



# Guidelines for the Use and evaluation of Artificial Intelligence in Talent Assessment

SHL.







# Guidelines for the Use and evaluation of Artificial Intelligence in Talent Assessment

Copyright © 2023 Federal Authority for Government Human Resources & SHL Company.

**All Rights Reserved.**

Not allowed to reprint any part of this document or copy it in any form or means without obtaining prior written approval from the Federal Authority for Government Human Resources or SHL.

PO Box 2350 Abu Dhabi, United Arab Emirates T +971 2 4036000  
PO Box 5002 Dubai, United Arab Emirates T +971 4 2319000

# Contents

|   |           |
|---|-----------|
| About SHL"  | 6         |
| Acknowledgment  | 7         |
| Executive summary                                       | 8         |
| AI Human Resources Strategy                             | 15        |
| What is Talent Assessment?                              | 16        |
| What is Artificial Intelligence?                        | 17        |
| How can AI benefit talent assessment?                   | 20        |
| What risks are involved in using AI to assess talent?   | 22        |
| Legal Risk  | 22        |
| Public Relations Risk                                   | 27        |
| Effectiveness Risk                                      | 27        |
| <b>Question 1: How relevant are the training data?</b>  | <b>34</b> |
| Data quality  | 34        |
| Data quantity   | 35        |
| Data representativeness                                 | 36        |
| Data privacy and protection                             | 36        |
| Applying Question 1 to the Hypothetical Case Study      | 36        |
| AI Assessment Evaluation Sheet: Training data relevance | 37        |
| Reviewing data quality                                  | 39        |
| Reviewing data quantity                                 | 40        |
| Reviewing data representativeness                       | 40        |
| Reviewing data privacy & protection                     | 41        |
| Question 2: How does the algorithm make decisions?      | 42        |



|   |           |
|---|-----------|
| <b>Applying Question 2 to the Hypothetical Case Study</b>           | <b>43</b> |
| AI Assessment Evaluation Sheet: Transparency                        | 45        |
| <b>Question 3: How biased are the decisions?</b>                    | <b>46</b> |
| Applying Question 3 to the Hypothetical Case Study                  | 48        |
| AI Assessment Evaluation Sheet: bias                                | 49        |
| <b>Question 4: How valid are the decisions?</b>                     | <b>50</b> |
| Content validation  | 51        |
| Construct validation  | 52        |
| Criterion-related validation  | 52        |
| Applying Question 4 to the Hypothetical Case Study                  | 56        |
| AI Assessment Evaluation Sheet: Validity                            | 57        |
| <b>Question 5: How final are the decisions?</b>                     | <b>58</b> |
| Applying Question 5 to the Hypothetical Case Study                  | 60        |
| AI Assessment Evaluation Sheet: oversight                           | 60        |
| Question 6: How are candidates informed?                            | 61        |
| Applying Question 6 to the Hypothetical Case Study                  | 62        |
| AI Assessment Evaluation Sheet: informing the candidate             | 63        |
| Scorecard for the use of AI by government entities to assess talent | 64        |
| A look ahead – the future of AI in HR                               | 69        |
| References  | 70        |
| Glossary of terms   | 71        |



## About SHL”

SHL, global leader in HR technology and psychometric science, transforms businesses by leveraging the power of people, science, and technology.

SHL unrivalled workforce data and highly validated talent solutions provide organizations with the workforce and scale to optimally leverage their people’s potential that maximize business outcomes. SHL equip recruiters and leaders with people insights at an organization, team, and individual level, accelerating growth, decision making, talent mobility, and inspiring an inclusive culture. To build a future where businesses thrive because their people thrive.

With 45 years of talent expertise, SHL trusted technology partner to more than 10,000 companies worldwide, across more than 150 countries, including 50% of the Fortune Global 500 and 80% of the FTSE 100. For more information, visit **shl.com**.

# SHL.



# Acknowledgment

**Mohamed Farid**, Managing Director, MEA

**David Edwards**, Head of Talent Solutions, Middle East

**Diala Jarrar**, Business Development Manager, Middle East

**Sara Gutierrez**, Chief Science Officer, SHL

**James Meaden**, Previous-Senior Research Scientist, SHL

**Jeff Johnson**, Principal Research Scientist, SHL

الهيئة الاتحادية للموارد البشرية الحكومية  
Federal Authority For Government Human Resources



## Executive summary



The main objective of this guide is to present best practices, general guidance, and a framework that can be used by the United Arab Emirates' (UAE) federal government entities to evaluate systems and technologies that incorporate and utilize Artificial Intelligence (AI) to assess individuals for employment-related decisions. AI is a field of study with the general purpose of developing digital programs and machines that can display some properties that are similar to human-level intelligence or judgment. Machine learning (ML), a subfield of AI, involves the use of mathematical algorithms that are deployed and adapted to maximize the prediction of patterns of relationships within datasets. ML enables computer algorithms to learn from datasets without being specifically programmed.

Talent assessment is a broad term for the process or activity of assessing key skills, traits, experience, and competencies of individuals (i.e., job applicants or employees) and using this information to make informed employment-related decisions. Talent assessment can occur either when an individual is applying to a job, known as pre-hire, or after an individual is hired and for development purposes, known as post-hire assessment. There are many different methods and forms of assessments, some

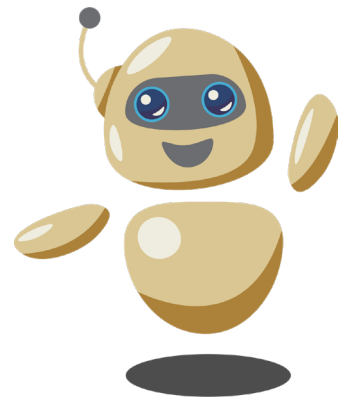






requiring little technology, such as an in-person interview, others as sophisticated as a highly realistic virtual reality simulation of on-the-job performance.

Like many other industries, talent assessment is experiencing a period of growth and innovation through the application of AI technologies. AI is enabling enhanced scoring of traditional assessments, with the promise of making the assessments more accurate, and is also enabling the development of novel assessments that capture information from candidates that was not available previously, such as video and audio data. The ability to include spoken language and video information in an assessment opens the possibility of developing highly realistic simulations of work scenarios to be used in an assessment. AI methods can also help to streamline the assessment process, enabling the candidate to have a better experience.



Although AI technology offers a variety of potential benefits, any kind of talent assessment carries some risk. This is also the case for assessments that utilize AI methods. Some of the risks associated with assessments that utilize AI methods are legal risk, public relations risk, and effectiveness risk.

**The legal risks associated with the use of assessments that utilize AI fall under two categories:**

- (A) Data protection
- (B) Bias

Many countries have implemented legal regulations that outline the permitted use of applying AI methods to analyze data about individuals. The UAE issued Personal Data Protection Law to enable artificial intelligence systems in the country and support their adoption. These regulations are related to legal protection surrounding the collection and use of personal data.






Bias in an assessment occurs when the assessment unfairly discriminates against an individual based on one or more of the individual's characteristics or background (e.g., race, ethnicity, religion, country of origin, gender, age, disability status). One of the benefits of AI methods in talent assessment is that they can lead to a reduction in such bias by reducing the influence of subjective human judgment. Incorporating AI into an assessment could also increase and solidify bias, however, if not done correctly and with expert oversight.

The use of AI methods in talent assessment also carries risks to the public's perception of the organization if the AI methods are not designed and developed in a careful and considered way. For example, candidates who are assessed using an AI-based assessment that they determine to be unfair or intrusive may share these negative reactions with their network and on social media. Such negative perceptions and reviews may damage the reputation of the organization and could have a negative effect on the quantity and quality of applicants to the organization.




Effectiveness risk is the risk of using an AI-based assessment that does not work as expected. The complexity of data sources and modeling techniques enabled by AI methods means that even experts who design an AI assessment may not know exactly how it works and how it is arriving at its predictions and/or decisions. The risk here is that the assessment may seem to work well in pilot studies, but the experts do not know how it is making decisions. In that case, they cannot accurately predict or anticipate how well it will perform when deployed in new scenarios, such as making decisions on real life data in new circumstances that the algorithm had not seen before.





To help avoid risk and understand what makes a well-designed AI assessment, these guidelines walk decision makers through the process of evaluating an AI-based assessment. The guidelines involve a series of six broad questions and a set of associated metrics for the evaluation of an assessment. The guidelines also present the following key questions to ask to gain clarity should the required information not be readily available or provided by an assessment provider:



# Question 1.

## How relevant are the training data?

It is important that the right quality and quantity of data are used when developing and evaluating an AI assessment. Data should be accurate, relatively free of error, and relevant to the assessment that is being developed. The quantity of data should be adequate for the data analyses that are required for algorithm development and validation. The data should also be representative of the intended pool of applicants and all relevant groups (e.g., age, gender, disability status). Any data used for the development or use of an AI assessment should be stored in a way that provides the greatest possible privacy and protection.





## Question 2.

### How does the algorithm make decisions?

It is important that any user of an assessment that uses AI methods knows how the assessment and the underlying AI work. This concept is known as transparency.

Assessments that are considered transparent are those for which assessment developers and users can explain how the AI algorithm works and how it arrived at a specific decision (e.g., to hire versus not hire a candidate). Using a highly transparent assessment will reduce the risk of legal, brand, and efficacy issues.

## Question 3.

### How biased are the decisions?

In the context of AI, the term “bias” is used to refer to an AI algorithm that results in discrimination against a certain group of people (e.g., a particular race, age group, or gender), regardless of whether that discrimination is fair or unfair. To avoid the accidental incorporation of bias into an AI assessment, a thorough consideration of the removal of bias must be made at each step of the assessment development process.





## Question 4.

### How valid are the decisions?

Validity is the degree to which evidence and theory support the interpretations of test scores. Assessment validation is the process through which the validity of an assessment is established, and the thorough validation of an assessment is best practice in both talent assessment and AI. We describe the three most common methods of collecting validation evidence to support the use of an assessment (content-, construct-, and criterion-related evidence) and how to evaluate validation studies.

## Question 5.

### How final are the decisions?

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





# Question 6.

## How are candidates informed?

Candidates should be informed when AI will be used to score their responses to an assessment, and a sufficient explanation of how the AI works should be provided.



For AI to deliver on the promise it has to yield more accurate and engaging talent assessments, these assessments 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. The guiding principles presented in this document can be used to help address the rapidly evolving and complex landscape of AI in talent assessment.



# AI Human Resources Strategy



## Vision:

That the UAE Federal Government be one of the most smartly run governments at the regional and global levels in leveraging AI to streamline HR functions.

## Mission:

To embed AI in HR to create a data-driven pan-organization value.



## Objectives:



Creating intelligent products and services and designing intelligent HR processes.



