Best Practices for the Ethical and Effective Use of Artificial Intelligence to Assess Talent

SHL.

Executive Summary

The use of Artificial Intelligence (AI) in Human Resources (HR) continues to increase. Estimates suggest 43% of organizations are now using AI to help complete HR tasks, up from 26% in 2024, led by Recruiting and Talent Acquisition functions (SHRM, 2025). In addition, 78% of organizations use AI in at least one business function (McKinsey, 2025). Talent assessment, in particular, is keeping pace with this adoption of AI, creating efficiencies and improvements through enhanced prediction of employee outcomes, reduced discrimination, and a more engaging candidate experience. Organizations will continue implementing AI-based assessment solutions at increasingly meaningful scale to drive efficiencies and enhance outcomes. Expertise is required to not only effectively incorporate AI technology into talent assessment processes, but also to navigate the legal landscape and candidate expectations if they are to realize their true potential. Therefore, an HR function that aims to successfully harness AI for talent assessment will benefit from expert guidance when navigating this complex and shifting landscape.

This document is intended to provide guidance regarding the application of existing talent assessment guidelines and regulations towards AI assessments, and to highlight some of the key steps and considerations when developing and/or using an AI assessment.

SHL's Best Practices for the Ethical and Effective use of AI to Assess Talent have been developed to align with current guidelines and legal regulations in both talent assessment and AI. These Best Practices are intended for HR practitioners, Industrial/Organizational (I/O) Psychologists, and other talent program owners who are interested in applying AI to assess candidates and/or employees.

This document begins with an introduction to AI and its current use in talent assessment. The remaining sections provide SHL's Core Principles and Best Practices as a guide for the design, development, and use of AI that is both ethical and effective in assessing talent.

Contents

Disclaimer: The contents of this document have been prepared in good faith as an introduction. It is not intended to provide legal advice, and you should not take, or refrain from taking, action based solely on its content. Neither SHL nor its affiliates accept any liability for any errors or omissions, nor for any loss incurred by placing reliance on the contents of this document. We do not owe a duty of care to any party in relation to the preparation of this document. Any person reading this document should use their own professional judgment regarding its content and should also seek independent advice specific to their circumstances.

Note: Throughout this document, italicized words or phrases are defined in the Glossary of Terms.

1. Introduction

Innovation in talent assessment has often been driven by technological developments in other fields. For example, the development of the computer, the internet, and the smart phone each led to a revolution in the way candidates and employees are assessed. Assessments have progressively moved further away from proctored pencil-and-paper testing, towards increasingly digital, more widely accessible, and higher-fidelity[1] methods. These moments of innovation have resulted in significant benefits to the organizations that effectively utilized this merging of new technology with the science of assessment.

A look at today's talent assessment landscape reveals the growing adoption and application of Artificial Intelligence (AI) technology (Microsoft and LinkedIn, 2024). It continues to spark innovation across the field and offers competitive benefits to organizations that successfully leverage this technology. AI enables the simulation of intelligent behavior in computers and has the potential to increase the validity and to enhance the candidate experience of assessments, while at the same time reducing unfair human bias. Such improvements to the assessment and selection process can have positive ripple effects across an organization, as the hiring of candidates with higher potential and better fit for the role, and the organization, will lead to increased employee performance and better business performance outcomes, as well as increased opportunities for previously underrepresented demographic groups.

However, when it comes to applying AI, as with many new technologies, there remain many unknowns and uncertainties, particularly regarding the use of AI in the assessment of talent. At the current time, it can be difficult to separate the AI hype from the facts, to comprehend the still-evolving legal regulations, and to make informed science-practice decisions with confidence, due to the lack of research and guidelines on the use of AI in assessment.

This has put HR teams in modern organizations into a difficult position - there may be pressure to quickly implement an AI tool without the necessary guidance and clarity required to do so. Therefore, a best practice document which highlights some of the critical considerations that should be made when developing or using an AI-based assessment provides tangible value to all HR professionals.

SHL has a track record of developing pragmatic best practices and recommendations for organizations, at key historical moments, when technology has significantly affected the science and practice of talent assessment[2]. In doing so, it is our hope that we have brought some clarity and useful guidance to organizations, and individuals, who have been involved in the application of these new technologies to HR. In this document, we present SHL's core principles and recommended best practices for the ethical and effective use of AI to assess talent.

SHL believes that AI has huge potential to improve assessment, people decisions, and outcomes for both organizations and individuals. However, the use of AI is not without effort or risk, and AI technology cannot simply be inserted into a process without careful forethought and the oversight of program owners and Subject Matter Experts (SME).

Our point-of-view paper, Harnessing AI in Talent Assessment, outlines how SHL develops, deploys, and monitors AI, and how responsible, transparent AI is shaping the future of talent assessment. Alongside the principles and practices presented in this paper, they should help organizations to develop an approach to minimize the risks, while increasing the benefits, of AI-based assessment.

[1] e.g., Multimedia-based Situational Judgment Tests (SJT) that closely resemble the tasks performed in the target job.

[2] Prior examples of SHL best practice documents and recommendations: 1) How to use competencies in testing (Bartram, 2005); 2) the validity of unproctored internet testing (Beaty, et al., 2011); 3) best practices for unproctored internet testing (Beaty, Dawson, Fallaw, & Kantrowitz, 2009); 4) techniques for cheating prevention in online cognitive ability testing (Burke, 2015); 5) device-equivalent mobile-first cognitive ability assessments (Grelle & Gutierrez, 2018).

2. What is AI?

Artificial Intelligence is a general term for any algorithm or computer program that attempts to simulate human-like intelligence or judgment (Poole & Mackworth, 1998). The definition of AI has evolved over time and tends to vary across fields. However, most definitions include the idea that AI is an effort to replicate tasks and processes, with computers, that are normally thought to require human intelligence. In other words, AI refers to attempts to make machines act intelligently.

What it means for machines to act intelligently has also varied with time. Early AI applications often used rule-based algorithms designed to follow clearly defined processes (e.g., playing a board game). More recently, many AI applications use *machine learning or deep learning*, which utilize quantitative models designed to learn patterns from observed data and then apply that information to new scenarios (i.e., new data). These quantitative models have evolved over time with early models being more similar to classical statistical models and newer models being more similar to artificial brains with billions of artificial neurons. These newer models are known as artificial neural networks. For example, online retailers use machine learning to make future purchase suggestions based on a user's past purchases. Machine learning algorithms, such as these, are what the term AI refers to in this document.

Many AI applications today are designed to utilize *Natural Language Processing* (NLP), which enables computer algorithms to parse and extract meaning from conversations and *natural language* (e.g., written or spoken text). Personal digital assistants, such as Amazon's Alexa or Apple's Siri, are examples of the sophisticated application of NLP combined with modern technology. A core technology within NLP is language modelling where a specially designed neural network is trained to understand written language through a process of predicting the next word in a string of words. The process of training increasingly sophisticated and powerful neural networks through language modelling gave rise to Generative Pretrained Transformers (GPT) and other Large Language Models (LLMs). AI tools based on these technologies have seen explosive adoption rates over the past few years given their wide-ranging applications and relative ease of use.

To learn and apply patterns in data, AI makes use of what are known as *features*. Features are quantifiable properties of a phenomenon being observed that are present in the data. In statistical models, features are sometimes referred to as "independent variables" or simply "predictors". AI algorithms find patterns in the relationships between features and an outcome variable (referred to as a *criterion* in I/O Psychology).

There are often multiple datasets and iterations involved in developing an AI application. The dataset from which the AI first identifies patterns amongst features and outcome variables is referred to as the *training dataset*, as this is the dataset which "trains" the AI. For example, in developing an AI-based video interview assessment, certain features from the recorded interview – such as words spoken or facial expressions – might be used to predict subsequent performance on the job for those who were hired. In this scenario the AI application would identify patterns between the video features and job performance in one dataset (the training data), and then attempt to apply those same patterns to another dataset, often called a test or *holdout* sample. If the same patterns do not apply to the test sample, then the AI may have identified the wrong patterns in the training data, and therefore another attempt at learning from the training data could be required. This process is known as cross validation, and it is a very important concept for the development of any assessment, but it is particularly important for developing and validating an AI assessment[3].

[3] See section IV. Rigorously Validate for more information on the cross validation of AI assessments

AI Assessment

The term *"AI assessment"* will be used frequently throughout the document, and refers to the following *– any non-human analysis of participants' responses that utilizes machine learning, NLP, or other related modeling approaches and techniques (e.g., deep learning, latent semantic analysis) to assign scores to attributes of people (e.g., KSAOs, competencies) or to individuals' expected work outcomes (e.g., probability of turnover).*

AI assessment does not refer to a specific item-type or assessment method. AI can be applied to scoring responses from any number of different assessment methods, many of which might generally be described as conversational assessments. These methods and assessment designs include but are not limited to:

- Asynchronous digital interviews (video, audio, textual)
- Live/synchronous digital interviews
- Constructed-response situational judgment tests (SJTs) and multimedia situational judgment tests (MMSJTs)
- Coding simulations
- Spoken and written language assessments
- Role plays, business cases, presentations, and other work samples or simulation exercises

3. The Promise of AI in Talent Assessment

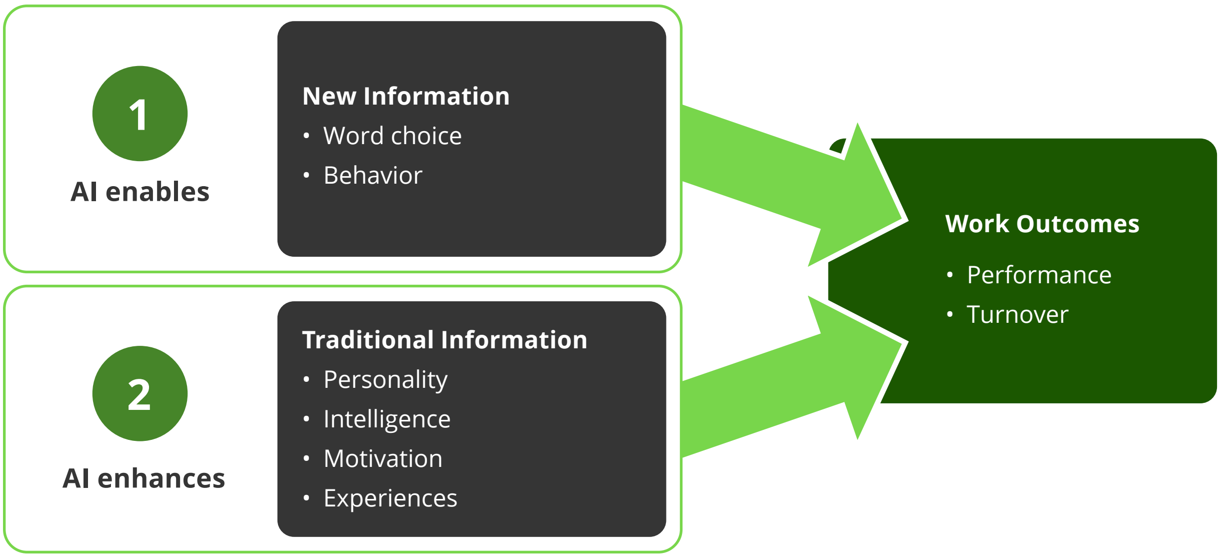
The field of talent assessment is constantly evolving, and many of the drivers of this evolution have come from outside of the field. For example, the widespread availability of computers and the creation of the internet led to the development of online assessments and applications of Computer Adaptive Testing (CAT). Likewise, the development of the smartphone has enabled mobile-based talent assessments.

