summary

Approaches to assessing job candidates and employees in talent management are experiencing rapid change. Much of the change is technologically driven by developments in machine learning and artificial intelligence (AI). For instance, automated video interview scoring, social media scraping, and gamification are relatively new methods deployed as assessments in talent acquisition. These methods are fast, can be more engaging for candidates, and represent an important evolution in testing beyond 20th century approaches. But, however compelling these new methods, practitioners should not forget they are selection tests. They therefore need to meet stringent criteria related to group differences, bias, and standards for reliability and validity (e.g., evidence that they predict job performance or quality of hire). In this paper, we give an overview of testing in talent acquisition and offer four specific guidelines for assessment practitioners evaluating new selection methods.

What is psychological testing and assessment?

Psychological testing describes the process of measuring people’s personal attributes, such as ability, personality, values, interests, knowledge, and technical and behavioral skills. When the process involves integrating information about individuals from different sources, practitioners often refer to assessment. Assessment is a broader process of forming a coherent overall picture of an individual using testing as well as other sources of information. But there are no firm boundaries, and many people use the terms assessment and testing interchangeably.

We should not underestimate the effect that scores from psychological testing and assessment processes have in society. Strong testing and assessment scores can open pathways to educational and occupational opportunities, such as admission to prestigious schools and well-rewarded jobs. Opportunities that follow from strong performance on tests can chart the courses of people’s lives. For this reason, we need to take seriously the question of how well psychometric testing approaches are working in talent acquisition processes in organizations.

In this paper, we focus on how to evaluate tests that are used in talent acquisition. We use the term talent acquisition to describe recruitment and selection from both internal and external sources. We emphasize the need to recognize that many new big data-based measurement methods are psychological tests and should be held to the same standards as other psychological tests before being used to make decisions in talent acquisition.

Why is testing and assessment used in talent acquisition?

Typical talent acquisition processes make extensive use of testing and assessment, and it’s easy to identify the reasons behind the popularity. One important appeal is that testing helps practitioners become more efficient by providing a way to reduce the size of applicant pools. This is a critical component of Quality of Hire (QoH) metrics in talent acquisition. But there is tremendous utility to be derived from testing and assessment processes beyond its efficiency benefit. In particular, it allows practitioners to be fair and principle-based in applying a common standard of evaluation against job requirements. For a recent review of the strength of the relationships between test scores and job performance see the article by Schmidt, Oh & Schaffer[1].

Technology's influence on testing in talent acquisition

The biggest developments impacting assessment experiences in talent management have been technological in nature. There have been two technological revolutions that have impacted testing in talent acquisition. The first revolution in the late 1990s coincided with the advent of widespread access to the Internet. The second revolution is going on right now. It involves a dramatic diversification of the data employed in assessment processes, and an extension of the analytical techniques employed to include machine learning and artificial intelligence for testing and assessment.

In the following sections, we describe these revolutions, with a focus on how assessment professionals working in talent acquisition can approach this second revolution to avoid getting lost in the wilderness of assessment using big data and machine learning.

The first testing revolution: from paper to online testing and assessment

The 1990s saw the advent of internet-based testing and an explosion of testing processes in talent management. Testing professionals around at the time will remember pronouncements about the impropriety of internet-based testing, because of threats to test content security (would people share the questions? would they get others to take their tests?).

Concern was also widespread about the purported lack of validity of many assessments produced by start-ups. Two-step authentication, where an initial un-proctored (i.e. unsupervised) score is verified in a subsequent proctored (i.e. supervised) session, parallel forms of tests, adaptive testing and camera technologies are three ways that the risk of cheating has since been managed.

The important lesson is that when technology challenged traditional assessment mechanisms, the industry did not relax its standards. Indeed, if guidelines from testing bodies such as the International Test Commission are anything to go by, standards have become more rigorous since internet testing began.

A second revolution: new methods and big data

Knowing that technological revolutions can lead to more rigor is important, because we now find ourselves in the midst of another revolution, and once again it's technological. New methods of assessment are being proposed based on technology that represents a dramatic diversification of the testing and assessment approaches traditionally used. For example:

- Voice and facial recognition in video interviewing that purports to measure emotion and allows concept extraction from candidate responses that can be evaluated against rating criteria
- Social media scraping of the digital exhaust we create as we participate in an online world, and which proponents claim can be used to assess personality traits
- Serious games (games with a primary goal other than entertainment[2]) offering data on user behavior that may be indicators of work relevant traits

Analyzing the data that these assessment methods produce requires skills that are more common among computer scientists and machine learning experts than among traditionally trained industrial-organizational (I-O) psychologists. The situation has likely left many I-O psychologists working in talent assessment wondering if they should have studied computer science.