Increasing staff productivity, boosting customer satisfaction, and achieving greater efficiency.



Data unification and automation in order to help in strategic decision-making process.



Creating synergies, cost optimization, and better resource allocation.

FAHR's vision to build a happy and innovative government workforce, and its mission to do so by using innovative and efficient solutions, can both be supported by using artificial intelligence in talent assessment.



# What is Talent Assessment?



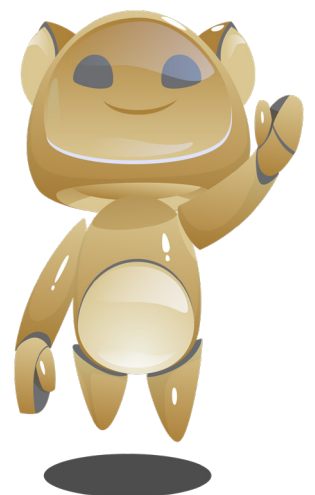
Talent assessment is a broad term for the process or activity of assessing key skills, traits, experience, and competencies of individuals (i.e., job applicants or employees) and using this information to make informed employment-related decisions. Talent assessment can occur either when an individual is applying to a job, known as pre-hire, or after an individual is hired and for development purposes, known as post-hire assessment. There are many different methods and forms of assessments, some requiring little technology, such as an in-person interview, others as sophisticated as a highly realistic virtual reality simulation of on-the-job performance.

There is ample academic research on the efficacy of talent assessment, dating back to the start of the last century. The consensus across this body of research is that talent assessments, when developed and used correctly, are an effective method for enhancing employment-related decisions regarding individual job candidates and/or current employees.

The science of psychometrics is heavily embedded in many forms of talent assessment. Psychometrics is the scientific practice of measuring psychological states, traits, and behavior.

Recent developments in talent assessment include a new focus on the experience of the candidates who take the assessments. This trend has been driven by the low unemployment rates in the decade since the 2008 recession and has coincided with growth in the field of user experience (UX) and the increased use of novel technology platforms (e.g., mobile devices) to deliver assessments.

Like many other industries, talent assessment is experiencing a period of growth and innovation through the application of AI technologies. AI is enabling enhanced scoring of traditional assessments, with the promise of making the assessments more accurate, and is also enabling the development of novel assessments that capture information from candidates that was not available previously, such as video and audio data.

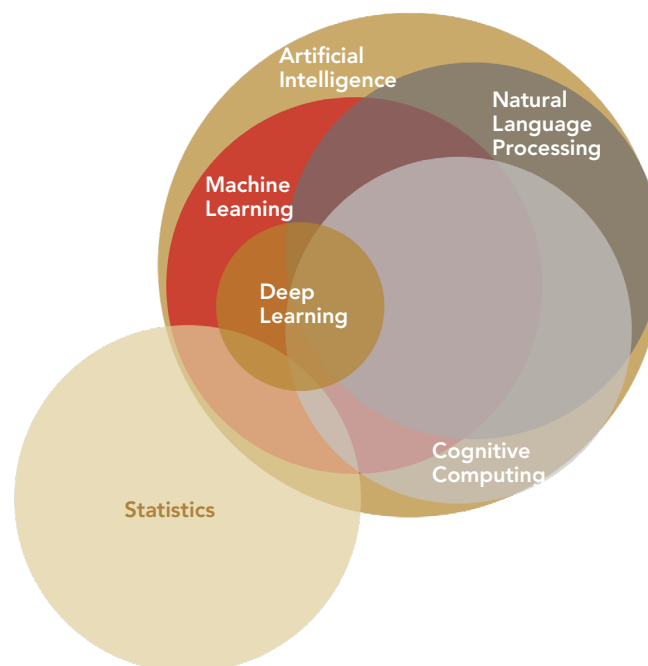




# What is Artificial Intelligence?



Artificial Intelligence (AI) is a broad term for which the definition has evolved over time, and the exact meaning tends to vary across fields. At its broadest, AI can be thought of as a field of study with the general purpose of developing digital programs and machines that can display some properties that are similar to human-level intelligence or judgment (Poole & Mackworth, 1998), and a field that encompasses many other sub-fields (see Figure 1) to achieve this general purpose.

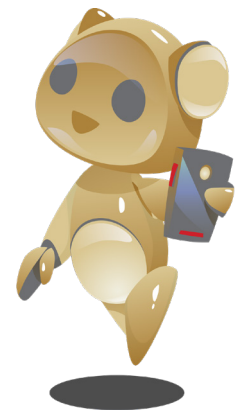


**Figure 1.** The many sub-fields within AI.





Early developments by AI researchers resulted in rule-based algorithms that were designed to follow clearly defined processes (e.g., playing a board game). These algorithms were hand coded to follow a long set of rules, which necessarily involved the input of domain experts. A key turning point in the history of AI was the development and use of machine learning. Machine learning (ML), a subfield of AI, involves the use of mathematical algorithms that are deployed and adapted to maximize the prediction of patterns of relationships within datasets. Machine learning, as its name implies, enables computer algorithms to learn from datasets without being specifically programmed. The importance of machine learning is that it enables a computer program to learn beyond human-level intelligence. For example, an AI application, or system, developed through machine learning will typically beat an AI application developed through hand coding, as the machine learning algorithm is able to iteratively learn from further exposure to data. It is the latter form of AI system development, which incorporates machine learning, that is the focus of this paper, and that is typically referred to when the term AI is used today.

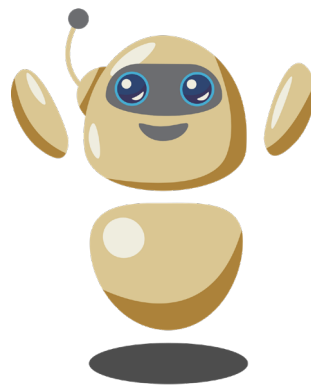


Central to the idea of machine learning is the concept of training, or teaching, the computer program. The program is initially taught on a training dataset. Then the program attempts to demonstrate its learning on a previously unseen (by the computer program) dataset, known as the test set. The computer program's performance on the test set is assessed by humans and deemed to be acceptable or unacceptable. If the performance is acceptable, the program may continue to further testing or may be deployed (depending on the specific situation). If the program does not perform adequately in the test dataset, then it will go back to receive further training. The use of machine learning, and in particular a subset of machine learning known as deep learning, has led to AI programs that can outperform human experts in narrowly defined tasks (e.g., Chess).





Another subfield of AI is Natural Language Processing (NLP), which involves the processing of spoken and written communication using human (i.e., “natural”) language. NLP enables the meaning conveyed by words to be transformed into data that can be used by machine learning algorithms. Many AI systems today use this combination of NLP and ML to enable the processing, understanding, and generation of speech by computer systems, such as Amazon’s Alexa and Apple’s Siri.



## How can AI benefit talent assessment?



The field of talent assessment has experienced multiple periods of innovation in the past century. Most notable are the move from paper-and-pencil testing to computer, then internet, then mobile-based assessments. Running in parallel to these technological developments have been developments in the mathematics behind psychometrics and AI, enabling new and more accurate models of measurement and prediction of key talent-related variables. Talent assessment is currently going through another pivotal moment of innovation with the incorporation of methods and technology from the field of AI. The application of these technologies and methods is leading to increases in the capabilities of psychometric assessments and the development of new methods that may be able to improve the accuracy of prediction beyond what was previously thought possible.

Some of the potential benefits that may be achieved through assessments that utilize AI technology and methods are shown in (Table 1.)





| Benefit           | Description   |
|-------------------|---|
| Better Prediction | AI methods, such as ML, can result in higher accuracy in the prediction of key talent outcomes (e.g., performance, turnover).   |
| Less Bias         | When used properly, AI methods can more easily identify and reduce the amount of bias in an assessment.   |
| New Methods       | The field of AI has a vast range of techniques, which are able to model relationships between audio and visual information.   |
| Spoken Response   | The combined use of NLP and ML methods enable candidates to speak their responses to an assessment, enabling a much more natural way for a candidate to engage with an assessment.            |
| More Realistic    | The ability to include spoken language and video information in an assessment opens the possibility of developing highly realistic simulations of work scenarios to be used in an assessment. |
| Better Experience | The use of ML methods can help to streamline the assessment process, enabling the candidate to have a better experience.  |

**Table 1.** Benefits of AI in Talent Assessment.



# What risks are involved in using AI to assess talent?



Talent assessment of any kind carries an associated risk. This is also the case for assessments that utilize AI methods. Therefore, the same general guidelines and regulations that inform the use of traditional assessments also apply to AI assessments (e.g., preventing bias, ensuring job-relatedness). However, there are some additional considerations regarding the use of assessments that utilize AI methods, which are discussed below.

## Legal Risk

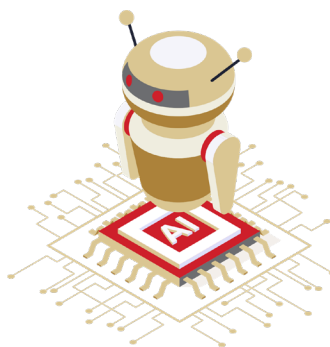
The legal risks associated with the use of assessments that utilize AI fall under two categories:

- A. Data protection.
- B. Bias.

The following sections describe these categories of legal risk.

## Data Protection

The first category of legal risk relates to data collection and use. Many countries have implemented legal regulations that outline the permitted use of applying AI methods to analyze data about individuals. Most of these regulations concern the use of AI within the UAE, European Union and the United States and are related to legal protection surrounding the collection and use of personal data. Examples of these regulations are presented in (Table 2.).



| Region         | Legislation/<br>Guidance                  | Description  |
|----------------|---|--|
| UAE            | The Personal Data Protection Law          | Federal Decree Law No. 45 of 2021 regarding the Protection of Personal Data gives the owner of the data the right to object to automated processing and decisions that have legal consequences or seriously affect the data subject, including profiling. Also, the data controller should include the human factor to review the decisions based on the data owner's request.   |
| European Union | General Data Protection Regulation (GDPR) | The GDPR requires processing of personal data to be fair, lawful, and transparent. Companies are required to disclose the use of AI to applicants and to provide sufficient information on how their data will be used for the applicant to make an informed decision to opt out or provide consent (Liem et al., 2018). The GDPR also includes the right not to be subjected to solely automated decision making, 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. |
|                | The ICO's AI Auditing Framework           | In Europe, there is no specific regulation regarding AI in force at present, but guidance has been released by data privacy regulators (in particular, the UK Information Commissioner's Office (ICO)) that contains specific recommendations.   |



| Region        | Legislation/<br>Guidance         | Description  |
|---------------|----------------------------------|--|
| United States | Illinois AI Video Interview Act  | The Illinois AI Video Interview Act requires employers to obtain the consent of applicants to use AI in the hiring process (Bologna, 2019). Additionally, it requires employers to explain the process and destroy data upon request.  |
|               | Fair Credit Reporting Act (FCRA) | The FCRA regulates the collection of consumer credit information and access to credit reports. It is also 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. |

**Table 2.** Regulations Concerning the Use of AI in UAE, EU and US

## Bias

The second category of legal risk is bias. Bias in an assessment occurs when the assessment unfairly discriminates against an individual based on one or more of the individual's characteristics or background (e.g., race, ethnicity, religion, country of origin, gender, age, disability status). One of the benefits of AI methods in talent assessment is that they can lead to a reduction in such bias by reducing the influence of subjective human judgment. Incorporating AI into an assessment could also increase and solidify bias, however, if not done correctly and with expert oversight.