AI is now driving another round of evolution in talent assessment. AI can draft attention-grabbing job descriptions (Sheng, 2019), identify optimal and diverse sourcing channels for candidates, and streamline candidate-recruiter interactions (Wisenberg Brin, 2019), and enhance or transform the assessment of candidates and employees, which is the focus of this paper. SHL has also implemented AI-driven capabilities to automate and augment various steps in the overall talent assessment and development process. One example of this is the ability to produce a skill or competency profile from a client's written job description instead of requiring more traditional or time-intensive steps.

The benefits of AI assessments, discussed below, will undoubtedly have a strong impact on the way organizations assess their talent. However, the benefits of AI assessment can extend beyond HR and have a much wider impact throughout the organization. For example, most organizations see employees as their greatest asset. Following this logic, then, the processes and tools involved in selecting and developing those employees must also be highly valuable. Any improvements in these processes and tools can have wide-reaching ripple effects that improve performance and efficiencies throughout the entire organization, and, ultimately, a boost in business performance metrics (e.g., sales revenue, gross margin, net profit margin, net promoter score). It is because of these expected benefits to bottom line performance that organizations are attempting to rapidly implement AI applications.

Yet, these benefits to business performance metrics start with localized benefits in talent assessment. The benefits of AI to assess talent can be grouped into two broad categories, which are described below, and presented in Figure 1.

Figure 1. Two Ways AI Improves Talent Assessment



1. Benefits From the Use of AI in Combination With Digital Technology, to Enable the Objective Measurement and Scoring of New Information

In addition to improving the accuracy, fairness, and candidate experience of traditional assessments, AI enables the measurement and scoring (i.e., modeling) of new sources of information regarding a candidate (e.g., word choice). However, it is not AI alone that enables this, but the combination of AI with advances in digital technology.

For example, using digital technology, data from a video interview can be stored and used to develop an AI assessment. In this scenario, the AI can learn patterns between features in the video interview data (e.g., words spoken) and an outcome variable (e.g., subsequent ratings of job performance). While features such as these may have been previously assessed (unsystematically, subjectively, and/or subconsciously) during in-person interviews, AI now enables the objective and systematic measurement of these features and their explicit inclusion in predictive models.

The ability to include these new sources of information in assessments offers a variety of potential benefits, from increasing validity to greatly enhancing the assessment experience by enabling candidates to respond in a natural format – by speaking or demonstrating their response as part of a conversational assessment, instead of selecting from a predetermined list of options.

2. Benefits From the Use of AI to Improve the Scoring of Traditional Information

AI can be used to improve the scoring of data from traditional assessments, such as personality questionnaires or intelligence tests. The goals of revised scoring could include better prediction of job-related outcomes (e.g., performance), reducing bias, reducing assessment length or administration time, or improving candidate experience (e.g., by enabling higher-fidelity simulations with natural language responses). An example of this approach is research showing how machine learning techniques can enhance the prediction of job-related outcomes, and while modeling information at the item-response level (which may reduce the number of questions required; Putka, Beatty, & Reeder, 2018). Notice that this second approach does not necessarily involve any changes to the way traditional assessments collect the data (i.e., the method of measurement). It is only the scoring of these assessments that is changed, and improved, with the use of AI. A summary of the benefits of AI assessments, that result from both of these categories, is presented in Table 1.

Table 1. Benefits of AI Assessments

1. Enhanced validity	AI provides an optimal method for scoring candidate responses and features in a way that can increase the validity with job-related outcomes.
2. Reduced bias	AI can learn to mitigate bias from an assessment.
3. Scoring of new information	AI can extract and model features that were previously very difficult to objectively analyze using more traditional modeling techniques (e.g., OLS regression). For example, the objective scoring of behavior, spoken language, and acoustic information. The use of LLM's can provide general scoring for a broad range of language-based assessments.
4. Natural response format	AI can enable a very natural way for candidates to respond to an assessment conversationally. For example, speaking a response, or demonstrating a behavior, instead of selecting from a narrow list of predetermined options (as required with traditional SJTs or personality assessments).
5. Higher fidelity	AI can lead to assessments that closely approximate the work conducted in the job. For example, an AI conversational assessment could be developed for a call center role which has the candidate verbally respond to simulated phone calls.
6. Better candidate experience	AI can increase candidate experience and positive perceptions of the organization by creating assessments that are engaging, highly job relevant, and that act as a Realistic Job Preview (RJP).

These benefits of AI assessments could lead to wider changes in the field of talent assessment, perhaps resulting in new standards for validity (e.g., the redefinition of small, medium, and large validity sizes; the use of different metrics to demonstrate validity), further development in psychometrics (e.g., computational psychometrics), and new theoretical contributions from the ability to study features that were previously very difficult to measure objectively and model at scale (e.g., acoustic features of spoken responses, DeGroot & Gooty, 2009).

In short, AI promises to enhance the scientific measurement and prediction of candidate and employee traits, attitudes, competencies, and behavior, resulting in wide-ranging benefits to both individuals and organizations.

4. Managing Risk

All forms of candidate or employee assessments carry an associated risk, and this is no different for AI assessments. Therefore, the same guidelines and regulations that inform the use of traditional assessments also apply to AI assessments. However, there are some additional considerations regarding risk that should be made when developing or using AI assessments. These additional risks have been grouped into three categories, presented below:

The Additional Risks Associated with AI Assessments

Legal Risks

Of the different categories of risk associated with AI assessments, organizations are perhaps most concerned with legal risks, as an organization that finds itself operating on the wrong side of legal regulations and guidance issued by regulators may face steep financial penalties. Described below are two of the various relevant components of legal risk regarding the use of AI assessments: data collection and usage, and bias.

The Collection and Use of Personal Data

The number of countries implementing legal regulations that outline the permitted use of AI in assessments has been growing in recent years. At this time, the majority of these regulations have emerged in the United States and Europe and are interconnected with legal protection around the collection and use of personal data. Examples of these regulations are presented in Table 2.

Table 2. Examples of Regulations Regarding the Collection and Use of Personal Data.

Region	Legislation	Description
European Union	General Data Protection Regulation (GDPR)	The GDPR requires processing of personal data to be fair, lawful and transparent. This requires companies to disclose the use of AI to applicants and to provide sufficient information on how their data will be used (and, if consent is required, in order for the applicant to make an informed decision to provide consent; Liem et al., 2018). Additionally, the GDPR includes the right not to be subjected to solely automated decision making in Article 22 GDPR, meaning that applicants have the right (in certain circumstances) to obtain human intervention, express their point of view about the decision and to have a right of appeal against the decision.
	EU AI Act	The EU AI act is a comprehensive legislation governing the use and compliance requirements of all AI systems made available to the EU market. The act takes a risk-based approach to AI systems placing them into a range of risk tiers depending on their technology and application. Most AI systems used in the recruitment/employee development domain are placed in the "high risk" tier. The act requires all high-risk AI systems to be extensively documented and registered with the EU to ensure all AI systems that are available on the EU market meet their rigorous requirements.
United States	Illinois AI Video Interview Act	The Illinois AI Video Interview Act requires employers to obtain the consent of applicants in order to use AI in the hiring process (Bologna, 2019). Additionally, it requires employers to explain the process and destroy data upon request.
	New York City Bill	Assessments used in recruitment that prioritize AI scoring above other methods (AEDT) must undergo yearly third-party bias audits and summaries of the audit results must be made available to all candidates that complete a recruitment process that includes the assessment. Employers must also advise candidates that they use AEDT at least 10 business days before they intend to use it and give candidates the opportunity to seek alternative assessment pathways.
	Colorado Consumer Protections for AI (S.B. 205)	The comprehensive Colorado bill requires that an employer must comply with high-risk AI system standards, bias audits for AI systems in employment and insurance.
	Texas H.B. 149	An employer is prohibited from developing or using an AI system to intentionally discriminate against a protected class in violation of federal law.
	Maryland H.B 1202	Employers must be transparent and follow ethical guidelines and obtain employee consent to use AI-based facial recognition technology in hiring.
	Illinois Biometric Information Privacy Act (BIPA)	BIPA requires employers to inform applicants of any biometric data collected and stored (e.g. vocal or facial characteristics), the reason for the collection, and how long it will be stored.
	Health Insurance Portability and Accountability Act (HIPAA)	While HIPAA is primarily intended to regulate healthcare information, it does have implications for AI assessment, in that companies using AI assessment must disclose or take steps to prevent the accidental collection of sensitive health information from non-clinical sources (e.g., social media) (Weintraub, 2017).
	California Consumer Privacy Act	The California Consumer Privacy Act requires businesses to provide individuals with notice of the personal data being collected about them, as well as the ability to opt out of data collection, request access to their personal data, or request its deletion. While the law currently excludes data related to employment assessment, it does require employers to provide notice to applicants regarding the categories of personal information collected and the intended use of the data.
	Fair Credit Reporting Act (FCRA)	The FCRA regulates the collection of consumer credit information and access to credit reports, and is relevant to talent assessment as it states that no organization should keep a secret database that is used to make decisions about a person's life, that individuals should have the right to see and challenge the information held in such databases, and that information in such a database should expire after a reasonable amount of time.
	Washington State Senate Bill 6280	This bill regulates the use of facial recognition technology by government agencies, and requires that state and local agencies report on their use of the technology, submit a notice to the state specifying the purpose for which the technology is to be used and provide a data accountability plan.

It is important to note that region and country-specific regulations regarding the collection and use of personal data will continue to develop, and talent assessment professionals should stay informed of relevant updates. Further regulations and directives in many countries such as South Africa, Japan, China, and Australia are also likely to consider the implications of AI. The European Union, the United Kingdom, and several states in the US are currently evaluating AI protocols, new guidance, and/or new laws regarding the use of AI as it relates to legal principles of fairness and personal protections of data subject’s rights.

Risk of Bias

Although AI assessments can help mitigate risk when developed properly (Section 3), AI assessments implemented without due diligence and best practices could actually increase the risk of bias in a selection process. This is more likely to occur if there is a lack of SME oversight during the development of the assessment. For example, in developing an AI-enabled video interview assessment, if an inappropriate criterion with possible bias is chosen, this could result in a concretization of the bias as a permanent element of the assessment. There are various technical and non-technical approaches which can be taken to mitigate discrimination risks in AI assessments, such as monitoring for algorithmic fairness using appropriate measures. It is important to take account of these risks in the development and deployment lifecycle of any AI assessment system.