However, many of these new methods are psychological tests according to Society for Industrial and Organizational Psychology (SIOP) guidelines. Therefore, psychological concepts like reliability (the repeatability and precision of test scores) and validity (the extent to which tests measure what we believe they are measuring) are critical to consider for these emerging tests. However, these concepts are as exotic to the developers of many of these big data and machine learning assessments as machine learning methods like neural nets and natural language processing are to psychologists. Psychologists therefore need to work closely with developers of these new techniques to ensure that traditional standards of rigor are maintained.

Psychologists need to work closely with developers of these new techniques to ensure that traditional standards of rigor are maintained.

Benefits and challenges of new approaches to assessment

The benefits of these new assessment methods receive considerable attention. They certainly have appealing features: hiring organizations can be seen as more attractive, new methods might be cheaper, and the experience of candidates can be enhanced. In particular, organizations find benefits in the ability to rapidly assess large volumes of candidates in the time it used to take to screen just a few. Furthermore, the interviewer and interviewee no longer need to be in the same room, as responses to digital interview questions can be recorded remotely. Managers can then review interview answers that are standardized for all applicants from anywhere, at any time, and even at high speed.

Obtaining candidate success profiles from digital footprints has the advantage that organizations can passively profile candidates who have not yet applied for jobs – although the legality and privacy of such an approach needs to be managed on a principle basis against evolving best practice, regulatory and legal guidance. Because candidates don't know they're being assessed, the thinking is the scores are less likely to reflect impression management efforts (although common sense suggests many people curate their online presence as carefully as they do their offline personas). In addition, the ability to identify prospective talent to convert to applicants is seen as a competitive advantage by employers akin to the function applied by agencies in many sourcing models.

When assessments are presented as serious games, the candidate experience is argued to be more favorable. This positive experience is important, because job applicants are often also customers of the hiring organization[3]. Businesses want candidates to tell their friends about opportunities, reapply for other roles if they are unsuccessful the first time around, and keep using their products even if they do not get the job.

Without detracting from the work that is occurring in these new areas of assessment, it is also important to state that many of these new methodologies do not yet produce assessment scores with the same psychometric characteristics of reliability and construct validity that have been established for traditional assessment methods using rating scales. What is unclear currently is whether this is simply a short-term calibration issue. Perhaps, with time, these methods will reach the higher traditional standards. On the other hand, it might be that the traditionally accepted re-test reliability, score precision, and construct validity standards are simply unachievable for these new techniques.

Industry accepted standards are detailed in documents such as SIOP's principles for the validation and use of personnel selection procedures[4] and the American Psychological Associations Standards for Educational and Psychological Testing[5]. These standards should remain firmly in the minds of talent management practitioners when evaluating the suitability of testing and assessment procedures in talent management. These standards have been established to be consistent with widely accepted interpretations of the state of psychological science. They also cover topics like the expected effectiveness of different procedures, ways to eliminate psychometric bias from selection procedures, and legal defensibility.

Industry accepted standards … should remain firmly in the minds of talent management practitioners when evaluating the suitability of testing and assessment procedures in talent management.

Four guidelines for evaluating new methods

The overall message of this article, then, is to not forget the basics when evaluating new assessment techniques. However, the basics outlined in various standards documents can be extensive and sometimes even advanced. Recognizing that not every practitioner has the time to remain familiar with all the standards and scientific research, here we offer four guidelines that practitioners should use to frame their thinking about testing and assessment, whether new or traditional. They do not replace the standards documents, rather, they reflect the essence of the ideas in the standards documents in a way that practitioners will be able to recall easily.