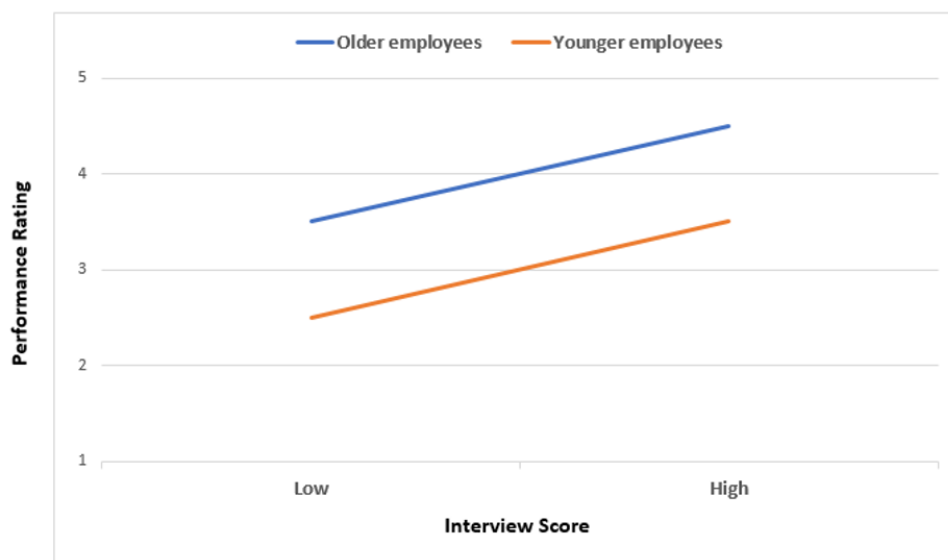
For example, a video interview assessment may be built to predict who will be a good performer from applicants' responses. The general process of linking responses to performance may be to have current incumbents complete the interview and collect performance ratings on those incumbents from their supervisors. The performance ratings are subjective judgments made by the supervisors, however, so their ratings may be influenced by factors that are not relevant to how interview responses made by candidates are related to later job performance. For example, supervisors may tend to give higher ratings to older employees for any number of reasons, such as (a) supervisors tend to be older and may give higher ratings to employees who are more like them, (b) older employees tend to have more experience, or (c) supervisors may be more familiar with the performance of a longer-tenured employee.

Figure 2 shows what the prediction lines for older and younger employees might look like in the situation above, with lower interview scores associated with lower performance ratings and higher interview scores associated with higher performance ratings at the same rate, but performance ratings are consistently higher for older employees. If these data are used to train an AI algorithm, the algorithm may give higher scores to responses containing words or phrases more often used by older people and lower scores to responses containing words or phrases less likely to be used by older people. This would result in a subtle form of bias in the algorithm that favors older people over younger people.

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, 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.



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 design an appropriate study for algorithm development, as well as inform the technological developers of the AI assessment on which features are likely to (a) be job-related, and (b) pose a risk of bias. 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.



**Figure 2.** Example of bias due to different predicted scores based on group membership.

## Public Relations Risk

The use of AI methods in talent assessment also carries risks to the public's perception of the organization if the AI methods are not designed and developed in a careful and considered way. For example, candidates who are assessed using an AI-based assessment that they determine to be unfair or intrusive may share these negative reactions with their network and on social media. Such negative perceptions and reviews may damage the reputation of the organization and could have a negative effect on the quantity and quality of applicants to the organization. The risk of public relations harm further underscores the importance of developing and delivering AI assessments with the guidance of experts in the field. Talent assessment experts can more easily recognize features of an assessment that could create negative reactions and recommend changes to improve candidate experiences.

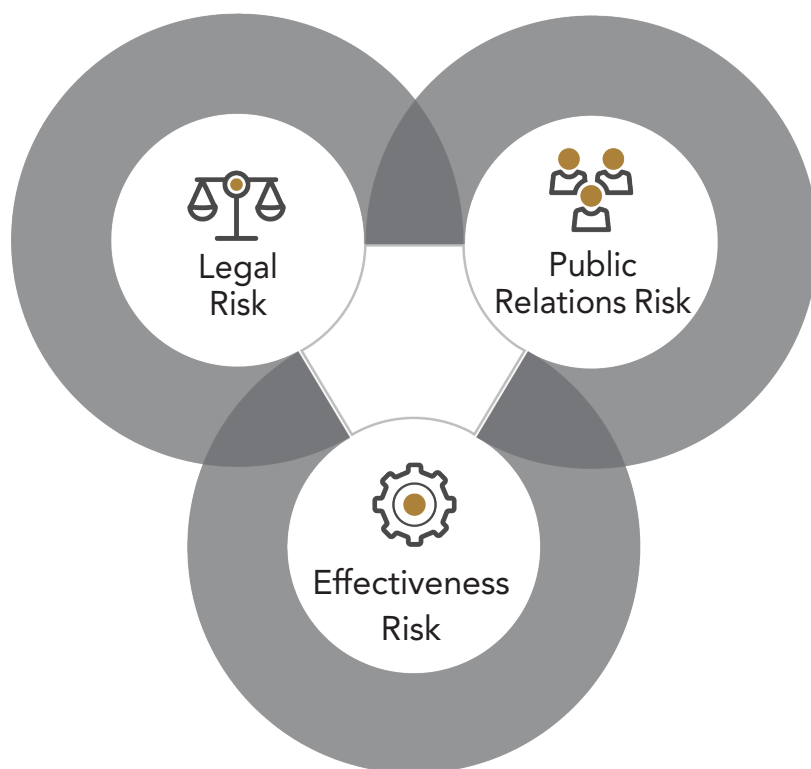
## Effectiveness Risk

The last of the three categories of risks is the risk of using an AI-based assessment that quite simply does not do what it claims to do – one that does not work. The complexity of data sources and modeling techniques enabled by AI methods means that even experts who design an AI assessment may not know exactly how it works and how it is arriving at its predictions and/or decisions. The risk here is that the assessment may seem to work well in pilot studies, but the experts do not know how it is making decisions. In that case, they cannot accurately predict or anticipate how well it will perform when deployed in new scenarios, such as making decisions on real life data in new circumstances that the algorithm had not seen before (e.g., new job roles, different types data input, different nationality of applicants, respondents with different types of accents).



Therefore, again, it is essential that experts in both AI and talent assessment are involved in the design, development, and delivery of an AI assessment. Talent assessment experts are well versed in different methods of validating the inferences made from an assessment. They can help ensure that the assessment measures what it is intended to measure, measures characteristics that are relevant to the job, and generalizes its ability to predict to new situations.

### Three Main risks in using AI for talent assessment:

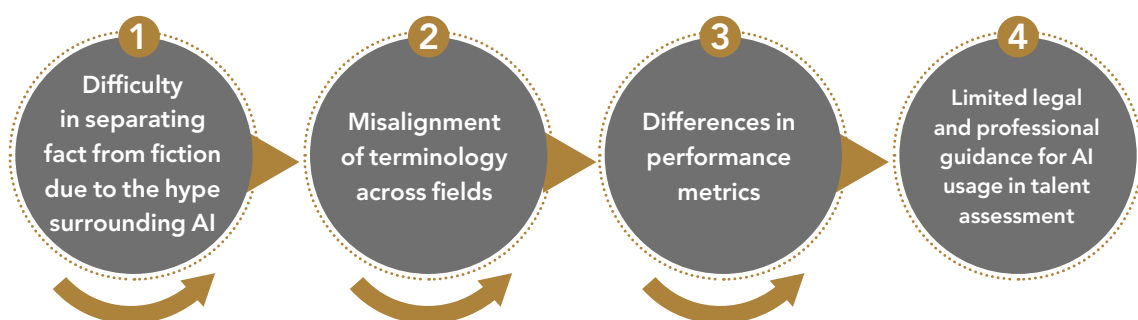


# Evaluating AI assessments

The previous sections have discussed the need for expertise in developing AI-based assessments. However, expertise alone is not sufficient for the development of industry-leading AI assessments. There also needs to be a structured and standardized way of reviewing an assessment, even those developed by experts, to determine their efficacy and risk of bias. This would not only allow for a thorough review of a single assessment, but it would also provide a standard against which to compare multiple AI assessments during a review process.

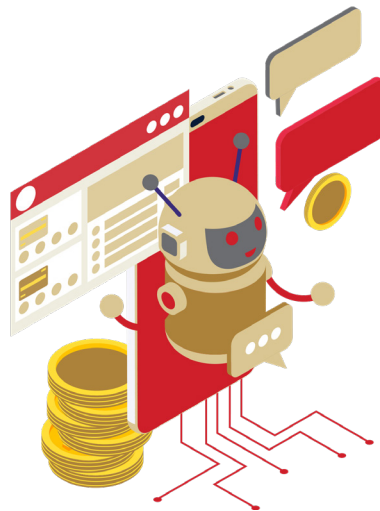
The challenge in doing this is multifaceted. First, there is plenty of hype in the technology and business industries surrounding the benefits and capabilities of AI technologies. It can be difficult to separate fact from fiction in this scenario. Second, AI and talent assessment are both technical fields, each with potentially confusing terminology, and often, terminology that does not align across fields (e.g., the term “bias” in AI means something different than the term “bias” in talent assessment). This can present challenges for non-technical users who are trying to understand how an assessment works. Third, different performance metrics are often used based on the method of assessment and how it was developed (e.g., R-squared, precision versus sensitivity, ROC), which leads to challenges when attempting to compare the predictive accuracy of two or more assessments. Fourth, guidance regarding best practices when using AI to assess talent is limited because this is a new development in the field. There is limited legal and professional guidance regarding how AI should and should not be used when assessing individuals.

## Four Challenges in Reviewing AI Assessments:



Because of the complexity in this situation, this set of guidelines was developed to help walk decision makers through the process of evaluating an AI-based assessment. The guidelines involve a series of six broad questions and a set of associated metrics for the evaluation of a given assessment. The guidelines also present key questions to ask to gain clarity should the required information not be readily available or provided by an assessment provider.