Brand & Public Relations Risks

The inappropriate use of AI in talent assessment could also have a negative impact on an organization’s brand and public image. Using AI in an unethical way – for example, collecting and storing information on employees, without their consent, in an effort to influence their opinions or behavior, or using AI to make blind decisions regarding talent management without incorporating the “human element” (e.g., SME oversight, appropriate communication from HR) – could result in a reduction in employee commitment to the organization, and, if made public, a reduction in the public opinion of the organization. For example, if an AI assessment is trained on an incumbent workforce that reflects historical inequalities (e.g., 90% male employees), this could result in a biased assessment which exacerbates adverse impact. Recent academic research has also found that the use of black box AI assessments result in negative candidate views towards the organization (Gonzalez, Capman, Oswald, Theys, & Tomczak, 2019).

Validity Risks

While the previous section mentioned that AI assessments can have higher validity than traditional assessments, if AI assessments are not appropriately designed and used, the validity could actually be lower. This can occur when an AI assessment has been developed on the training data, and not thoroughly tested on new samples (i.e., not cross validated). Cross validation are procedures that are important for any assessment, but are especially important for AI assessments because: 1) there is often a very large number of features that are considered for use in an AI assessment (e.g., the words that someone uses during a video interview); and 2) some of the features that can be included in an AI assessment currently lack strong theoretical backing (e.g., facial expressions from a video interview). Developing models with many features, without strong theory to guide feature development and selection, increases the likelihood of "capitalization on chance" and developing models that do not generalize beyond the training sample.

Another factor that can cause an AI assessment to have low validity is when the assessment is not validated with high quality criterion data. For example, low quality outcome data, such as events that occur prior to an employee's onboarding (e.g., hiring decision) or data with little variation (e.g., typical annual performance reviews), can also result in an AI assessment that might not be very predictive of the behaviors that an organization is ultimately interested in predicting. Although criterion data quality is also a risk with traditional assessment development and validation, the reliance of AI assessments primarily or exclusively on empirical feature selection can exacerbate the problems resulting from poor quality criteria. Therefore, the use of such an assessment would be unlikely to predict performance in the real world and would not deliver the benefits the organization is seeking.

Mitigating the Risks Associated with AI Assessments

The typical risks associated with assessing candidates and employees, as well as the unique risks for AI assessments mentioned above, can be mitigated when developing or using an AI-based assessment.
We present the above information as examples of key risk categories, and the core principles and best practices (in subsequent sections) as key considerations for mitigating these risks.

5. Core Principles for the Ethical and Effective Use of AI to Assess Talent

To help organizations reap the benefits of AI assessment, while helping to mitigate some of the potential risks, SHL presents three Core Principles in this section to guide the design, development, and deployment of AI assessments in organizations. These Core Principles apply to any assessment method that utilizes AI to score any form of response data including text, audio, and/or video.

Table 3 provides an overview of the three core principles followed by a detailed description of each of them.

Table 3. SHL's Core Principles for the Ethical and Effective Use of AI to Assess Talent

I. AI Assessment is Still Assessment
AI assessment needs documented evidence of reliability and validity just like all other assessments, and employers' AI assessment programs need to be demonstrably job-related
II. AI Assessment Should be Explainable
Users and participants should be able to understand what is being assessed and how the AI assessment is job-related
III. Big Claims Require Big Evidence
AI assessment adds new tools and possibilities to psychology's 140-year study of talent and human performance, but new approaches with new claims require more diligence and more empirical support

I. AI Assessment is Still Assessment

Regardless of whether a scoring algorithm utilizes machine learning, NLP, LLM's, rational or empirical scoring keys, human judgment informed by scoring rubrics, or any other method for assigning numbers to people's responses – if response behaviors of any kind are scored in a standardized manner and used in employment decisions, then that process is an assessment. The SIOP Principles (SIOP, 2018) concur: *"Scores produced by algorithms based on structured inputs (e.g., closed-ended assessment items) or unstructured inputs (e.g., resumes, open-ended text responses, or oral responses to stimuli) that are used to make selection decisions should also be recognized as predictors" (p.13).*

Therefore, although AI is a fairly new technology being applied to talent assessment, we can still examine AI assessment within the well-established scientific, pragmatic, and legal frameworks for evaluating assessment tools and employers' assessment programs. All of the professional standards for development, validation, and use of assessments still apply to AI assessment. Additionally, all of the legal requirements that apply to employers' use of assessments in making employment decisions (e.g., Uniform Guidelines on Employee Selection Procedures (EEOC, 1978) and other local employment laws) also apply to AI assessment.

Simply put, AI assessment needs documented evidence of reliability and validity (like all other assessments), and employers' AI assessment programs need to be demonstrably job-related.

II. AI Assessment Should be Explainable

Explainable AI assessment means that both the “why” and the “how” of the assessment process can be sufficiently described to users:

- Why should this assessment be used? AI assessment should target job-relevant criteria to justify its use in specific employment decisions; and
- How does this assessment work? Response characteristics or features that drive the AI scoring algorithm should be identifiable and explicitly linked to work performance or outcomes.

Like all assessment, **AI assessment should be job related**, meaning it should be used: to predict important or critical aspects of job performance; to predict important work outcomes; or to measure essential knowledge, skills, or abilities that are prerequisites to successful performance. The SIOP Principles (2018) concur: *“In cases where scores from such algorithms are used as part of the selection process, the conceptual and methodological basis for that use should be sufficiently documented to establish a clear rationale for linking the resulting scores to the criterion constructs of interest”* (p.13). There should be clear evidence and explanation that the scores produced by the AI assessment are directly related to job performance.

Scoring algorithms usually are not described in complete detail in published research in order to protect trade secrets and other confidential information. Nonetheless, SHL believes that there should be a **reasonable explanation of AI assessment scoring** to help users and participants understand the job relatedness of the end-to-end assessment process.

Organizations should be open and candid about their use of AI-enabled decisions, when they choose to use them and why they choose to do this, including proactively making people aware of the specific AI-enabled decision being made, in advance of the decision being made. A reasonable explanation should describe broadly how the AI assessment scoring works, and should specify what features of people’s data, performance, and/or responses are being scored. The explanation should be truthful and meaningful, written or presented appropriately and delivered at the right time. Regarding scoring, the SIOP Principles prescribe that: *“Methods and algorithms used to score content should be fully described… When performance tasks, work samples, or other methods requiring some element of judgment are used, a description of the type of rater training conducted and scoring criteria should be provided”* (p.34). Thus, some level of detail about what factors are being scored, and the features indicative of good and bad responses, should be documented for AI assessments (like all other assessments).

It is worth noting that technical documentation for assessments can vary depending on different intended audiences and their corresponding assessment usage, documentation, or evaluation requirements (e.g., User Guides vs. Technical Manuals vs. project-specific Technical Reports vs. Score Reports for participants and decisions makers).

III. Big Claims Require Big Evidence

Science advances over time through the refinement and elaboration – or disconfirmation and replacement – of theories and models on the basis of new evidence and discoveries. The field of psychology has studied the measurement of talent and human performance for 140 years, developing and refining theory-based practices with proven real-world effectiveness. All research on AI assessment of talent and prediction of human performance should be evaluated and understood in the broader context of the scientific study and professional practice of psychology and psychometrics.

In any area of scientific exploration and discovery, **bold new claims** that seem remarkable in the context of the accumulated body of scientific knowledge **require more evidence** to be credible. And likewise, less remarkable findings that are clearly linked to prior research and consistent with current theories should be easier to accept with less new evidence. Commonly referred to as the “Sagan standard” after the phrase was popularized by astrophysicist Carl Sagan on the TV show Cosmos, the idea that “extraordinary claims require extraordinary evidence” has been utilized as an evidentiary standard for scientific progress for hundreds of years. This standard is routinely applied by journal editors and peer reviewers to gauge the sufficiency of research evidence to support the researchers’ inferences in the context of the current scientific knowledge on the research topic.

“More evidence” could mean larger sample sizes, which arguably does improve claims of generalizability for specific research findings. More importantly, though, “more evidence” also means establishing evidence and theory to support every inferential link in the hypothesized causal chain connecting a person’s assessment responses to their future work outcomes. For example, a bold claim about the validity of a new predictor (e.g., that measures of facial action units predict job performance) that has little or no prior research support, and that doesn’t have strong links to accepted theories of human performance at work, should be held to higher evidentiary standards than validity claims that simply replicate well-established peer-reviewed findings (e.g., that measures of cognitive ability predict job performance).

For AI assessments to become more credible, more peer-reviewed research will be required. In the meantime, as with all other assessments, evidence of the validity of a specific AI assessment for particular uses should be documented in a corresponding technical manual consistent with professional standards that would enable qualified reviewers to evaluate the scientific validity of the AI assessment process.

6. Best Practices for the Use of AI to Assess Talent

This section presents SHL's six Best Practices for the use of AI to assess talent. These Best Practices are informed by SHL's three Core Principles listed in section five of this paper, and are aligned with guidelines and standards within the fields of talent assessment (e.g., SIOP Principles, 2018; Uniform Guidelines, 1978) and AI (e.g., European Commission's Guidelines for Trustworthy AI, GDPR). The Best Practices are presented in the order in which consideration is typically given during the development and deployment of an AI assessment, and are focused on key considerations and recommendations from the standpoint of talent assessment professionals. Organizations developing or using AI for assessments or other activities should consider additional relevant frameworks, such as the ICO's AI auditing framework.

An overview of the best practices is shown in Table 4. Following this is a detailed description of each best practice.

Table 4. SHL's Six Best Practices for the Use of AI to Assess Talent

I. Identify Data Requirements
Consider data quality, representativeness, and security.
II. Prioritize Transparency
Develop transparent AI - no "black box" algorithms.
III. Design for Fairness
Build fairness into the assessment from the beginning.
IV. Rigorously Validate
Hold AI assessments to a high standard regarding validity evidence.
V. Incorporate Human Oversight
No AI assessment should make decisions without human oversight.
VI. Disclose Intent
Provide a notification, explanation, and request consent (where and when required) from candidates who will be assessed by AI.

6.I Identify Data Requirements

The ethical and effective use of AI starts with considerations about the data. Using poor quality or inappropriate data, data that isn't sufficiently representative of the population, or failing to keep data private and protected could result in potentially severe consequences, in terms of both financial penalties and brand reputation. Therefore, a thorough consideration of the data requirements prior to designing and developing an AI assessment is crucially important. SHL regards the following as essential considerations for determining the data requirements prior to developing an AI assessment.