Guideline 1: Focus on the validity of constructs, rather than methods

Confusion reigned regarding the efficacy of different predictors in industrial-organizational psychology until the late 1970s. A consistent picture of what works and what does not work in predicting job performance did not emerge until the field realized an important point: different terms were being used for the same concepts, and different concepts were being referred to with the same terms. Once the field addressed this problem, a coherent picture began to emerge. To avoid a replay of this situation from the ‘70s, today’s practitioners should differentiate between measurement methods (e.g., web scraping, video interviewing, serious gaming) and the concepts that the measurement methods assess (i.e., abilities, personality, values, interests, and behavioral and technical competencies). Practitioners should ask how well a concept like personality is assessed using, for example, web scraping, rather than focusing on the efficacy of web scraping as a predictor of future job performance. Without this conceptual distinction, practitioners risk drifting into measurement of surface features with no clear rationale for why a ‘like’ here and a ‘retweet’ there matters for future job performance. The generalizability of results (i.e., whether scores will have validity beyond the sample in which a relationship is originally discovered) are likely to be stronger if this point is kept in mind.

Guideline 2: Look for evidence of reliability and multi-trait, multi-method validity

In psychometrics, test developers show scores are reliable by demonstrating that they are repeatable and precise. We also show they are valid by demonstrating that they are related to other measures of the same or similar things. For instance, if a new selection game purports to measure a candidate’s level of sociability, sociability scores from the game might be expected to relate to sociability scores from a personality questionnaire.

To that end, ensure you ask your test providers what evidence they have that their tests are reliable and valid.

Reliability

To confirm reliability, request information about the scores for the same people on two measurement occasions.

Validity

The most effective way to confirm validity is to ask for a validity study where relevant concepts are checked by at least two methods, such as a standardized questionnaire, and via web-scraping.

Correlations between the two different methods of measuring the trait should be high. Correlations between different traits measured using the same method should be low. This approach is referred to as multi-trait, multi-method validity and it is the gold standard for evidence by talent management practitioners considering emerging assessment methodologies in talent acquisition.

In summary, guideline two is for practitioners to ask for both reliability and multi-trait, multi-method evidence when considering new testing methods.

Guideline 3: Require local validity evidence for new methods

Where summary evidence does not exist (e.g., meta-analyses which quantitatively summarize the efficacy of a new method of measuring a work-related attribute), organizations should require local validity evidence – from their own organization. This evidence should show how well the selection technique is predicting performance in your organization. Keep in mind that small sample sizes, unreliability of measures, and range restriction (i.e., a limited distribution of scores on a measure) can impair your ability to successfully evaluate the effectiveness of psychological test validity. However, corrections can be made for these problems and, in the absence of meta-analytic evidence, these corrections should be applied to local data to assess the effectiveness of testing procedures.

To help, as you evaluate your own local validation of new assessment methods, we recommend requiring a correlation of at least .20 with job performance for a selection method to be acceptable for operational use if the job performance outcome is measured on a scale (e.g., 5 points or more). If the outcome variable is categorical, (e.g., a value that is a name or label rather than a number), it is important not to focus on the overall accuracy, but rather consider whether there is any improvement over guessing the most common category. For instance, if I am predicting turnover, and 80% of workers never leave, it is easy to be 80% accurate by guessing that nobody ever leaves. But more interesting is how well we can predict those who do leave, so we want to know what improvement we can make beyond predicting that nobody will leave.

Guideline 4: Ensure similar information is considered for all candidates

Our fourth guideline is around ensuring consistency in the content provided by candidates or gained via techniques like scraping. AI capability can only be deployed on available data, so inconsistent availability of data may handicap some candidates. Missing data may not necessarily be a sign of a poor candidate. Given AI capability like matching job profiles to resumes, we need to ensure candidates have the ability to provide relevant and complete information on themselves in order to evaluate and make decisions on their applications.

Summary

Traditional standards of reliability and validity developed by psychologists were intended to ensure unbiased treatment and effectiveness of assessment processes in talent management. When evaluating new talent assessment techniques, talent management professionals should retain their traditional standards for reliability and validity.

While there is a lot to consider in this domain, the four core guidelines presented in this paper can help in your decisions about what to adopt in your own organization. New techniques that cannot pass these requirements may provide general insights into public opinions and attitudes, but should not be used for decision making about individual job candidates in talent acquisition settings.