Best practices and recommendations for reviewing AI-based assessments have been developed (SHL, 2020). See (Table 3.) for a summary. The guidelines in this document have been developed to align with these best practices. (Table 4.) presents a mapping of the best practices on the AI assessment review framework questions we present in the rest of this document.



| Best Practice                  | Description   |
|--------------------------------|---|
| 1. Identify Data Requirements  | Consider data minimization, quality, diversity, and security.   |
| 2. Prioritize Transparency     | Develop transparent AI that demonstrates the interpretability of the results (i.e., no “black box” algorithms). The decisions and methodologies of AI systems are interpretable, to the extent permitted by available technology. |
| 3. Design for Fairness         | Build fairness into the assessment from the beginning.  |
| 4. Rigorously Validate         | Hold AI assessments to a high standard regarding validity evidence.   |
| 5. Incorporate Human Oversight | No AI assessment should make decisions without human oversight.   |
| 6. Disclose Intent             | Provide a notification, explanation, and request consent (where and when required) from candidates who will be assessed by AI.  |

**Table 3.** SHL’s Best Practices for the Use of AI to Assess Talent



| AI Assessment Best Practice    | FAHR AI Assessment Review Framework Questions      |
|--------------------------------|--|
| 1. Identify Data Requirements  | 1. How relevant are the training data?             |
| 2. Prioritize Transparency     | 2. How does the AI make decisions?                 |
| 3. Design for Fairness         | 3. How biased are the AI's decisions?              |
| 4. Rigorously Validate         | 4. How accurate, or valid, are the AI's decisions? |
| 5. Incorporate Human Oversight | 5. How final are the AI's decisions?               |
| 6. Disclose Intent             | 6. How much information do candidates receive?     |

**Table 4.** Mapping of FAHR's Assessment Review Framework to SHL's Best Practices





The following sections present a description of the guidelines, along with a hypothetical case study of how the guidelines and scoring metrics can be used.

## Hypothetical case study

Due to recent changes in the local economy and labor market, an organization has started receiving a three-fold increase in applications for its entry-level roles. This increase in applications is straining the organization's HR team, who are unable to review each application before making decisions on candidates that should enter the second phase of the selection process - a phone interview. The HR leadership decides to invest in technology that will enable all applications to be reviewed, with the goal being an increase in overall quality of hire.

A review of available technologies to solve this need reveals an AI-based assessment that uses information on a resume to predict job performance. The marketing materials from the organization that developed this assessment claim that the use of the assessment leads to a 20% increase in manager ratings of employee performance, while reducing recruiting costs.

You are responsible for identifying the assessment technology to achieve the goal set by the HR leadership. How would you go about determining if you should use this AI resume assessment?

Use the framework in the following sections to arrive at an answer.



## Question 1:

How relevant are the training data?



The first consideration regarding the development of the assessment is about the data that were used, or will be used, to develop it. It is important that the right quality and quantity of data are used, and this will vary for each assessment. Some domain expertise will be required when reviewing the data used to develop the assessment.

### Data quality

One of the most important considerations regarding the development of an assessment is the quality of the data. The data that an AI assessment are built upon can be considered the foundation for the performance of the assessment. This is because the assessment will be trained on, and therefore learn from, the data. If the data lack sufficient quality, then the quality (i.e., predictive accuracy) of the resulting assessment will be limited.

How can you distinguish between high-quality data and poor data? High-quality data are accurate, relatively free of error, and relevant to the assessment that is being developed. For example, if NLP is used to analyze spoken responses (e.g., to a video interview or role-play exercise), it is imperative that the translation of the audio to text accurately represents the original responses.

If the translation software produces a relatively high percentage of errors, the AI algorithm will be based on faulty information and respondents may not be evaluated fairly.



## Question 1:

How relevant are the training data?



The specifics of how to evaluate data quality will vary by each situation and type of data. For example, the quality of resume content is evaluated differently from the quality of spoken responses. Assessment developers should provide as much information as possible about the quality of the data that were used to develop the AI assessment. If this information is not available (for example, in a technical manual), then an inquiry into the quality of the data used to train the algorithm should be sent to the assessment developer.

### Data quantity

A second consideration regarding the data on which an AI assessment is developed is the quantity of data. Assessments that utilize AI often, but not always, require much larger amounts of data for development than traditional assessments. This is especially true when a new form of assessment is developed that does not have prior support in the scientific literature or a predecessor that has been widely used in practice in the talent assessment industry. An example of such an assessment is applying AI algorithms to extract facial expressions during a video interview and using this information to predict job performance. In this scenario, as there is limited prior research supporting the use of facial expressions as a basis for hiring decisions, an assessment developer would need to provide compelling evidence that this is a sound hiring practice. This compelling evidence could be achieved through a sufficiently large amount of data used to train the algorithm (though other factors mentioned in this guide would also need to be thoroughly considered in this scenario).

### How much data are required?

The exact number of data points required will vary according to multiple criteria (e.g., purpose of assessment, type of data, novelty of assessment). A clear rationale for the number of data points used to train the AI algorithm should be presented in a technical manual. If this information is not readily available, an inquiry into the rationale behind the quantity of the data used to train the algorithm should be sent to the assessment developer.



## Question 1:

How relevant are the training data?



### Data representativeness

The data used to develop an AI assessment should be representative of the intended pool of applicants and all relevant groups (e.g., age, gender, disability status). To achieve this, data from individuals from all relevant groups must be included in the training of the AI assessment. This may require strategic oversampling, in which additional data for a group of people that do not represent a large percentage of applicants – for example individuals who have a disability – are included to ensure that sufficient data points are included in the training process. Assessment developers should provide information regarding the approach used to ensure data representativeness during the development of the AI assessment.

### Data privacy and protection

It is important that any data used for the development or use of an AI assessment are stored in a way that provides the greatest possible privacy and protection. Many new assessments that use AI methods, such as AI-scored video interviews, capture and store video data that reveal the individual's identity and might contain sensitive information. Recent regulations, such as the European Union's GDPR, specify the requirements for protecting and storing such information. Assessment developers should provide adequate information regarding the methods used to secure and protect the assessment data, and any standards that they meet (e.g., GDPR).

### Applying Question 1 to the Hypothetical Case Study

This section explores how each of the training data considerations (data quality, quantity, representativeness, and privacy/protection) can be applied to the hypothetical case study - the review of an AI assessment that uses information from resumes to predict employee outcomes.



## Question 1:

How relevant are the training data?

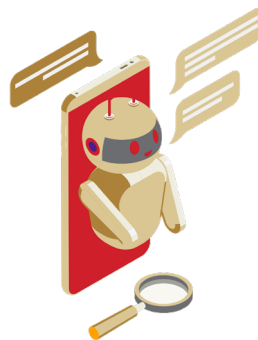


The first question to ask in this scenario is whether the assessment would be custom built or a generic, off-the-shelf product. Developing a custom-built version of the AI resume assessment would require that the customer provides the assessment vendor with a sufficient amount and quality of data that would be used to train the AI assessment. Following the development of the assessment, a validation study would need to be conducted by the vendor in collaboration with the customer. An off-the-shelf version of the assessment would have been previously validated with a different set of data that is not specific to the customer. There are tradeoffs between each of these methods, and the best decision for an organization will vary based on factors specific to their situation.

### AI Assessment Evaluation Sheet:

#### Training data relevance

(Table 5.) presents an evaluation sheet that can be used to evaluate the relevance of the training data collected for an AI assessment. For each data relevance metric, two scores are generated. The first score is either 1 or 0 based on whether a detailed description of the metric is provided in a technical manual or other document provided to the customer. The second score is based on a subjective evaluation of the extent to which the metric meets a standard, rated on a 0 to 3 scale. A subject matter expert (SME) with expertise in talent assessment should make this judgment. The final column contains example scores based on our hypothetical case study. The rationale behind these scores is presented in the following sections.



# Question 1:

How relevant are the training data?



| Metric   | Description   | Score |
|--|---|-------|
| Data quality   | A. A detailed description of the reasons for the selection and inclusion of all data types is available (Yes = 1, No = 0).  | A: 1  |
|  | B. Appropriateness of data quality for the use case (Not Appropriate = 0, Low Appropriateness = 1, Medium Appropriateness = 2, High Appropriateness = 3).   | B: 2  |
| Data quantity  | A. A detailed description of the reasons for the size of the sample used to train the AI assessment is available (Yes = 1, No = 0).   | A: 1  |
|  | B. Appropriateness of data quantity for the use case (Not Appropriate = 0, Low Appropriateness = 1, Medium Appropriateness = 2, High Appropriateness = 3).  | B: 3  |
| Data representativeness  | A. A detailed description of the representativeness of the sample used to train the AI assessment is available (Yes = 1, No = 0).   | A: 1  |
|  | B. Representativeness of the data for the use case (Not Representative = 0, Low Representativeness = 1, Medium Representativeness = 2, High Representativeness = 3).  | B: 1  |
| Data security  | A. A detailed description of how the data gathered by the assessment will be stored and protected is available (Yes = 1, No = 0).   | A: 1  |
|  | B. Appropriateness of the storage and protection of the data gathered for the use case (Not Secure/Protected = 0, Low Security/Protection = 1, Medium Security/Protection = 2, High Security/Protection = 3). | B: 3  |
| Overall training data relevance score (sum all values in the "Score" column) |   | 13/16 |

**Table 5.** Assessment Evaluation Sheet for Evaluating Training Data Relevance



## Question 1:

How relevant are the training data?



### Reviewing data quality

In the AI resume screen case study, the quality of the data used to develop the assessment could be investigated by inquiring into the accuracy of the transformation from resumes in their original form (e.g., .PDF, Word) into the version stored in a database (e.g., a text vector). For example, is each word on the original resume present in the word vector, and in the exact same order? Additionally, the quality of the outcome data should be scrutinized. If annual supervisor performance ratings are used, then this may present a challenge as these types of ratings were not designed for this type of predictive task. Ideally, a custom job performance measure will be developed that provides a more thorough assessment of the performance of the employee against the competencies required for that job. The reason for this is that annual performance reviews typically lack variance – for example, perhaps 70% of employees at an organization receive an annual performance score of “3,” 10% receive a “2,” 10% receive a “4,” 5% receive a “1,” and the other 5% receive a “5.” This lack of distinction involving most of the workforce will make it difficult for the AI algorithm to learn the relationships between elements on the resume and job performance. Therefore, while there may be high quality data on the side of the predictors (the resume data), the outcome measure may still be of low quality. Ideally, in this situation, a custom-built performance rating scale would be developed to collect outcome data against which the assessment can be validated, although this can be expensive to do at scale.