Data Quality

The quality of AI assessments is dependent on the quality of data. The term "garbage in, garbage out" is frequently used among computer and data scientists when referring to the fact that when poor quality data is put into a predictive algorithm, then the output is unlikely to be accurate or informative. This general concept – a recognition of the importance of data quality – has been a core part of the history of talent assessment and I/O Psychology, and has influenced the development of psychometrics and guidelines for assessment development. When developing an AI assessment, a thorough consideration of the quality requirements of the data is vital. Consider adopting standards and protocols to ensure the quality and integrity of data and consider the risks of data sets being compromised or hacked.

In talent assessment, quality data is data that results from a measure that is valid and reliable – a measure that produces data with as little "noise" as possible (i.e., with little measurement error). For example, the AI scoring of asynchronous video interviews typically requires that the words a candidate speaks are first transcribed into written text. To maintain quality data when using AI-scored video interviews, it is imperative that this transcription is as accurate and reliable as possible. In the development, and ongoing maintenance, of this type of AI assessment, a reasonable threshold for the acceptable accuracy and reliability of transcription should be determined and monitored. Additionally, this transcription should work well across and within all relevant groups (these will vary by country, and may include race, ethnicity, gender, national origin, disability status, and/or age).

A second example of the importance of data quality is in regard to the outcome variable, or criterion. If an AI algorithm is trained on a criterion that is not well measured or not related to actual performance on the job, then the scores produced by the algorithm may not predict actual job performance (see section 6.IV, 'Rigorously Validate' for further information regarding quality criterion). In addition to identifying the quality requirements of the predictor and criterion data, a plan for attempts to fake the assessment should be conducted. For example, if, during an asynchronous, AI- scored, video interview, a candidate repeats certain key terms in a nonsensical way, their resulting data will be of low quality. Efforts should also be made to flag plagiarized or non-participant responses.

Finally, it should be noted that the use of increasingly interactive technologies and requirement for openended responses, including recorded/streaming voice and video, also increases opportunities for technical issues to interfere with accurate assessment. Accordingly, efforts should be made to identify unusable responses (e.g., no/low volume, garbled speech, frozen video) and/or to advise participants when conditions may be unsuitable for further assessment (e.g., low bandwidth, too much background noise).

Data Representativeness

Consideration should be given to ensure that the data used for developing AI assessments are representative of the intended pool of applicants and all relevant groups (e.g., job level, age, gender, disability status). For example, an assessment that is developed using a sample of recent graduates should not be used to make decisions on mid-level managers. Having a representative sample also reduces the risk of bias against a particular group. To achieve representativeness, a sufficient amount of quality data must be collected from each relevant group, which may require strategic oversampling of smaller groups to ensure that sufficient data points are included in the training process. Building an AI assessment on data that are not sufficiently representative of all groups may result in the assessment being invalid for its intended purpose or biased against a protected group, thus exposing the organization to legal risks.

Data Privacy and Protection

Data from candidate and employee assessments often contains private and sensitive information. This is especially true for assessments such as asynchronous AI-scored video interviews, which may capture, store, and assess facial, acoustic, verbal, and behavioral information. Depending on how and where the data from a video interview is stored, it may not be possible to guarantee anonymity. Therefore, the data collected by these assessments must be processed and stored in a way that provides the greatest possible privacy and protection. The SIOP Principles state that data confidentiality is an ethical responsibility of the testing professional, and suggests that the testing professional *"provides the maximum confidentiality feasible in the collection and storage of data, recognizing that identifying information of some type is often required to link data stored in different databases, collected at different times, or collected by different methods"* (SIOP Principles, 2018, pg. 30).

In addition to following professional guidelines, assessment users and developers will need to follow the legal regulations that apply within their region of operation. In recent years, many countries have introduced legal regulations regarding the protection of personal data. These regulations typically require that data on individuals are:

- Collected only where it is required to fulfil a specific purpose
- Processed lawfully, fairly and in a transparent manner
- Approved for use and storage by the individual (e.g., consent in certain circumstances)
- Securely stored
- Accurate and, where necessary, kept up to date
- Appropriate for and minimized to the intended use case
- Only used for the intended use case
- Deleted when it is no longer necessary for the purposes for which it was collected
- Accessible to the individual (e.g., if they exercise their rights)

Particular care should be taken regarding sensitive information. This could include any record of the person's ethnic origin (e.g., if collected for monitoring purposes such as adverse impact) as well as information relating to the person's health, disability, trade union membership, or religious views, among other information. The type of information considered sensitive varies by country, so it is advisable to refer to relevant employment and privacy law and regulations for your region.

Data Protection Checklist

This will vary depending on the local data protection laws in your jurisdiction. However, some suggested steps may include:

- Consider if consent is required to use video applications or AI assessment systems
- Depending on region-specific requirements, check whether you need to be registered with the local regulator or pay a fee (for example, the data protection fee required to be paid to the UK Information Commissioner's Office https://ico.org.uk/for-organisations/data-protection-fee/)
- Develop a data privacy governance framework for how you acquire, store, access, check, and delete personal data (including specific policies on the development and deployment of AI, where relevant)
- Document key decisions on the use of AI assessment systems (for example, conducting a Data Protection Impact Assessment (DPIA) prior to the deployment of a system)
- Consider both electronic and paper systems, as both typically fall under relevant data protection legislation
- Ask candidates to sign an agreement regarding your holding and processing of their data
- Check that assessors and line managers are not retaining copies of candidates' personal data

Section Summary

When developing an AI assessment, it is imperative to begin with considerations regarding the data. Making the right decisions in this step will reduce the risks associated with developing or using an AI assessment, and will increase the ultimate value of the assessment. Developing a plan that at least addresses each of the three considerations presented in this section will help to reduce these three categories of risk associated with AI assessments (refer to section 4 for an overview of these risks):

- Ensuring that the AI assessment is built on quality data will lead to an assessment with higher validity.
- Having data that is representative of all relevant groups will reduce the risk of bias (and legal issues).
- Maintaining data privacy and protection will reduce the risk of legal issues.

Identify Data Requirements: Best Practice

Maintain a high standard regarding all aspects and considerations of the data that go into developing an AI assessment. Ensure compliance with existing legal standards (e.g., GDPR, AI Video Interview Act) and guidelines (e.g., SIOP Principles, European Commission's Guidelines for Trustworthy AI) during the design, development, and operation of an AI assessment.

6.II Prioritize Transparency

Transparency is critical when developing or using AI for talent assessment. Being able to explain how an AI assessment works, and why it reached specific conclusions (i.e., scores), will help to safeguard an organization against potential ethical or legal issues. In addition, knowing what information the AI assessment is using to score candidates can help an HR team plan for how potential changes to the economy, candidate pool, or job requirements may affect the scores on the assessment, as well as impact the continued suitability of the AI assessment for a given role. This can be somewhat complicated when using third party AI and LLMs as part of talent assessment, whether content generation or scoring, as users of third-party tools will be dependent on those third parties for relevant documentation around transparency, development, etc.

Being transparent about how an AI assessment works is fundamental to the principles and guidelines in talent assessment. For example, the SIOP Principles (2018) state that *"variables chosen as predictors should have a theoretical, logical, or empirical foundation. The rationale for a choice of predictor(s) should be specified. A predictor is more likely to provide evidence of validity if there is good reason or theory to suppose that a relationship exists between it and the behavior it is designed to predict"* (SIOP Principles, pg. 12).

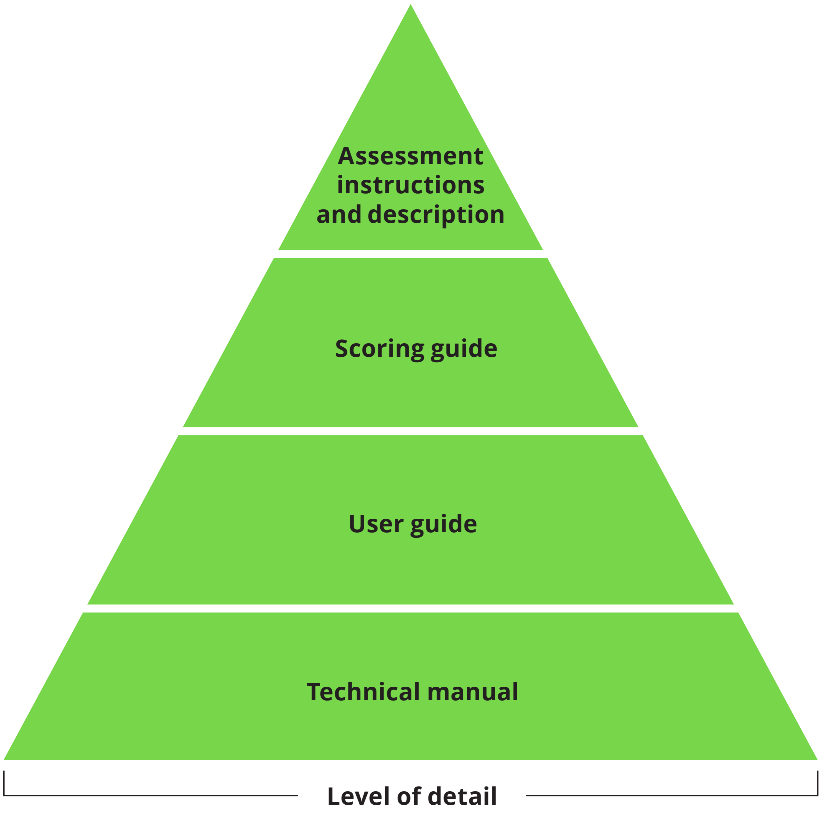
Therefore, the way an AI assessment makes decisions should be known to the assessment developer and user, and should be documented (e.g., in a technical manual). While this is best practice for any assessment, it is a particularly important consideration when designing and developing AI assessments, as they have the possibility of being completely black box – meaning that the way the assessment makes decisions is not fully understood by the assessment developer or the user. Because of this possibility, and the increasing use of AI within organizations, there is growing pressure for developers of AI to make their algorithms more transparent. The need for explainable AI has been recognized in the EU's Ethics Guidelines for Trustworthy AI (European Commission, 2019) and The Public Voice's Universal Guidelines for AI (2018). In addition, the EU's GDPR and associated regulatory guidelines amount to a right for individuals to be given an explanation for and to contest a solely automated decision by an algorithm. Under EU GDPR there is a more general right to be informed about how an individual's right is processed (e.g., by way of a privacy notice) and there may have to be a specific tailored notice used for an AI assessment. Therefore, not only is transparency a best practice according to the traditional guidelines for talent assessment, it has now become a legal requirement in many jurisdictions and is likely to form an important basis of analysis for any claim by a plaintiff.

One way to increase the transparency of an AI assessment is to only allow features that have a conceptual linkage to the target job to be included in the model training process. Identifying features that have a conceptual linkage to the job can be achieved through a thorough job analysis. This up-front approach to feature selection reduces the likelihood of including features which will not generalize to new data (i.e., will cross validate poorly), and can increase the face validity of the assessment.