Learn more about
IBM Kenexa Employee Assessments →

IBM Smarter Workforce Institute

The IBM Smarter Workforce Institute produces rigorous, global, innovative research spanning a wide range of workforce topics. The Institute's team of experienced researchers applies depth and breadth of content and analytical expertise to generate reports, whitepapers and insights that advance the collective understanding of work and organizations. This paper is part of IBM's ongoing commitment to provide highly credible, leading edge research findings that help organizations realize value through their people. To learn more about IBM Smarter Workforce Institute, visit http://ibm.biz/Institute

How IBM can help

IBM leverages the power of innovation, data, and expertise to improve business and society. By bringing together behavioral science, artificial intelligence, and expert consulting, IBM helps companies attract, hire, and develop the talent they need to grow their business. For more information, visit ibm.com/talent-management

About the authors

Nigel Guenole, Ph.D. is an Executive Consultant with the Smarter Workforce Institute and a Senior Lecturer in Management at Goldsmiths, University of London. He is known for his work in workforce analytics, statistical modeling and psychological measurement. Nigel's work has appeared in leading scientific journals including Industrial Organizational Psychology: Perspectives on Science and Practice and Frontiers in Quantitative Psychology & Measurement, as well as in the popular press. Nigel is the current external examiner for organizational behavior programs at London School of Economics (LSE) and University College London (UCL). He is a Chartered Occupational Psychologist and an Associate Fellow of the British Psychological Society (BPS). He is registered with the Health & Care Professions Council (HCPC) in the United Kingdom, is a member of the Academy of Management (AoM), and is an international affiliate of the Society for Industrial and Organizational Psychology in the United States (SIOP). At Goldsmiths Nigel teaches courses on leadership and statistical modelling. Nigel is also co-author of the book The Power of People: Learn How Successful Organizations Use Workforce Analytics To Improve Business Performance (Pearson, 2017).

Sheri Feinzig, Ph.D. is the Director, IBM Talent Management Consulting and Smarter Workforce Institute and has over 20 years' experience in human resources research, organizational change management and business transformation. Sheri has applied her analytical and methodological expertise to many research based projects on topics such as employee retention, employee experience and engagement, job design and organizational culture. She has also led several global, multi-year sales transformation initiatives designed to optimize seller territories and quota allocation. Additional areas of expertise include social network analysis, performance feedback and knowledge management. Sheri received her Ph.D. in Industrial-Organizational Psychology from the University at Albany, State University of New York. She has presented on numerous occasions at national and international conferences and has co-authored a number of manuscripts, publications and technical reports. She has served as an adjunct professor in the Psychology departments of Rensselaer Polytechnic Institute in Troy, New York and the Illinois Institute of Technology in Chicago, Illinois, where she taught doctoral, masters and undergraduate courses on performance appraisal, tests and measures. Sheri is also co-author of the book The Power of People: Learn How Successful Organizations Use Workforce Analytics To Improve Business Performance (Pearson, 2017).

References

1 Schmidt, F. L. and Oh, I-S., and Shaffer, J. A., The Validity and Utility of Selection Methods in Personnel Psychology: Practical and Theoretical Implications of 100 Years of Research Findings (October 17, 2016). Fox School of Business Research Paper. Available at SSRN: https://ssrn.com/abstract=2853669

2 Armstrong, M. B., Landers, R. N., & Collmus, A. B. (2016). Gamifying recruitment, selection, training, and performance management: Game-thinking in human resource management. In Emerging research and trends in gamification (pp. 140-165). IGI Global.

3 IBM Smarter Workforce Institute (2017) The Far-Reaching Impact of Candidate Experience. https://www-01.ibm.com/common/ssi/cgi-bin/ssialias?htmlfid=LOW14341USEN

4 SIOP (2003). Principles for the Validation and Use of Personnel Selection Procedures. Fourth Edition. http://www.siop.org/_principles/principlesdefault.aspx

5 APA (2014). Standards for Educational and Psychological Testing.

IBM Corporation
New Orchard Road,
Armonk, NY 10504

Produced in the United States of America
May 2018

This document is current as of the initial date of publication and may be changed by IBM at any time. Not all offerings are available in every country in which IBM operates.

THE INFORMATION IN THIS DOCUMENT IS PROVIDED “AS IS” WITHOUT ANY WARRANTY, EXPRESS OR IMPLIED, INCLUDING WITHOUT ANY WARRANTIES OF MERCHANT¬ABILITY, FITNESS FOR A PARTICULAR PURPOSE AND ANY WARRANTY OR CONDITION OF NON-INFRINGEMENT. IBM products are warranted according to the terms and conditions of the agreements under which they are provided.

♻ Please Recycle