To present an example for evaluating data quality for the evaluation sheet shown in (Table 5.), we assume that the AI resume screen in the case study has a very high transformation and transcription accuracy rate of 99.7%. This means that the resume data used to develop the assessment is very close to its original form, and we therefore consider the resume data to be of high quality. We further assume that the assessment will be validated against annual employee performance review scores. As discussed above, this type of performance data has several limitations, so we consider this aspect of the data to be of low quality. Combining the high quality of the data on the predictor side with the low quality of the data on the criterion side, we arrive at a rating of “medium appropriateness” (2 out of 3) for the data quality score in (Table 5.).





## Question 1:

How relevant are the training data?



### Reviewing data quantity

Given that resume data are abundant within organizations, particularly for high-volume entry-level roles, there might be a very large amount of readily available resumes on which to train an algorithm. In fact, there could easily be tens if not hundreds of thousands of resumes that the organization (the customer) has collected over time. The number of cases used to train the algorithm and a rationale for why this number was chosen should be provided by the assessment developer.

For the case study, we assume that the assessment will be validated on 10,000 resumes and associated performance review scores based on three years of data. Therefore, the assessment scores high (3 out of 3) on data quantity in (Table 5.).

### Reviewing data representativeness

While there may be a high volume of resumes on which to train an algorithm, most resumes would not include information that would be required to assess the representativeness of the applicant pool (such as race or disability status). Because of this, there may not be sufficient labeled data regarding these types of demographic variables to include in the algorithm training process. Another important aspect to consider is whether the pool of candidates would change over time. For example, shifts in the labor market may lead to differences in education or experience among candidates over time. The result could be that the candidate pool on which the algorithm was developed is not representative of the candidate pool to which the algorithm is applied.

For the case study, we assume that the pool of candidates is expected to remain relatively stable over time. We further assume, however, that the data lack sufficient demographic information regarding group membership. Therefore, it will not be possible to conduct analyses to assess potential bias or fairness for certain demographic groups. For this reason, the assessment scores low (1 out of 3) on data representativeness in (Table 5.).





## Question 1:

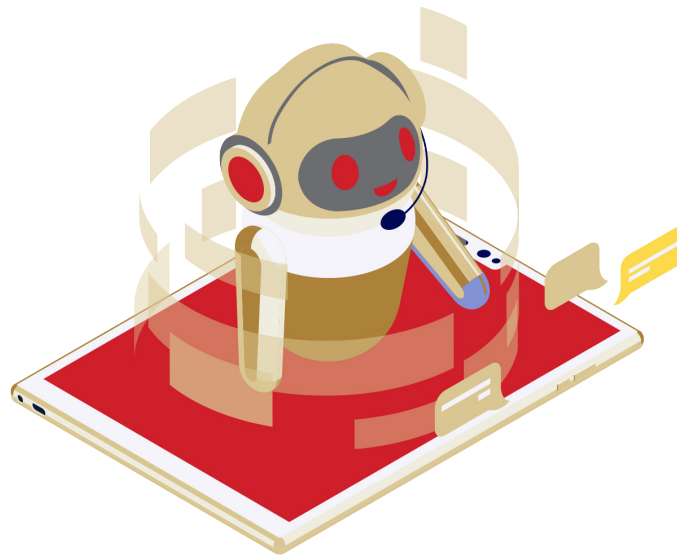
How relevant are the training data?



### Reviewing data privacy & protection

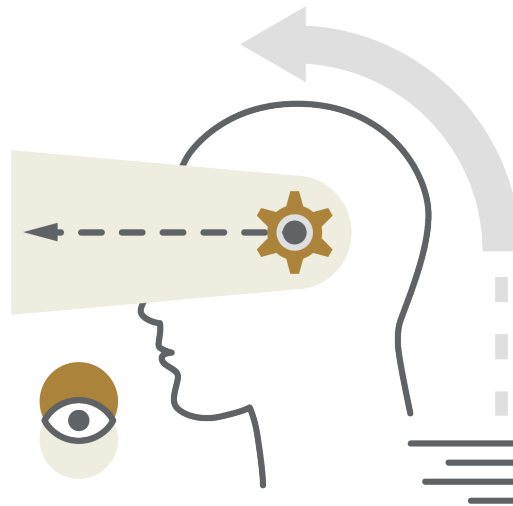
The assessment provider should have a statement regarding how they store and protect any data that are gathered as part of the assessment process. This statement should describe how strong the provider's data anonymization and security processes are and how they compare to the highest international standards (e.g., GDPR).

For the case study, we assume that the organization providing the assessment operates in accordance with the GDPR standards with respect to how they obtain and store personal data. Therefore, the assessment scores high (3 out of 3) on data privacy and protection in (Table 5.).



## Question 2:

How does  
the algorithm make decisions?



It is important that any user of an assessment that uses AI methods knows how the assessment and the underlying AI work. That is, how does the AI algorithm make decisions? This concept is known as the transparency of an assessment. Assessments that are considered transparent are those for which assessment developers and users are able to explain how the AI algorithm works and how it arrived at a specific decision (e.g., to hire versus not hire a candidate). Using a highly transparent assessment will reduce the risk of legal, brand, and efficacy issues, and enable the user to plan ahead regarding the continued suitability of the assessment for the particular roles for which it is used.

International standards support the use of transparent assessments. For example, the Society for Industrial and Organizational Psychology (SIOP) Principles for the Validation and Use of Personnel Selection Procedures (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”.

## Question 2:

How does  
the algorithm make decisions?



Two other examples are the European Union'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 include a right for individuals to receive an explanation for, and to contest, a solely automated decision by an algorithm. Therefore, the inner workings – the way an assessment makes decisions – should be known to the assessment developer and user, and should be sufficiently documented (e.g., in a technical manual).

An AI assessment can be made more transparent through many different approaches. One way is to only allow variables that have a conceptual linkage to the target job to be included in the model training process. Another is to only use simple and easy to interpret algorithms (such as logistic regression) in the assessment. A third option is to use relatively new and evolving methods, known as explainable AI (XAI), which aim to provide an explanation of even the most complex algorithms. In practice, a combination of the above options may be required to develop a sufficiently transparent assessment.

### Applying Question 2 to the Hypothetical Case Study

Transparency of an AI assessment is an important concept that reflects the ability of a user to understand how decisions are made (e.g., to hire versus not to hire an individual). The specifics of the algorithm that the assessment uses to make decisions are confidential and is therefore unlikely to be known by the assessment user. Nevertheless, the assessment developer should provide some degree of appropriate insight into the inner workings of the assessment. The degree of transparency required may vary by use case (e.g., selection decisions versus employee training), but the AI assessment developer should provide a thorough description of how the assessment makes decisions to potential users. This transparency should include information about what aspects of an individual the algorithm to make decisions uses, and how these variables are weighted and combined.



## Question 2:

How does  
the algorithm make decisions?



In our resume scoring hypothetical scenario, an investigation into the transparency of the assessment could begin with a review of any supporting materials of the assessment, such as a technical manual. The review should first focus on identifying what information from a candidate's resume is included in the AI algorithm. For example, perhaps the algorithm is looking only for key words or phrases that match the current job for which the assessment is being used. Or perhaps it is looking only for certain types of information, such as education. In addition, information should be provided regarding the process through which this information was identified as being important. Was the process completely data driven? Or was there domain and SME input into the types of variables that should be identified and used in building the algorithm? In addition, some information regarding the way that these variables are weighted and combined should be described. For example, perhaps an OLS regression model is used to weight and combine the resume variables.

For our resume assessment case study, we assume that domain experts in talent assessment were used to identify variables that were expected to have a relationship with job performance in the jobs under consideration. The experts developed a taxonomy of job titles that could be used to identify experience in prior jobs that were similar to the current job in question. For example, for a customer service role, prior experience in a call center was seen as having at least some of the competencies required for the new role. In addition, responsibilities listed underneath each job title on the resume can also be mapped to this taxonomy. The domain experts also chose to include time in each amount of job in the algorithm. The result is an estimate, for each candidate, of the amount of time they have spent in each of the job categories in the taxonomy. This information was then used to predict the subsequent job performance ratings of that individual.

We also assume that the weighting and combination of this information was developed through the testing of multiple different models. The best performing model was OLS regression, and metrics regarding the performance of the model are presented (i.e., model R-squared).



## Question 2:

How does the algorithm make decisions?



Given these assumptions, the assessment in this case study scores high on transparency. Information is provided regarding what information on the resume is used by the algorithm to create a predicted score and there is a description as to how the algorithm is weighting and combining this information (i.e., an OLS regression model). In addition, domain experts in talent assessment were involved in the identification of information on resumes that should be used by the assessment. This resulted in a taxonomy of jobs that further helps to make the assessment more transparent. Therefore, our evaluation sheet, presented in (Table 6.), yields a score of 3 for the elements of information used and weighting and combining of information.

### AI Assessment Evaluation Sheet: **Transparency**

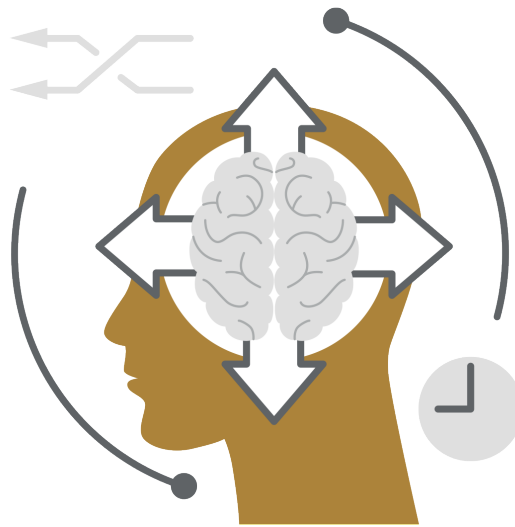
| Metric                                 | Description   | Score |
|--|---|-------|
| Information used                       | A. A description of the information used by the AI algorithm to arrive at a prediction is available (Yes = 1, No = 0).  | A: 1  |
|  | B. Information is appropriate for the use case (Not Appropriate = 0, Low Appropriateness = 1, Medium Appropriateness = 2, High Appropriateness = 3).                    | B: 3  |
| Weighting and combining of information | A. A description of the way that the information mentioned above is weighted and combined by the AI algorithm to arrive at a prediction is available (Yes = 1, No = 0). | A: 1  |
|  | B. Methodology is appropriate for the use case (Not Appropriate = 0, Low Appropriateness = 1, Medium Appropriateness = 2, High Appropriateness = 3).                    | B: 3  |
| Overall transparency score             |   | 8/8   |

**Table 6.** Assessment Evaluation Sheet for Evaluating Transparency



## Question 3:

How biased are the decisions?