Another way of increasing transparency is through the use of Explainable AI (XAI) models. XAI refers to a set of tools, methods, and algorithms in AI model development that allow each decision made during the machine learning process to be traced and explained in an understandable way. XAI helps to generate interpretable and intuitive explanations that make AI models more understandable and builds trust in the results generated by these models. For example, techniques can be used to determine what features (e.g., words and phrases identified through NLP) are driving the predictions generated by the model. If the most important features are logical and conceptually appropriate, the model is interpretable and very likely generalizable. If there are important features that are not easily explainable, that suggests a problem and makes it difficult to build trust in the model. XAI offers the benefits of building trust and confidence in the AI model, continuous evaluation, and improvement of models, and mitigating legal and ethical risk.

Given that transparency is a core element of the best practices in talent assessment (e.g., SIOP Principles) and AI (e.g., European Commission's Guidelines for Trustworthy AI, The Public Voice's Guidelines for AI) and is increasingly recognized in legal regulations (e.g., GDPR, AI Video Interview Act), SHL believes that AI assessments should be designed and developed with sufficient transparency that assessment users can have a reasonable level of understanding of how the assessment works, and why it assigned a certain score to an individual. A "reasonable level of understanding" may vary by the purpose of the assessment (e.g., used for selection versus training), the complexity of the underlying algorithm (e.g., a random forest model versus a neural network), and the requirements of the end user. For example, the requirements for a candidate taking an AI assessment are very different from the requirements of the recruiter receiving the candidate's score. The candidate may need to know basic information regarding the purpose of the assessment, how long it will take, high-level information about any constructs that are measured, and any information that may affect their score (e.g., instructions to speak for at least 30 seconds so that an accurate score can be calculated). However, the recruiter may need to know more specific information regarding the competencies assessed, and what elements of the candidate's response were most strongly related to those competencies. Therefore, when developing documentation to make an AI assessment more transparent, consideration should be paid to the various end users and the level of detail that each will require. Various documentation can then be produced to suit the needs of each end user. An example of such documentation and the level of detail required is presented in Figure 2.

Figure 2. Documenting the Transparency of an AI Assessment



Section Summary

Developing and/or using AI that is relatively transparent and explainable is quickly moving from good practice to a legal requirement. Having insight into how an assessment works has long been a best practice in the field of talent assessment, and this practice still applies to AI assessments.

Prioritize Transparency: Best Practice
Design, develop, and use AI assessments that are appropriately transparent. Transparency into an AI assessment may include a description of the features scored by the assessment, any constructs and/or characteristics that the features are related to, and some information regarding how the features are combined to produce a score. The level of information provided should vary by the end user (e.g., hiring manager, candidate).

6.III Design for Fairness

AI assessments should be ethical and fair. In talent assessment, "fairness" is a broad term that encompasses the following (SIOP Principles, 2018):

1. Equal treatment of all candidates in the selection process
2. Equal access to the constructs being measured by an assessment (i.e., "accessibility")
3. Hiring and selection processes that are nondiscriminatory (i.e., without bias)

SHL believes that when AI assessments are designed and developed to meet this standard of fairness, they can provide benefits to individuals (e.g., candidates and employees), organizations, and society at large. Therefore, developers of AI assessments and organizations' assessment program owners have a responsibility in supporting the achievement of these fairness goals.

Each of the three aspects of fairness mentioned above are describe further below.

Equal Treatment

All candidates should receive the same treatment during the assessment process. This includes the information that is shared regarding the role and any specific assessments, and any feedback on a candidate's progress through the selection process. Access to the assessment itself is also a potential source of inequity. For example, in the US, some minority groups and young adults are disproportionately reliant on mobile phones as their primary household computer and/or their only tool to access the internet, and therefore to apply for jobs and take assessments. Accordingly, assessments that are not mobile-ready – or where mobile versions are not equivalent to desktop versions (as with many mobile cognitive ability assessments, Arthur, Doverspike, Muñoz, Taylor, & Carr, 2014) – will not provide equal treatment to all candidates.

Modern technology can enable the automation of information and feedback presented to candidates. This can reduce the burden for recruiters and can help to standardize the treatment of each candidate (or employee). However, be careful not to create an experience that is perceived as too impersonal or mechanical. Maintaining some degree of human-human interaction, where appropriate, remains important for the overall candidate experience.

Equal Access to the Constructs Being Measured

All candidates should have equal opportunity to demonstrate their ability on the constructs being assessed for a particular role. This concept is also known as the accessibility of an assessment. Accessibility is an important concept, as ideally scores on an assessment are largely a result of a candidate's true score on the construct measured by an assessment (always with a little measurement error). However, some candidates may have difficulty taking an assessment, often due to the method (e.g., spoken versus written response) or technology used, and this can unduly impact their score on the assessment.

For example, if a role is determined to require skills in written communication, then candidates should have equal access to demonstrate their written communication skills in the assessment for this role. If a candidate, with strong written communication skills, who also has a visual impairment, takes an online written communication assessment that is not tailorable to their needs (e.g., increasing font size), then their score on this assessment may be more influenced by these technology limitations than their actual ability. This would present an accessibility problem. In this scenario, the candidate does not have an equal opportunity to demonstrate their ability. The designers and developers of the assessment should have foreseen this problem and designed the technology platform such that it can accommodate people with various visual impairments (e.g., ability to increase or decrease font size), thereby making the assessment more accessible.

When developing AI assessments to be as widely accessible as possible, designers and developers must pay particular attention to the new features that can be included in AI assessments (e.g., words used, facial expressions, voice intonation, behavior). The benefit of including these new features in AI assessments is that they enable candidates to provide information in a more natural way. For example, information regarding a candidate's relevant past experience can now be objectively measured through the candidate's spoken response to an AI-scored video interview question, instead of the candidate completing a multiple-choice biographical data inventory. In another example, a candidate could demonstrate their persuasion skills in an AI-scored simulation which assess their behavior and choice of words, instead of completing a multiplechoice Situational Judgement Test (SJT). Being able to respond in a way that is much more natural for a candidate can increase their enjoyment of the experience, and can result in assessments that have higher validity and fidelity. However, some individuals may be unjustly negatively impacted by the inclusion of these new features. For example, an individual who has a neurological condition that impacts their ability to communicate in social contexts (e.g., a neurodiverse candidate) may have difficulty with an AI-scored video interview.

Therefore, AI assessment developers, and users, should be aware that some of the benefits of AI assessments (e.g., enabling a natural and open response format) can bring unintended challenges for some candidates and employees (and that this is also true for more traditional assessment formats). To the extent possible, the design of an AI assessment should take these challenges into consideration, and if the assessment is not able to be designed in a way that these challenges are mitigated or removed, accommodations should be made available. These accommodations might involve allowing a longer response time, human scoring of a recorded interview, or the identification and offering of a suitable alternative method of assessing the same necessary constructs, among other options. When considering whether the AI assessment can accommodate a wide range of individual preferences and abilities, developers should take into account, and may even consult with, the potential user audience (including those with special needs or disabilities or those at risk of exclusion).

It is important to note that while the development of an AI assessment can involve unique accessibility considerations, this is not necessarily always the case. Some AI assessments use AI to score responses on traditional questionnaires and tests. With this form of AI assessment, the accessibility considerations will be no different than they were for the traditional assessments.

The first step in assisting individuals who might have difficulty taking an AI assessment is to provide enough information about how the assessment works, what it measures, and the technology involved. Candidates taking the assessment can use this information to determine whether their disability or medical condition might negatively impact their performance on the assessment. A description of how to request an accommodation should be easily visible to candidates.

Non-Discriminatory Hiring Practices

As with all employment-related assessments, AI assessments should not result in biased hiring decisions. Multiple countries have legal regulations and guidelines that prohibit the use of any assessment that is found to discriminate against various groups. AI assessment developers and users should be familiar with the regulations and guidelines that apply to their area of operation. Table 5 presents an example of some of these regulations at the time of publication of this paper.

Table 5. Examples of Hiring-Related Legislations

Australia	Age Discrimination Act 2004, Human Rights Commission Act 1986, Disability Discrimination Act 1992, Racial Discrimination Act 1975, Sex Discrimination Act 1984
European Union	Race Equality Directive, 2000; Equality Framework Directive, 2000; Equal Treatment Directive, 2006
South Africa	Employment Equity Act, 1988 and Labor Relations Act, 1995
United Kingdom	Equality Act 2010
United States	Uniform Guidelines 1978, Americans with Disabilities Act, 1990, Age Discrimination in Employment Act, 1967

It is important to understand what is meant by the term bias when applied to employment assessments. In traditional assessment development and validation, bias is distinguished from adverse impact. Adverse impact refers to differences in group outcomes (e.g., different pass rates for different groups). In the Uniform Guidelines on Employee Selection Procedures (EEOC, CSC, DOL, & DOJ, 1978), adverse impact is defined as a substantially different rate of selection in hiring, promotion, or other employment decision that indicates a disparate impact of the selection procedure on members of a specific group (e.g., based on race, sex, ethnic, age, disability). A “substantially different” rate is typically defined in government enforcement or Title VII litigation settings using the adverse impact ratio (AIR), statistical significance tests, and/or practical significance tests. The AIR is the ratio of the selection rate of a protected group to the selection rate of the unprotected group, which is usually evaluated using the four-fifths rule. The four-fifths rule is based on the rates at which job applicants pass an assessment. For example, if 50 percent of the men pass while only 40 percent of the women pass, one could look at the ratio of those two passing rates to judge whether there might be a discrimination problem. The ratio of 40:50 means that the rate of passing for female applicants is only 80 percent of the rate of hiring for male applicants. An AIR below 80% is typically regarded as a meaningful level of adverse impact.

When interpreting group differences, it is important to note that findings of adverse impact do not violate EEOC guidelines, provided the characteristic being measured is job relevant and no other assessments are available that measure the same construct with less adverse impact. If the assessment demonstrates a predictive relationship with job-related criteria, it is legally defensible (e.g., Uniform Guidelines on Employee Selection procedures; EEOC et al., 1978, Section 4D). For example, a test of upper body strength will likely have adverse impact on women compared to men, but it may measure a construct that is necessary for a job such as firefighter.

In contrast, the term bias in assessment development and validation has traditionally referred to “systematic error in a test score that differentially affects the performance of different groups of test takers (SIOP, 2018; p. 39). This error could be measurement bias (irrelevant variance in the assessment, or the criterion it is intended to predict, that negatively impacts scores for a particular group), or predictive bias (irrelevant variance affecting predictor-criterion relationships, such that the prediction of criterion scores from assessment scores differs for different groups). These types of bias would put an assessment user at legal risk.

In the context of AI, however, the term “bias” is commonly used to refer to an AI algorithm that results in discrimination against a group, regardless of whether that discrimination is fair or unfair. Such an algorithm is considered biased or having bias. Thus, in the world of AI assessment development, there is no distinction between adverse impact and predictive or measurement bias. The goal is to minimize group differences for any reason, although in some cases this goal may be impossible to meet while still measuring the most important aspects of the job.