It is important that AI assessments are developed in a way that does not result in biased decision-making. In the context of AI, the term “bias” is used to refer to an AI algorithm that results in discrimination against a certain group of people (e.g., a particular race, age group, or gender), regardless of whether that discrimination is fair or unfair. Such an algorithm is considered biased or having bias, and these biases may reflect larger societal biases (e.g., racism, ageism, sexism).



## Question 3:

### How biased are the decisions?



To avoid the accidental incorporation of bias into an AI assessment, a thorough consideration of the removal of bias must be made at each step of the assessment development process. For example, 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.

A documentation of this approach and the results should be provided in supporting documentation. For example, data should be presented that show mean scores on the assessment by relevant groups of people (e.g., race, age, gender) and any significant differences should be noted and investigated further.





## Question 3:

How biased are the decisions?



### Applying Question 3 to the Hypothetical Case Study

In the resume assessment example, an investigation of bias in the assessment should begin with the technical manual. The technical manual should have information regarding investigations that were conducted to test for bias resulting from the use of the assessment, along with key metrics and a supporting interpretation. Two forms of bias that should be inquired about are differences in assessment scores for different groups of the population (e.g., age, race, gender; the groups of focus may vary by region) and differences in predictive accuracy of the assessment for different groups. For example, it would be important to confirm that the scores from the assessment are not consistently higher or lower for men compared to women. If this was to occur, then an explanation for why it is happening, along with attempts to mitigate the gap in scores, should be described. The second type of bias to inquire about in the resume assessment example is whether the scores from the assessment are equally predictive of job performance for both men and women. For example, if the scores are twice as predictive of job performance for men than for women, despite men and women receiving equal scores (i.e., no difference in mean scores for men and women), the resume assessment would still have a form of bias.

For the resume assessment case study, we assume that a description of the mean and standard deviation of scores is provided for all relevant groups. These metrics reveal a small but significant difference between mean scores such that older candidates tend to receive higher scores than younger candidates. The technical manual explains that this is occurring because older candidates tend to have more job experience, and therefore are receiving higher scores as job experience is a key predictive variable used by the algorithm. As the difference in assessment scores is explainable, related to the variables identified by the SMEs as related to job performance and therefore relevant to the assessment, the risk of bias is documented but not considered prohibitive. Therefore, the assessment receives a moderate score for the first bias metric (2\3) in the evaluation sheet displayed in (Table 7.).

We also assume that the technical manual contains no reference to any investigations of differences in predictive accuracy across different groups. Therefore, the assessment receives the lowest score of 0 for the second bias metric in (Table 7.).





## Question 3:

How biased are the decisions?



### AI Assessment Evaluation Sheet: bias

| Metric   | Description   | Score |
|--|---|-------|
| Significant group differences in assessment scores   | A. A description of the mean and standard deviation of scores by different groups (e.g., race, sex, age) is available (Yes = 1, No = 0).                                    | A: 1  |
|  | B. Any score differences between groups are acceptable for the use case (Not Acceptable = 0, Low Acceptability = 1, Medium Acceptability = 2, High Acceptability = 3).      | B: 2  |
| Significant group differences in assessment accuracy | A. A description of the accuracy of assessment prediction by different groups (e.g., race, sex, age) is available (Yes = 1, No = 0).  | A: 0  |
|  | B. Any prediction differences between groups are acceptable for the use case (Not Acceptable = 0, Low Acceptability = 1, Medium Acceptability = 2, High Acceptability = 3). | B: 0  |
| Overall fairness (lack of bias) score                |   | 3/8   |

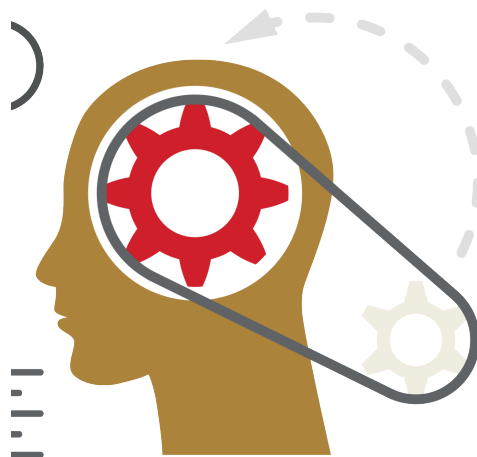
Table 7.

Assessment Evaluation Sheet for Evaluating Bias



## Question 4:

How valid are the decisions?



Validity is a technical term in talent assessment and refers to “the degree to which evidence and theory support the interpretations of test scores for proposed uses of tests” (AERA et al., 2014; p. 11). Assessment validation is the process through which the validity of an assessment is established, and the thorough validation of an assessment is best practice in both talent assessment (e.g., SIOP Principles, 2018) and AI (e.g., the EU’s Ethics Guidelines for Trustworthy AI).

Both quantitative evidence that an assessment adequately predicts its intended outcome and theory supporting the reason why it should predict is important in the validation process. This concept is described in the Standards for Psychological Testing, which states 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).

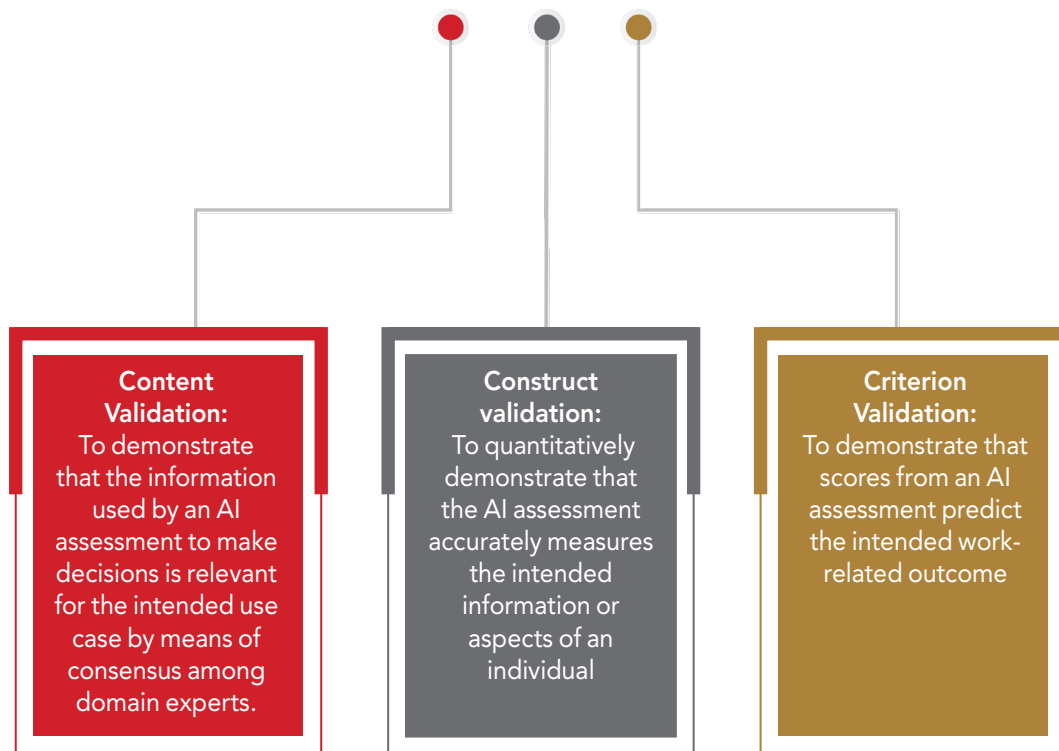
The most common methods of collecting validation evidence to support the use of an assessment, including assessments that use AI, are discussed below.

## Question 4:

How valid are the decisions?

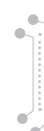


### Three Most common methods of assessment validation



### Content validation

Content validation focuses on demonstrating that the information used by an AI assessment to make decisions is relevant for the intended use case by means of consensus among domain experts. For example, if a group of domain experts agree that an AI simulation assessment adequately assesses the core competencies required of a data analyst, as demonstrated via their subjective ratings, then the assessment can be said to have content-related validity evidence supporting its use in selecting candidates for a data analyst job.



## Question 4:

How valid are the decisions?



### Construct validation

Construct validation focuses on quantitatively demonstrating that the AI assessment accurately measures the intended information or aspects of an individual – known as psychological constructs. For example, if an AI assessment is designed to measure an individual's communication skills, then scores from the assessment can be compared to scores from a different assessment also designed to measure communication skills. A high correlation between the two sets of scores produced from these two different assessments will demonstrate what is known as convergent validity – that the two conceptually related assessments (in that they are both designed to measure communication skills) are also quantitatively related in that their scores are highly correlated. This convergent validity is one type of evidence to support construct validity. Another form of construct validity evidence is known as discriminant validity. Discriminant validity is used to show that the assessment is not related to scores from an assessment that measures an unrelated construct. For example, if an AI assessment accurately measures communication skills, and only communication skills, then it should be unrelated, or very weakly related, to scores on a measure of conscientiousness. The combined demonstration of convergent and discriminant validity in this way provides evidence of construct validity.

### Criterion-related validation

Criterion-related validation focuses on demonstrating that scores from an AI assessment predict the intended work-related outcome (e.g., supervisor ratings of job performance). This prediction is expressed via a statistical metric (most often a correlation coefficient), and the value of the metric along with its statistical significance demonstrates the strength of the evidence supporting the criterion-related validity of the AI assessment. This is often considered the strongest type of evidence for demonstrating validity.

There are five important questions to consider when reviewing a criterion-related validation study for an AI assessment, see (Table 8.).



## Question 4:

How valid are the decisions?



| Study Element       | Question  |
|---------------------|---|
| 1. Job analysis     | Was a thorough job analysis conducted?  |
| 2. Criterion        | Was an appropriate criterion metric used?   |
| 3. Sample size      | Was the sample size large enough to provide sufficient statistical power and allow hold-out samples for cross validation?       |
| 4. SME input        | Was SME input included in the job analysis, selection or development of the criterion metric, and predictive feature selection? |
| 5. Cross validation | Was the performance of the AI assessment tested on one or more holdout samples?   |

**Table 8.** Elements Used to Evaluate a Criterion-Related Validation Study

The following descriptions of each of the elements in (Table 8.) provide guidance on how to design a strong criterion-related validation study for an AI assessment.

### 1. Job analysis

A job analysis provides assessment developers with key information about the performance domain of the job. This typically involves identifying the primary tasks, major work behaviors, competencies, and worker characteristics that are important to successfully performing the job. A job analysis helps to inform the choice of criterion and the information about a candidate that should be included in the AI assessment.



## Question 4:

How valid are the decisions?



When reviewing the information on an AI assessment, look for evidence that an appropriate job analysis was conducted. There are varying methods for conducting a job analysis, usually involving focus groups, surveys, or a combination. The exact method used may vary by type of assessment and type of job. What is important is that a detailed description of the methods behind the job analysis, and the results of the job analysis, are presented and available (e.g., in a technical manual).

### 2. 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 validity study is crucially important. High-quality criterion metrics are often those that have been informed by a job analysis, represent important aspects of the performance domain of the job, and were developed specifically to measure the performance domain of the target job.

When reviewing an AI assessment, it is very important to conduct an inquiry into the relevance and appropriateness of the criterion variable and the way it was measured. The large amount of data on which AI models are typically built often means that the criterion variable is readily available and not designed for this specific purpose (e.g., annual job performance ratings).

### 3. Sample size

The sample size is an important consideration for a criterion-related validity study. The sample size must be large enough to provide stable estimates of the weights associated with an AI algorithm and to test for their statistical significance. Power analyses are available to determine minimum sample size to ensure that statistical significance tests are appropriate (e.g., Cohen, 1992).

Informal guidelines frequently used to determine adequate sample



## Question 4:

How valid are the decisions?



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. For example, studies of statistical bias in prediction frequently require over 400 cases to have adequate power (Aguinis & Stone-Romero, 1997).

When reviewing the sample size used to develop an AI assessment, look for evidence of a rationale behind the decision to use a certain sample size. Did the assessment developers put careful thought and domain experience in AI algorithm development into this decision? Or was the choice for sample size based on what was convenient or easily available?

### 4. 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. The incorporation of domain experts during the design and development of an AI assessment often leads to better performing assessments.

When reviewing an AI assessment, look for evidence that SMEs in talent assessment, and not just AI, were used throughout the design and development of the assessment.

### 5. Cross validation

Cross validating an assessment involves testing the accuracy of the scores produced by the assessment in a new dataset. This is a very important step because the accuracy of an AI algorithm is typically much higher within the dataset on which it was developed, and often declines when the algorithm's performance is tested on a new dataset. Therefore, cross-validation analyses are an important way of making sure that scores from an assessment will remain valid when used operationally to assess candidates.



## Question 4:

### How valid are the decisions?



There are multiple ways to cross validate an AI algorithm, and the choice of cross-validation strategy may vary by assessment type, job type, and intended assessment use case, among other things. What is important to know is that a thorough and appropriate cross validation study was conducted, the outcome of which is reported and available for review.

In summary, the criterion-related validity approach is perhaps the most frequently used to validate AI assessments. Criterion-related validity studies that meet the majority of the above five criteria can be determined to have stronger evidence of validity regarding the use of the scores from the assessment to assist in making employment-related decisions.

### Applying Question 4 to the Hypothetical Case Study

When reviewing an AI assessment for evidence of validity, first look for the type of validity evidence on which the AI assessment is built.

For the resume assessment case study, we assume that a criterion-related validation approach was used. Consultation with SMEs in talent assessment confirms that this type of validation is highly appropriate for this assessment, but the addition of content- or construct-oriented validation evidence would have provided stronger rationale for the use of the assessment. Therefore, we give a rating of 2 for the first metric in the validity evaluation sheet, the appropriateness of the validation strategy (Table 9.).

The next step is to review the results of the validation study. Because the AI resume assessment used a criterion-related validation study, the factors in (Table 8.) should be reviewed to determine the acceptability of the validation results. We make the following assumptions for the AI resume assessment against the five factors in (Table 8.).

**Job analysis:** The technical manual for the AI resume assessment provides a detailed description of the job analysis that was conducted.

**Criterion:** The AI resume assessment used annual performance review scores for the criterion, and there are issues with using this type of metric as a criterion (as described previously in this document).





## Question 4:

How valid are the decisions?



**Sample size:** 50,000 resumes and associated first-year performance review ratings by supervisors, which is a very large sample size and more than adequate for all data analyses.

**SME input:** The technical manual provides detailed descriptions of the areas in which SME input was provided during the design and development of the assessment

**Cross validation:** The technical manual does not mention the use of unseen holdout samples as part of the validation process

The assessment is assigned a score of 2 after review of the factors in (Table 8.) This is because there are issues with the criterion used and a cross validation using an unseen holdout sample was not conducted.

### AI Assessment Evaluation Sheet: **Validity**

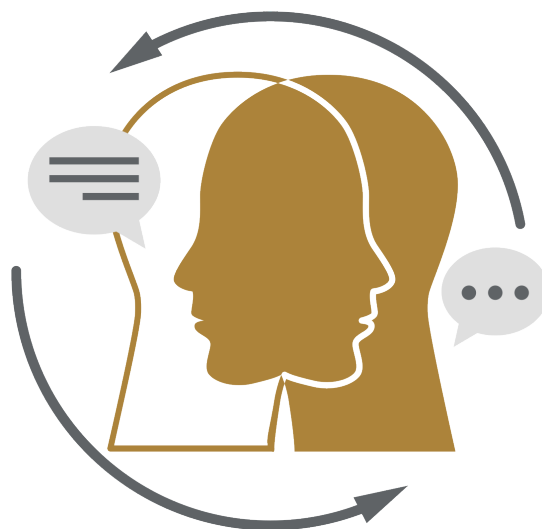
| Metric                         | Description   | Score |
|--------------------------------|---|-------|
| Assessment validation strategy | A. A description of the strategy used to validate the assessment is available (Yes = 1, No = 0).  | A: 1  |
|                                | B. The validation strategy is appropriate for the assessment and the use case (Not Appropriate = 0, Low Appropriateness = 1, Medium Appropriateness = 2, High Appropriateness = 3). | B: 2  |
| Assessment validation study    | A. A description of the results of the validation study is available (Yes = 1, No = 0).   | A: 1  |
|                                | B. The validation study is well designed and executed (Not Acceptable = 0, Low Acceptability = 1, Medium Acceptability = 2, High Acceptability = 3).                                | B: 2  |
| Overall validity score         |   | 6/8   |

Table 9. Assessment Evaluation Sheet for Evaluating Validity



## Question 5:

How final are the decisions?



There should be the ability for human oversight and intervention over any decisions made by an AI assessment. Industry-leading guidelines and regulations across the globe 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” (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 - the AI has complete autonomy. On the higher end of the spectrum, human oversight can be built into an AI assessment such that a human must first approve any decision the AI makes. In practice, the amount of human oversight required for most AI assessments will fall somewhere in the middle of this spectrum.

## Question 5:

### How final are the decisions?



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 that determine the candidates' outcomes from taking the assessment, and set the parameters under which the AI can act (e.g., with full autonomy, or with some human oversight).

The assessment user (e.g., a recruiter or hiring manager) provides oversight of an AI assessment by using the AI's recommendations as information that is combined with information from other sources in making 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 candidate. The reverse is also true. If an AI assessment determines that a candidate is not a good fit, but the recruiter thinks otherwise, the recruiter can override the AI assessment'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). The important point is that a human has the ability and opportunity to intervene when needed.

The key point is that 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.



## Question 5:

How final are the decisions?



### Applying Question 5 to the Hypothetical Case Study

When reviewing an AI assessment for evidence of human oversight, look for the extent to which decisions are driven by the algorithm output at each decision point in the hiring process.

We make the following assumptions for the resume screening example: The candidate completes a job application and submits a resume. The resume is automatically scored by the AI assessment and this score is attached to the application for the recruiter to review. The score provides a number on a scale of 1-100, along with a threshold -- Scores below 50 are considered a high risk of lower-than-average performance. This score is then used by recruiters as a decision aide, along with other information in the application, to help them determine whether to invite the candidate for a phone interview. Because the assessment user in this scenario can determine whether to follow the advice of the assessment report regarding progressing a candidate onto the phone screen, it receives a high score for human oversight (3) in the evaluation sheet shown in (Table 10.).

#### AI Assessment Evaluation Sheet: **oversight**

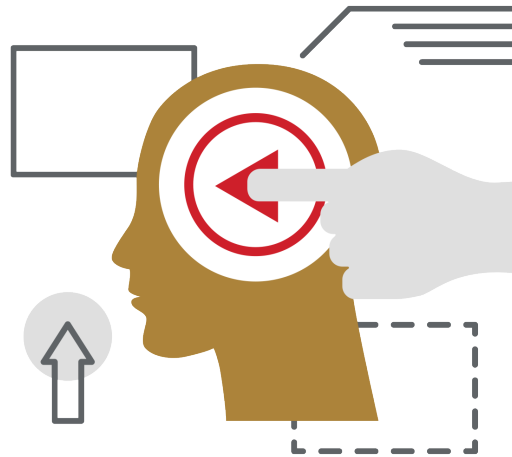
| Metric                             | Description  | Score |
|------------------------------------|--|-------|
| Opportunity for human intervention | A. A description of the level of human oversight incorporated into the assessment is available (Yes = 1, No = 0).  | A: 1  |
|                                    | B. The level of human oversight is appropriate for the assessment and the use case (Not Appropriate = 0, Low Appropriateness = 1, Medium Appropriateness = 2, High Appropriateness = 3). | B: 3  |
| Overall oversight score            |  | 4\4   |

**Table 10.** Assessment Evaluation Sheet for Evaluating Oversight



## Question 6:

How are candidates informed?



Industry-leading best practices suggest that candidates are informed when AI will be used to score their responses to an assessment, and that a sufficient explanation of how the AI works is provided.

A consideration as to whether informed consent from candidates should be required for the AI to score their responses should be made. In this situation, candidates that do not provide their consent should be provided with an alternative and equivalent method of assessment (e.g., a traditional in-person interview in place of an AI-scored virtual interview). When this occurs, a candidate's decision to not provide consent must not be detrimental to their chances of being selected.



## Question 6:

How are candidates informed?



### Applying Question 6 to the Hypothetical Case Study

When reviewing an AI assessment for how candidates are informed, look for a description or screen shot of what the candidate sees. Is there a statement that AI will be used? Is there a description of how the AI works? Is there a request for informed consent from the candidate? Is there a description of an alternative option for assessment if the candidate does not provide consent for their responses to be scored by AI?

For the resume screening example, we assume that a notification is provided on the page where the applicant uploads their resume. The notification states that an AI algorithm, which has been developed to help identify individuals who are a good fit for the role based on the information in their resume, will be used to produce a job fit score from their resume content. It explains that this score will be visible to the HR team at the company and will be used, along with other information from their application, to determine whether the individual will be invited to participate in the next step of the selection process.

We further assume, however, that this is a situation in which informed consent should be required for the use of the AI assessment. The candidates are not presented any information regarding how the AI works, are not required to provide consent to be scored by AI, and there is no mention of an alternative assessment that could be used to score candidates who would prefer it. Based on these assumptions, the evaluation includes a high score for informing the candidate of the use of AI but low scores for the other metrics on the evaluation sheet displayed in (Table 11.).