The techniques used to test for bias in traditional assessments can, and should, also be applied to AI assessments. For example, the four-fifths rule can easily be applied to scores resulting from AI assessments. There is no reason why the traditional techniques used to assess for bias in an assessment cannot be applied to AI assessments, regardless of how sophisticated a particular AI assessment might be. During the development and deployment of the AI assessment system, mechanisms for flagging issues relating to bias, discrimination or poor performance of the AI system can be embedded as well as ways in which the fairness of the system is monitored and ensured.

AI Assessments & Workforce Representativeness

In addition to following the regulations regarding non-discriminatory hiring practices to reduce legal risks, organizations are increasingly interested in the use of fair hiring practices to increase the representativeness of their workforce (LinkedIn, 2018). AI assessments can help to support and advance initiatives to improve representativeness within organizations by reducing subjective human judgment, thereby minimizing human error and bias. However, done improperly, AI assessments can unintentionally counter these goals. For example, developing an AI assessment that learns to predict hiring decisions could result in a biased assessment, should these past hiring decisions have been made with a degree of bias (e.g., members of one group were hired significantly more often than members of another group).

Therefore, to achieve the potential benefit of AI to reduce bias, AI assessments must be designed for fairness from the beginning. This consideration of fairness should be present throughout all stages of the development process, instead of relying on a single test for bias after the assessment has been developed. Fortunately, just as an AI assessment can “learn” how to best predict a work-related outcome (i.e., criterion), it can also learn to avoid bias. By designing an assessment to be both valid (effective) and fair (ethical) from the very beginning, the risk of bias is greatly reduced. The following steps may be taken during the development and evaluation of an AI-based assessment to help minimize potential bias:

1. Bias can emerge through language or contexts that may be relatively inaccessible to a particular group, or an item may offend individuals from certain groups or make them uncomfortable (e.g., by promoting a particular stereotype). Multicultural experts should review all questions, scoring rubrics, or other text for potential bias, offensiveness, or cultural effects.
2. If assessment scores are generated by raters, all raters should complete extensive training on how to properly use the scoring rubrics to provide ratings of responses that are as objective as possible. Responses should be rated by at least two raters. Multiple raters help to remove individual biases that may exist. Depending on the assessment, it may be wise to provide only the audio of the responses to raters to make sure that facial features or appearance do not introduce any bias.
3. When developing algorithms to score responses automatically, examine between-group score differences for evidence of adverse impact against protected groups. If an AI algorithm appears to exhibit adverse impact, we look at the constituent features in the algorithm to identify and remove those features that appear to be producing the unintended score differences.

An AI assessment that is successfully designed and developed with fairness in mind from the beginning will require input and oversight from SMEs in talent assessment. These SMEs will be able to inform the technological developers of the AI assessment on which features are likely to be 1) job related, and 2) pose a risk of bias (e.g., including pitch of voice as a predictor variable, which has a high correlation with gender). In addition to SME input, rigorous and continued testing throughout the development process is crucially important to prevent bias from creeping into an AI assessment.

Section Summary

SHL recommends that AI assessments are designed with fairness in mind from the beginning. To achieve this, consideration must be paid to each of the three dimensions of fairness presented in this section. An AI assessment should result in the equal treatment of all candidates, provide equal access the constructs being measured, and result in employment decisions that are minimize bias.

Design for Fairness: Best Practice
Begin the development of an AI assessment with fairness in mind. Take steps to proactively remove or reduce bias from the AI assessment during its design and development. Do not rely only on a single test of bias (e.g., the four-fifths rule) after the AI assessment has been developed.

6.IV Rigorously Validate

The thorough validation of AI assessments is a key requirement in the best practices and guidelines in talent assessment (e.g., SIOP Principles) and AI (e.g., the EU's Ethics Guidelines for Trustworthy AI). Validation is a process through which an assessment's validity is examined. The Standards define validity as "the degree to which evidence and theory support the interpretations of test scores for proposed uses of tests" (AERA et al., 2014, p. 11). This definition of validity is also supported by SIOP.

An important point mentioned in this definition is that both evidence and theory support interpretations of test scores. This means that the optimal demonstration of validity involves both quantitative evidence of the assessment's relationship with key variables (e.g., a correlation coefficient demonstrating a significant relationship between assessment scores and outcomes of selected candidates) and a supporting rationale explaining why this relationship exists. The Standards also note that validation starts with "an explicit statement of the proposed interpretation of test scores, along with a rationale for the relevance of the interpretation to the proposed use. The proposed interpretation includes specifying the construct the test is intended to measure" (AERA et al., 2014, p. 11). Essentially, according to The Standards and the SIOP Principles, the use of scores from an assessment to make decisions regarding potential or current employees requires a strong rationale and specification of the constructs being measured, in addition to any evidence regarding the empirical relationship between assessment scores and work-related outcomes (e.g., job performance)[4].

This can present a dilemma with AI assessments that include new features (e.g., words used, facial expressions, vocal intonation), as these features currently have relatively little research and theory that supports their use (i.e., their job relatedness). This does not mean that the inclusion of these predictors in AI assessments is not justified, it means only that the scientific research has yet to catch up with the application of these technologies. Because of this, and the evolving guidelines and regulations regarding AI (see section 4), a thorough validation process is required when developing an AI assessment. SHL believes that this validation process should be conducted to at least the same standard that is required for traditional assessments.

When conducting a validation study for an AI assessment, or any assessment, it is important to remember that it is the resulting scores, and not the assessment procedure itself, which are being investigated (see AERA definition of validity, above). For example, scores from an assessment that measures mechanical comprehension may significantly predict future performance for engineering roles, but not for sales roles. When assessing the validity of this mechanical comprehension assessment for sales roles, it is not the assessment itself which is not valid (as it is a valid measure of the construct of mechanical comprehension) but rather the use of scores produced from this assessment for selecting individuals for sales roles. Likewise, if an AI-based simulation is determined to have sufficient validity evidence supporting its use to assess candidates for a particular job, it is not the simulation method itself that is found to be valid, but that particular derivation and application of scores.

[4] Some assessments are designed in a way that the items do not necessarily measure a construct but instead are used to directly predict a criterion (e.g., biographical data scales). This approach has been known as "empirical keying." However, even with empirically-keyed biographical data scales, calls for the inclusion of job relevance and SME input into the items have long been recommended (Pace & Schoenfeldt, 1977).

Validating an AI Assessment

The validation of an AI assessment, as with any other form of assessment, requires the use of methods that are supported by current legal and professional standards. Four such methods of validation are:

- 6.IV.a Content-related validation
- 6.IV.b Construct-related validation
- 6.IV.c Criterion-related validation
- 6.IV.d Generalizing validity evidence

The choice of which of these four methods to use in the validation of an AI assessment will depend on a number of factors, and input from talent assessment SMEs will be required in selecting the right method for a given situation. A description of each of these methods is presented below.

6.IV.a Content-Related Validation

A content-related validation strategy focuses on demonstrating that the content of the assessment (i.e., the features or constructs being assessed) is relevant to the work requirements of the target job. This typically involves input and judgment from SMEs. For example, a word processing assessment can be validated for an administrative assistant role via the consensus from a panel of SMEs that the operation of the word processing software, as measured by the assessment, is an important requirement of the job. Content-related validation relies on this SME-based evidence of the correspondence between the tasks or competencies measured by the assessment, and the tasks or competencies performed on the job.

Using a content-related validation strategy to validate an AI assessment would involve the conclusion, from a group of SMEs, that the features measured by the AI assessment were relevant to the work requirements of the target job. For example, an AI assessment designed to measure communication skills could be validated for a customer service job via a thorough examination and documentation, by SMEs, of the required competencies for that role, and a decision that the communication skills measured by the AI assessment are appropriately reflective of one or more of the required competencies. The SMEs required to judge the appropriateness of some of the novel features capable of being included in AI assessments, such as word choice and facial expressions, would be expected to possess experience and expertise in areas such as linguistics and human micro expressions.

A final note on content-related validation – because the constructs and/or characteristics measured by an AI assessment must be known for the SMEs to assess their job relevance, black box assessments will not be able to demonstrate content validity.

6.IV.b Construct-Related Validation

A construct-related validation strategy focuses on demonstrating that the assessment accurately measures the target construct(s). This demonstration typically involves a SME's judgment of sufficient evidence supporting inferences of measurement. One common means of producing construct validity evidence is through quantitative analyses comparing scores from the new assessment with scores from previously validated measures of the construct, or closely related constructs, and scores from previously validated measures of unrelated constructs (known as convergent and discriminant validity analyses, respectively). To the degree that scores from the new assessment demonstrate a relationship with measures of the target construct, and show a relative lack of a relationship with measures of unrelated constructs, then the new assessment can be viewed as having some evidence of construct validity.

In the context of AI assessments, construct-related validation evidence would demonstrate that the assessment is measuring the constructs that it is designed to measure. For example, if a multimedia based simulation was developed to assess aspects of a candidate's personality, then scores from the assessment could be compared with scores from a personality questionnaire to examine how closely the scores are related. Typically, construct-related validation studies are conducted when the constructs that an assessment is designed to measure are known in advance.

6.IV.c Criterion-Related Validation

A criterion-related validation strategy relies on demonstrating a quantitative relationship between assessment scores and a criterion (e.g., employee performance). This relationship is expressed via a statistical metric (often a correlation value), and the value of the metric determines the validity of the assessment. This is one of the strongest and, therefore, most preferred methods for demonstrating validity.

Table 6 shows five important factors to consider when conducting a criterion-related validation study. These factors are standard considerations when developing any type of assessment, but particular consideration should be paid to them when developing an AI assessment that incorporates novel features. This is due to the current lack of research and theoretical explanation for why some of these features predict job-relevant outcomes. Due to this lack of theoretical support, some AI assessments will need to rely heavily, or even solely, on criterion-related evidence. Therefore, this evidence must be robust and able to withstand close scrutiny.

Table 6. Important Considerations for the Criterion-Related Validation of an AI Assessment.

Job analysis	A thorough job analysis is conducted.
Quality criterion	A job-specific performance measure is developed.
Sample size	Large enough to provide sufficient statistical power and allow hold-out samples for cross validation. Sample sizes may need to be very large, compared to the requirements for traditional assessments.
SME input	Job analysis, criterion development, and predictive feature selection are conducted with SME input.
Cross validation	The performance of the AI assessment is tested in one or more holdout samples.

The following descriptions of each of the elements in Table 6 will provide guidance on how to design a strong criterion-related validation study for an AI assessment.

Job analysis

A job analysis provides the foundation for the demonstration of validity evidence for any assessment. The design of a criterion-related validation study for the development of an AI assessment should begin with a thorough job analysis. The job analysis will provide insight into the competency and performance requirements for a given role, and will therefore inform the choice of criterion, as well as the constructs and features to be measured by the AI assessment.

Quality criterion

The criterion refers to the metric that is used for the outcome variable that the AI assessment is designed to predict, such as job performance or turnover risk. The quality of the criterion in a criterion-related validation study is crucially important. Inappropriate criterion measures, or criterion measures of poor quality, will affect the conclusions drawn regarding the validity of an AI assessment (e.g., an assessment that is not a valid predictor of actual job performance, could be mistakenly determined to have appropriate validity).

Essential requirements for selecting a high-quality criterion metric include: 1) that the choice of criterion has been informed by a job analysis, 2) the criterion represents an aspect of on-the-job outcomes[5], and 3) the criterion is collected via a measure that has been tailored to the job in question[6].

If a criterion-related validity study is the only method used to demonstrate the validity of an AI assessment (which will often be the case, due to the previously mentioned lack of theoretical support for features such as choice of words, voice intonation, etc.), then the quality of the criterion is critical. An AI assessment that produces scores that have a strong relationship with a criterion should still be looked at with suspicion, if this criterion is not representative of the job's performance domain.

Sample size

When conducting a criterion-related validation study, it is important to have a sample size that is large enough to provide stable estimates of coefficients and to test for their significance. While this is ultimately a complex statistical question affected by expected effect sizes, number of features, and other considerations, some informal guidelines frequently used to determine adequate sample sizes for an assessment validation study may range from about 100 to 300, although this number may be higher or lower depending on the type of AI algorithm, the number of variables included, and the data analyses required.

In following this basic guideline, validation studies of AI assessments may require a sample size that is an order of magnitude larger than those of traditional assessments, because AI assessments often contain many more features than traditional assessments (e.g., the words a person speaks during a recorded interview versus items on a personality questionnaire). While the specific requirement will vary by situation, the general guidance when conducting a criterionrelated validation study for an AI assessment is to use sample sizes that are larger than what may be sufficient for traditional assessments.

[5] As opposed to a metric that represents pre-hire outcomes, such as hiring decisions (i.e., hired vs not hired).

[6] Customized performance measures are recommended over existing performance data in an HRIS as such performance data was not designed for predictive modeling and often lacks sufficient variance.

SME input

While the ultimate test of validity in a criterion-related validation study relies on the empirical relationship between assessment scores and the criterion, the input of SMEs is still crucially important throughout the design and development of the assessment.

When developing an AI assessment, the early and continued input of SMEs, including talent assessment professionals and any other experts that may be required, can result in an assessment with higher validity and lower bias, than would be expected without the input of these individuals. This can become particularly evident during the cross-validation stage, where features that were identified as job-related often perform better compared to features that were chosen solely based on their relationship with the criterion in the training data.

Cross validation

Cross validating an assessment involves testing the validity of an assessment with new data. Typically, the validity will be reduced when an assessment is tested in a different dataset than the one it was built upon. Therefore, cross validation analyses are an important way of making sure that scores from an assessment will remain valid when used in operation to assess candidates.

When designing a criterion-related validation study for an AI assessment, the incorporation of rigorous cross validation is essential. This is due to the large number of features often included in an AI assessment, some of which may lack strong theoretical support. Without a thorough cross validation, it will be very difficult to know which features are truly related to the criterion, and which are not (i.e., it will be difficult to identify Type I errors). This can lead to a false degree of confidence in the expected performance of an AI assessment. Reports of validity coefficients for AI assessments should include results from at least one cross validation study.

Criterion-related validation studies that meet these recommendations will be stronger and, therefore, carry more weight in demonstrating the validity of an assessment. The importance associated with each of these factors may vary depending on the situation (e.g., type of AI assessment, type of role, assessment designed for selection versus development).

6.IV.d Generalizing Validity Evidence

Another option for investigating the validity of using scores from an assessment to make decisions on candidates for a particular role is to demonstrate evidence of generalizability. Generalizability is a term that has a similar conceptual meaning in both talent assessment and AI, although there are some important differences. In essence, generalizability refers to the degree to which an assessment or AI algorithm is expected to perform in a new scenario – that is, with "unseen" data. In AI, tests for generalizability typically involve the cross validation of an algorithm, in which the predictive accuracy of the algorithm is assessed in the test or holdout sample. In talent assessment, generalizability refers to the application of existing validity evidence from the use of an assessment in a particular job, or jobs, to a new target job, based on similarities between those jobs. There are multiple ways that evidence for the generalizability of an assessment can be determined (e.g., meta-analysis, transportable validity study, synthetic validity). The focus of this validation strategy is the demonstration of a high degree of similarity between the job in question (typically understood to be a collection of competencies and the requisite KSAOs and tasks), and the job(s) for which the assessment has been determined to be valid.

Generalizability is a concept with many specific forms. In one form, generalizability might suggest that the cumulative amount of evidence for a relationship is so strong that further evidence in additional contexts or roles is not necessary and obviates the need for additional data collection and research efforts. In another form, generalizability may refer more narrowly to transporting validity evidence from one job to another based on overall similarity of the jobs. Even more narrowly, generalizability could simply refer to some component of a job that is demonstrated to be similar in another job even if, in total, the jobs might be considered substantially different.

For example, if an AI assessment that measures general coding proficiency has sufficient criterionrelated validity evidence for use in selecting candidates for website developer roles, this assessment could also be used to identify strong candidates for data scientist roles – if it can be demonstrated that those data scientist roles also require strong general coding skills. Generalizing validity evidence from an existing AI assessment to a new role will typically, as a best practice, require input from SMEs. In all cases, however, the core concept of generalizability involves the use of prior data to make inferences about the suitability of a predictive model or specific elements of such a model, to new situations and contexts.

Section Summary

SHL believes that AI assessments should be validated to a high standard that meets or exceeds existing guidelines and best practices, and should be designed and developed with guidance and oversight from SMEs in talent assessment.

Rigorously Validate: Best Practice
AI assessments should be held to the current standards and guidelines regarding the evidence required for demonstrating validity. When conducting criterion-related validation studies for AI assessments (which are highly encouraged), careful consideration should be given to each of the five factors presented in Table 6.

6.V Incorporate Human Oversight

One of the key benefits of AI is the ability to analyze vast amounts of data and detect patterns that would be difficult for a human to detect. When developing AI assessments, this means that a large amount of data, containing information about candidates, can be “fed” into an AI algorithm and used to maximize the prediction of employee outcomes. This can all happen without much insight from the AI assessment developers as to how the AI is making decisions. As discussed in section 6.II of this paper, this would be an example of an AI assessment with low transparency. Ideally, the inner workings of any AI assessment should be relatively explainable to a human. However, merely knowing how an AI assessment makes decisions is not sufficient. Humans should have a degree of oversight into the decisions that an AI assessment is making.

The current guidelines and regulations regarding the use of AI support this proposition. For example, guidelines from Europe state that *“all individuals have the right to a final determination made by a person” (The Public Voice, 2018), and that “proper oversight mechanisms need to be ensured, which can be achieved through human-in-the-loop, human-on-the-loop, and human-in-command approaches”*[7] (European Commission’s Guidelines for Trustworthy AI, 2019). Under EU GDPR, if a decision is taken by solely automated means (i.e., there is no meaningful human input into the decision), then an individual has the right to request human intervention, to express their point of view and to contest the decision.

The amount of human oversight required for a particular AI assessment will vary. On the lowest end of the spectrum, the AI could be free to make decisions without any human oversight (e.g., the AI sets the cut scores that determine who passes the assessment, or even choose which of those individuals who passed the cut score should be selected to continue in the process). The AI is completely free to act. On the higher end of the spectrum, human oversight can be built into an AI assessment such that any decision the AI makes must first be approved by a human. In practice, the amount of human oversight required for most AI assessments will fall somewhere in the middle of this spectrum.

The human oversight of an AI assessment can occur during both the development and ongoing use of the assessment. Developers oversee the creation and validation of the AI assessment, may set the cut scores which determine the candidates’ results from taking the assessment, and also set the parameters under which the AI can act (e.g., with full autonomy, or with some human oversight). The assessment users (e.g., a recruiter or hiring manager) provide oversight of an AI assessment by using the AI’s recommendations as information which is then combined with information from other sources to make a decision. For example, if an AI assessment predicts that a candidate has high potential to be a good performer in a particular role, but the hiring manager disagrees based on other information available on the candidate, then the hiring manager is free to intervene and choose not to hire this particular individual. The reverse is also true. If an AI assessment decides that a candidate is not a good fit, but the recruiter thinks otherwise, they can override the AI’s decision. These two examples demonstrate the design and development of an AI assessment that has human oversight. In these scenarios, the number of times that the human user chooses to override the recommendation by the AI assessment may be quite rare and occur only under specific circumstances (e.g., a defined exception or escalation process). However, the fact that the output from the AI is used as a recommendation, and not the final decision, makes all the difference – as a human has the opportunity to intervene when needed.

[7] These different approaches involve varying levels of human oversight and control over the AI system, with human-in-command having the most human control, human-in-the-loop requiring human approval before an AI makes a decision, and human-on-the-loop allowing AI to automatically make decisions, still with some human oversight. The choice of approach will depend on the use case and severity of the decision (e.g., deciding who is selected for a job versus providing individual feedback for development purposes).

Section Summary

AI assessments should be designed to provide information that is used, along with information from other sources (when applicable), by a human to make decisions regarding current or potential employees of an organization. AI assessments should not be designed to make these decisions without human oversight.

Regardless of guidelines or regulatory requirements, having human oversight of an AI assessment is good practice for an organization, its employees, and society at large. AI assessments should be developed with human oversight throughout the entire process – data gathering, data cleaning, feature extraction and development, model training and testing, and model deployment.

Incorporate Human Oversight: Best Practice
Design AI assessments to have human oversight throughout its development and during its deployment. No AI system should make a final decision in a highstakes situation (e.g., determining who to hire) without the possibility of human intervention.

6.VI Disclose Intent

Candidates participating in an AI assessment should be provided with sufficient detail regarding what constructs or features they are being assessed on, and how the AI works. Providing candidates with such information is considered good practice and increases the candidate's assessment experience. In some cases, collecting formal candidate consent is legally required. In particular, consent may be required for video assessments or for any use of AI that results in automated decision making (see section 4). Consent, where required, must be freely given (e.g., under GDPR). Candidates that do not provide their consent should be provided with an alternative method of assessment that measures the same constructs. A candidate's decision to not provide consent may not be detrimental to his/her chances of being selected.

For example, recent legislation in the U.S. state of Illinois regarding the use of AI video interviews requires the following:

1. A notification that AI will be used to analyze responses,

2. An explanation of how the AI works, and the characteristics it uses to evaluate the candidate, and

3. Obtaining consent from the candidate for their recorded video to be evaluated by AI.

Similar regulations exist in Europe (e.g., GDPR). Under the GDPR, there is a high threshold for what constitutes candidate consent: it needs to be specific, informed, freely given and unambiguous. The responsibility for collecting the consent will rest with the controller of the collected information, usually the employer. As with the Illinois law, the requirement for formal consent is likely to be implemented in other regions and countries which adopt GDPR-like laws.