## Question 6:

How are candidates informed?



### AI Assessment Evaluation Sheet: **informing the candidate**

| Metric  | Description  | Score                   |
|---|--|-------------------------|
| Candidates are informed of use of AI          | <p>A. A description of the degree to which candidates are informed of the use of AI is available (Yes = 1, No = 0).</p> <p>B. The disclosure of the use of AI is appropriate for the assessment and the use case (Not Appropriate = 0, Low Appropriateness = 1, Medium Appropriateness = 2, High Appropriateness = 3).</p>   | <p>A: 1</p> <p>B: 3</p> |
| Candidates are informed of how AI works       | <p>A. A description of the degree to which candidates are informed of how the AI works is available (Yes = 1, No = 0).</p> <p>B. The disclosure of how the AI works is appropriate for the assessment and the use case (Not Appropriate = 0, Low Appropriateness = 1, Medium Appropriateness = 2, High Appropriateness = 3).</p>   | <p>A: 0</p> <p>B: 0</p> |
| Informed consent is requested from candidates | <p>A. A description of whether, and why/why not, informed consent is required from candidates is available (Yes = 1, No = 0).</p> <p>B. The inclusion or exclusion of informed consent is appropriate for the assessment and the use case (Not Appropriate = 0, Low Appropriateness = 1, Medium Appropriateness = 2, High Appropriateness = 3).</p>                      | <p>A: 0</p> <p>B: 0</p> |
| Alternative assessment is available           | <p>A. A description of whether, and why/why not, an alternate form of assessment can be used for those who decline to be scored by AI is available (Yes = 1, No = 0).</p> <p>B. The inclusion or exclusion of an alternative form of assessment is appropriate (Not Appropriate = 0, Low Appropriateness = 1, Medium Appropriateness = 2, High Appropriateness = 3).</p> | <p>A: 0</p> <p>B: 0</p> |
| Overall informed candidate score              |  | 4\16                    |

**Table 11.**

Assessment Evaluation Sheet for Evaluating Informing the Candidate





# Scorecard for the use of AI by government entities to assess talent



In this section, we provide blank evaluation sheets for each of the questions that can be used to judge the appropriateness of AI assessments.

| Question 1: How relevant are the training data?   |   |           |
|---|---|-----------|
| Metric  | Description   | Score     |
| Data quality  | A. A detailed description of the reasons for the selection and inclusion of all data types is available (Yes = 1, No = 0).  | A: __ / 1 |
|   | B. Appropriateness of data quality for the use case (Not Appropriate = 0, Low Appropriateness = 1, Medium Appropriateness = 2, High Appropriateness = 3).   | B: __ / 3 |
| Data quantity   | A. A detailed description of the reasons for the size of the sample used to train the AI assessment is available (Yes = 1, No = 0).   | A: __ / 1 |
|   | B. Appropriateness of data quantity for the use case (Not Appropriate = 0, Low Appropriateness = 1, Medium Appropriateness = 2, High Appropriateness = 3).  | B: __ / 3 |
| Data representativeness   | A. A detailed description of the representativeness of the sample used to train the AI assessment is available (Yes = 1, No = 0).   | A: __ / 1 |
|   | B. Representativeness of the data for the use case (Not Representative = 0, Low Representativeness = 1, Medium Representativeness = 2, High Representativeness = 3).  | B: __ / 3 |
| Data security   | A. A detailed description of how the data gathered by the assessment will be stored and protected is available (Yes = 1, No = 0).   | A: __ / 1 |
|   | B. Appropriateness of the storage and protection of the data gathered for the use case (Not Secure/Protected = 0, Low Security/Protection = 1, Medium Security/Protection = 2, High Security/Protection = 3). | B: __ / 3 |
| Overall training data relevance score (sum of all assigned scores in the "Score" column / sum of all possible scores) |   | __/16     |







## Question 2: How does the algorithm make decisions?

| Metric                                 | Description  | Score                               |
|--|--|-------------------------------------|
| Information used                       | <p>A. A description of the information used by the AI algorithm to arrive at a prediction is available (Yes = 1, No = 0).</p> <p>B. Information is appropriate for the use case (Not Appropriate = 0, Low Appropriateness = 1, Medium Appropriateness = 2, High Appropriateness = 3).</p>  | <p>A: ___ / 1</p> <p>B: ___ / 3</p> |
| Weighting and combining of information | <p>A. A description of the way that the information mentioned above is weighted and combined by the AI algorithm to arrive at a prediction is available (Yes = 1, No = 0).</p> <p>B. Methodology is appropriate for the use case (Not Appropriate = 0, Low Appropriateness = 1, Medium Appropriateness = 2, High Appropriateness = 3).</p> | <p>A: ___ / 1</p> <p>B: ___ / 3</p> |
| Overall transparency score             |  | ___/8                               |





### Question 3: How biased are the decisions?

| Metric   | Description  | Score                             |
|--|--|-----------------------------------|
| Significant group differences in assessment scores   | <p>A. A description of the mean and standard deviation of scores by different groups (e.g., race, sex, age) is available (Yes = 1, No = 0).</p> <p>B. Any score differences between groups are acceptable for the use case (Not Acceptable = 0, Low Acceptability = 1, Medium Acceptability = 2, High Acceptability = 3).</p>  | <p>A: __ / 1</p> <p>B: __ / 3</p> |
| Significant group differences in assessment accuracy | <p>A. A description of the accuracy of assessment prediction by different groups (e.g., race, sex, age) is available (Yes = 1, No = 0).</p> <p>B. Any prediction differences between groups are acceptable for the use case (Not Acceptable = 0, Low Acceptability = 1, Medium Acceptability = 2, High Acceptability = 3).</p> | <p>A: __ / 1</p> <p>B: __ / 3</p> |
| Overall fairness (lack of bias) score                |  | __/8                              |





#### Question 4: How valid are the decisions?

| Metric                         | Description   | Score     |
|--------------------------------|---|-----------|
| Assessment validation strategy | A. A description of the strategy used to validate the assessment is available (Yes = 1, No = 0).  | A: __ / 1 |
|                                | B. The validation strategy is appropriate for the assessment and the use case (Not Appropriate = 0, Low Appropriateness = 1, Medium Appropriateness = 2, High Appropriateness = 3). | B: __ / 3 |
| Assessment validation study    | A. A description of the results of the validation study is available (Yes = 1, No = 0).   | A: __ / 1 |
|                                | B. The validation study is well designed and executed (Not Acceptable = 0, Low Acceptability = 1, Medium Acceptability = 2, High Acceptability = 3).                                | B: __ / 3 |
| Overall validity score         |   | __/8      |

#### Question 5: How final are the decisions?

| Metric                             | Description  | Score     |
|------------------------------------|--|-----------|
| Opportunity for human intervention | A. A description of the level of human oversight incorporated into the assessment is available (Yes = 1, No = 0).  | A: __ / 1 |
|                                    | B. The level of human oversight is appropriate for the assessment and the use case (Not Appropriate = 0, Low Appropriateness = 1, Medium Appropriateness = 2, High Appropriateness = 3). | B: __ / 3 |
| Overall oversight score            |  | __/4      |





### Question 6: How are candidates informed?

| Metric  | Description  | Score                             |
|---|--|-----------------------------------|
| Candidates are informed of use of AI          | <p>A. A description of the degree to which candidates are informed of the use of AI is available (Yes = 1, No = 0).</p> <p>B. The disclosure of the use of AI is appropriate for the assessment and the use case (Not Appropriate = 0, Low Appropriateness = 1, Medium Appropriateness = 2, High Appropriateness = 3).</p>   | <p>A: __ / 1</p> <p>B: __ / 3</p> |
| Candidates are informed of how AI works       | <p>A. A description of the degree to which candidates are informed of how the AI works is available (Yes = 1, No = 0).</p> <p>B. The disclosure of how the AI works is appropriate for the assessment and the use case (Not Appropriate = 0, Low Appropriateness = 1, Medium Appropriateness = 2, High Appropriateness = 3).</p>   | <p>A: __ / 1</p> <p>B: __ / 3</p> |
| Informed consent is requested from candidates | <p>A. A description of whether, and why/why not, informed consent is required from candidates is available (Yes = 1, No = 0).</p> <p>B. The inclusion or exclusion of informed consent is appropriate for the assessment and the use case (Not Appropriate = 0, Low Appropriateness = 1, Medium Appropriateness = 2, High Appropriateness = 3).</p>                      | <p>A: __ / 1</p> <p>B: __ / 3</p> |
| Alternative assessment is available           | <p>A. A description of whether, and why/why not, an alternate form of assessment can be used for those who decline to be scored by AI is available (Yes = 1, No = 0).</p> <p>B. The inclusion or exclusion of an alternative form of assessment is appropriate (Not Appropriate = 0, Low Appropriateness = 1, Medium Appropriateness = 2, High Appropriateness = 3).</p> | <p>A: __ / 1</p> <p>B: __ / 3</p> |
| Overall informed candidate score              |  | __ / 16                           |



# A look ahead – the future of AI in HR



The arrival of AI-based assessments promises to revolutionize talent assessment, but 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, it is possible 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 may 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 to which HR has access. 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.

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 not only be able to speak their responses in a natural way, 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.

For AI assessments to deliver on these promises, however, 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. The guiding principles presented in this document can be used to help address the rapidly evolving and complex landscape of AI in talent assessment.



# References



**Aguinis, H., & Stone-Romero, E. F. (1997).** Methodological artifacts in moderated multiple regression and their effects on statistical power. *Journal of Applied Psychology*, 82, 192206-.

**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.

**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>.

**Cohen, J. (1992).** A power primer. *Psychological Bulletin*, 112, 155–159.

**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. 197253-). Cham, Switzerland.

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

**SHL (2020).** The ethical and effective use of artificial intelligence to assess talent. Arlington, VA: Author.

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

**The Public Voice (2018).** Universal Guidelines for Artificial Intelligence. <https://thepublicvoice.org/ai-universal-guidelines/>



# Glossary of terms



**AI Assessment:** In the context of talent assessment, refers to assessments that 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., knowledge, skills, 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.

**Black box:** Any AI system for which the underlying computational processes or algorithms are unknown.

**Criterion:** In the context of talent assessment, the outcome variable against which an assessment is validated (when using a criterion-related validation study).

**Deep learning:** A sophisticated form of machine learning, sometimes referred to as an “artificial neural network,” that is inspired by the structure and functioning of biological neurons.

**Fairness:** In the context of talent assessment, a broad term that encompasses equal treatment of all candidates, equal access to the constructs being measured by an assessment, and non-discriminatory hiring practices or outcomes of hiring practices.

**Features:** A term used in computer science to represent independent variables.

**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.

**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.

**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.

**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.