In situations where candidates are required to be asked for formal consent, and do not consent to be evaluated by AI, they must be offered an alternative means of assessment or evaluation and it must be readily available. For example, if a candidate does not provide consent for their recorded video interview to be assessed by AI, then their responses to the interview questions should only be assessed by a trained human rater.

As mentioned previously, the guidelines and regulations involving the use of AI in talent assessment are still evolving. While not all regions, countries, or states require that consent be obtained from a candidate before scoring their assessment responses using AI, SHL recommends that AI assessment developers and users hold a high standard, and therefore always notify candidates where AI is used, how it is used, and determine whether collecting consent for the use of AI is required. Additional legal regulations, where they exist, should also be confirmed to inform best practice regarding candidate consent to be scored by AI.

Disclose Intent: Best Practice

- Notify candidates that AI will be used to analyze responses
- Explain how the AI works (e.g., the constructs/ features measured - for example, through a privacy notice)
- Obtain consent to be assessed by AI where and when required
- Provide alternative assessment to those who do not provide consent

7. Looking Forward

The arrival of AI assessments promises to revolutionize talent assessment, however, just how much of an impact AI will have on the field remains to be determined. It is possible that the current validity ceiling will be surpassed, bias reduced, and assessments will become more engaging and enjoyable for candidates. Regardless of the exact outcomes, SHL believes that talent assessment is about to make its next great evolutionary leap with the incorporation of AI.

Beyond improvements in talent assessment and selection, AI will bring additional benefits to HR via the increased efficiency and automation of tasks, and the ability to make informed strategic decisions from the ever-increasing amount of data that HR has access to (SHL, 2018). These benefits will in turn enable HR to continue to deliver more and more value to organizations. This is all expected to occur over a relatively short time period.

Over a longer term, expect to see the ever-increasing sophistication of AI assessments, taking talent assessment into the worlds of both augmented and virtual reality. In these virtual worlds, candidates will be able to not only speak their responses in a natural way, but they will also be able to move and behave in a natural way. AI assessments developed to successfully harness the combination of new technology, AI, and assessment science could result in incredibly rich and high-fidelity simulations that further redefine the thresholds for acceptable levels of validity and candidate experience.

However, for AI assessments to deliver on these promises, they must be developed, and used, according to strong guiding principles and practices. As legal regulations continue to develop around the world, the inappropriate use of AI in assessments could lead to legal and ethical violations, which could substantially impede the development of AI assessments. Therefore, in our attempt to address the rapidly evolving and complex landscape of AI in talent assessment, SHL has developed the Core Principles and Best Practices contained in this document.

Disclaimer: The contents of this document have been prepared in good faith as an introduction. It is not intended to provide legal advice, and you should not take, or refrain from taking, action based solely on its content. Neither SHL nor its affiliates accept any liability for any errors or omissions, nor for any loss incurred by placing reliance on the contents of this document. We do not owe a duty of care to any party in relation to the preparation of this document. Any person reading this document should use their own professional judgment regarding its content and should also seek independent advice specific to their circumstances.

Note: Throughout this document, italicized words or phrases are defined in the Glossary of Terms.

References

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2014). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association.

Arthur Jr, W., Doverspike, D., Muñoz, G. J., Taylor, J. E., & Carr, A. E. (2014). The use of mobile devices in high-stakes remotely delivered assessments and testing. *International Journal of Selection and Assessment*, 22(2), 113-123.

Artificial Intelligence Video Interview Act, Illinois HB 2557.

Bartram, D. (2005). The great eight competencies: A criterion-centric approach to validation. *Journal of Applied Psychology*, 90(6), 1185-1203.

Beaty, J. C., Dawson, C. R., Fallaw, S. S., & Kantrowitz, T. M. (2009). Recovering the scientist–practitioner model: How IOs should respond to unproctored internet testing. *Industrial and Organizational Psychology*, 2(1), 58-63.

Beaty, J. C., Nye, C. D., Borneman, M. J., Kantrowitz, T. M., Drasgow, F., & Grauer, E. (2011). Proctored versus unproctored internet tests: Are unproctored noncognitive tests as predictive of job performance? *International Journal of Selection and Assessment,* 19(1), 1-10.

Bologna, M.J. (2019, May 30). 'Hiring robots' restrictions passed by illinois legislature. Bloomberglaw. https://news.bloomberglaw.com/daily-laborreport/ hiring-robots-restrictions-passed-by-illinois-legislature.

Burke, E. (2009). Preserving the integrity of online testing. *Industrial and Organizational Psychology*, 2(1), 35-38.

DeGroot, T., & Gooty, J. (2009). Can nonverbal cues be used to make meaningful personality attributions in employment interviews? *Journal of Business and Psychology,* 24(2), 179-192.

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.

European Parliament and Council of the European Union (2016). Regulation on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (Data Protection Directive). L119, 1-88.

Gonzalez, M. F., Capman, J. F., Oswald, F. L., Theys, E. R., & Tomczak, D. L. (2019). "Where's the IO?" artificial intelligence and machine learning in talent management systems. *Personnel Assessment and Decisions,* 5(3), 33-44.

Grelle, D. M., & Gutierrez, S. L. (2019). Developing Device-Equivalent and Effective Measures of Complex Thinking with an Information Processing Framework and Mobile First Design Principles. *Personnel Assessment and Decisions,* 5(3), 21-32.

High-Level Expert Group on Artificial Intelligence. (2019). *Ethics guidelines for trustworthy artificial intelligence*. Office for Official Publications of the European Communities.

Liem, C. C., Langer, M., Demetriou, A., Hiemstra, A. M., Wicaksana, A. S., Born, M. P., & König, C. J. (2018). Psychology meets machine learning: Interdisciplinary perspectives on algorithmic job candidate screening. In H. J. Escalante, S. Escalera, I. Guyon, X. Baró, Y. Güçlütürk, U. Güçlü, & M. van Gerven, (Eds.) Explainable and interpretable models in computer vision and machine learning (pp. 197-253). Cham, Switzerland.

LinkedIn. (2018). Global recruiting trends 2018: The 4 ideas transforming how you hire. Linkedin. https://business.linkedin.com/talent-solutions/ recruiting-tips/2018-global-recruiting-trends.

Mercer. (2019). *Global Talent Trends 2019*. Mercer.

Pace, L. A., & Schoenfeldt, L. F. (1977). Legal concerns in the use of weighted applications. *Personnel Psychology*, 30(2), 159-166.

Poole, D. L., & Mackworth, A. K. (2010). Artificial intelligence: *foundations of computational agents*. Cambridge University Press.

Putka, D. J., Beatty, A. S., & Reeder, M. C. (2018). Modern prediction methods: New perspectives on a common problem. *Organizational Research Methods*, 21(3), 689-732.

Sheng, E. (2019, May). The job postings enticing workers the most are being written by A.I. CNBC. https://www.cnbc.com/2019/05/24/job-postings-youwant-most-will-be-written-by-artificial-intelligence.html

Society for Industrial Organizational Psychology, Inc (2018). *Principles for the validation and use of personnel selection procedures* (5th ed.) Bowling Green, OH.

The Public Voice. (2018). Universal guidelines for artificial intelligence. The Public Voice. https://thepublicvoice.org/ai-universal-guidelines.

Weintraub, A. (2017, June 5). What you should know about HIPAA Compliant Servers. Medstack. https://medstack.co/blog/what-you-should-know-abouthipaa-compliant-servers.

Wisenberg Brin, Dinah. (n.d.). Employers Embrace Artificial Intelligence for HR. SHRM. https://www.shrm.org/resourcesandtools/hr-topics/global-hr/pages/employers-embrace-artificial-intelligence-for-hr.aspx

Glossary of Terms

Accessibility: In the context of talent assessment, refers to the design of an assessment so as to be useable by people with disabilities.

Accuracy: The proportions of predictions made by a machine learning model which were correct. Used as a metric for evaluating model performance.

Adverse impact: Refers to employment practices that have a discriminatory effect on a protected group, resulting in members of the protected group being chosen or selected at a rate deemed unfair relative to that of the reference group.

AI Assessment: In the context of talent assessment, refers to assessments which utilize AI. More specifically, “AI assessment” refers to any non-human analysis of participants’ responses that utilizes machine learning, NLP, or other related modeling approaches and techniques (e.g., deep learning, latent semantic analysis) to assign scores to attributes of people (e.g., KSAOs, competencies) or to individuals’ expected work outcomes (e.g., probability of turnover).

Algorithm: A process or sequence of steps followed by a computer to complete a task.

Artificial Intelligence (AI): A branch of computer science dealing with the simulation of intelligent behavior in computers.

Bias: In the context of talent assessment, bias refers to the qualities of an assessment that unfairly penalize a group of candidates due to their gender, race, ethnicity, age, disability status or other legally protected characteristic.

Biographical data: Information about an individual’s background, prior experiences, and behavior. This information can be used to assess candidates.

Job analysis: The systematic study and documentation of the tasks and responsibilities of a job, as well as the knowledge, skills, abilities, and other characteristics (KSAO) required to perform the job.

KSAO: This acronym refers to “knowledge, skills, abilities, and other characteristics”, which are also often referred to as competencies or elements of competencies.

Machine learning: An automated method of data analysis, pattern recognition, and model building, that can learn from data and make decisions with minimal human intervention.

Measurement error: The difference between an observed (or measured) value and the true value of any object under study.

Natural language: Any language that has developed naturally through use, as opposed to a computer language.

Natural Language Processing (NLP): A subfield of linguistics, computer science, and AI that studies the processing and analysis of natural language data.

Neural network: See “Deep Learning”.

Random forest: A modeling method in machine learning which combines the results of multiple decision trees to enhance the prediction of an outcome.

Realistic Job Preview (RJP): A method used during recruiting which provides information regarding the job to candidates, and is used to help candidates determine whether they might be a good fit for the job.

Situational Judgment Test (SJT): A type of assessment which presents candidates with hypothetical scenarios to which the candidate selects the most appropriate response.

Statistics: The science of collecting, analyzing, and drawing inferences from quantitative data.

Test data(set)/holdout sample: The data used to test (or cross validate) a model.

Training data(set): The data used to train a model.

User (of an assessment): In this document, the term “user”, when referring to an assessment, means an individual within an organization with a need for the assessment information (e.g., a recruiter or hiring manager), unless otherwise specified.

Validity: The degree to which evidence and theory support the interpretations of test scores for proposed uses of tests.

SHL provides trusted Talent Intelligence that empowers organizations to make confident data-driven people decisions. Built on 45+ years of expertise, our innovative assessment solutions combine proven behavioral science, predictive analytics, and transparent AI to deliver objective, actionable insights across the talent lifecycle – from hiring and development to internal mobility. We help leaders measure skills and potential, unlock performance, and mobilize talent to drive transformation.

shl.com