Considerations and Recommendations for the Validation and Use of AI-Based Assessments for Employee Selection

Society for Industrial and Organizational Psychology (SIOP)

January 2023

Considerations and Recommendations for the Validation and Use of AI–Based Assessments for Employee Selection

Ad Hoc Task Force on AI-Based Assessments

Christopher D. Nye, PHD (Chair)
Michigan State University

Leaetta Hough, PHD
The Dunnette Group, LTD.

Kisha Jones, PHD
Florida International University

Richard N. Landers, PHD
University of Minnesota

Toni S. Locklear, PHD
APTMetrics

William Macey, PHD
CultureFactors, Inc.

Frederick L. Oswald, PHD
Rice University

Dan J. Putka, PHD
Human Resources Research Organization

Ann Marie Ryan, PHD
Michigan State University

Ryne A. Sherman, PHD
Hogan Assessment Systems

Nancy T. Tippins, PHD
Nancy T. Tippins Group, LLC

ABOUT THE AUTHORS

The Society for Industrial and Organizational Psychology (SIOP) is the premier professional association for the science and practice of industrial and organizational (I-O) psychology. While an independent organization with its own governance, SIOP also represents Division 14 of the American Psychological Association and is an organizational affiliate of the Association for Psychological Science.

I-O psychology is a dynamic and growing field that addresses workplace issues at the individual, team, and organizational levels. I-O psychologists apply research that improves the well-being and performance of people and the organizations that employ them. This involves everything from workforce planning, employee selection, and leader development to studying job attitudes and job motivation, implementing work teams, improving diversity and inclusion, and facilitating organizational change. Particularly relevant for the current document, I-O Psychologists are also rigorously trained in the development and evaluation of tests, assessments, and other selection procedures that are used to make hiring and promotion decisions. Through their expertise, I-O Psychologists have been working to improve the accuracy and fairness of hiring procedures for decades.

SIOP's Task Force on Artificial Intelligence (AI)-Based Assessments was launched in October 2021 to address issues related to the development and implementation of AI for assessing and hiring talent and to increase awareness of scientific research on this topic. The task force comprises SIOP members with expertise in a broad range of related areas such as employee selection and assessment, psychometrics and measurement bias, and the use of AI-based technologies in the workplace.

CONTENTS

Executive Summary

Artificial Intelligence (AI) is changing the way that organizations assess and hire talent. These changes are happening rapidly yet with little guidance about how to effectively validate, implement, interpret, and use scores produced from AI-based assessments in this context. Therefore, the purpose of the current document is to provide scientifically based recommendations for the effective use of AI for assessing and hiring talent.

A key theme that emerges in this document is that AI-based assessments used to make hiring and promotion decisions require the same level of scrutiny and should meet the same standards that traditional employment tests have been subjected to for decades. However, the way that these standards are evaluated and met may be unique to AI-based assessments. Therefore, this document discusses the unique challenges and considerations that arise in the development, evaluation, use, and interpretation of AI-based assessments. To be clear, the aim here is to provide recommendations rather than mandates for the use of AI-based assessments. These recommendations are discussed in detail in the following five sections:

Section 1. AI-Based Assessments Should Produce Scores that Predict Future Job Performance or Other Relevant Outcomes Accurately

Section 2. AI-Based Assessments Should Produce Consistent Scores that Reflect Job-Related Characteristics (e.g., upon re-assessment)

Section 3. AI-Based Assessments Should Produce Scores that are Considered Fair and Unbiased

Section 4. Operational Considerations and Appropriate Use of AI-Based Assessments for Hiring

Section 5. All Steps and Decisions Relating to the Development and Scoring of AI-Based Assessments Should be Documented for Verification and Auditing.

Introduction

Artificial Intelligence (AI) is changing many aspects of our lives, including the way that employees perform their jobs and interact with their work environments. Similarly, advances in AI are also beginning to change the way that organizations assess and hire talent. These changes are happening rapidly and with little guidance about how to effectively validate, implement, interpret, and use scores produced from AI-based assessments to make hiring decisions. At the same time, there have been increasing calls for legal regulations, ethical guidelines, and other forms of scrutiny of AI-based selection procedures, reflecting concerns over privacy, fairness, lack of transparency, and the accuracy of their predictions. Therefore, the purpose of the current document is to provide scientifically based recommendations for the effective use of AI for assessing and hiring talent.

AI refers to a broad range of technologies and statistical techniques that have the potential to identify patterns of behavior and predict outcomes, such as job performance and retention, that are important to organizations. This broad technology is being applied in numerous ways, but the focus of the current document is on AI-based assessments that are used to assess job applicants or make hiring decisions. This includes technology-based hiring and promotion procedures that incorporate AI, machine learning, and other novel assessment techniques (e.g., game-based assessments, automated scoring of video interviews, and evaluation of social media). Although this document is intended to cover many of these applications, it does not elaborate on specific AI applications and issues that can arise in each of these areas. In addition, this document does not cover other important workplace applications of AI, such as employee-AI interactions, employee-AI teams, AI-based decision support systems, or incorporating AI into products, among other applications. Instead, this document is intended to cover general principles that can guide the development, evaluation, and implementation of AI for assessing and hiring talent.

This document builds on previous standards and best practices outlined in the *Principles for the Validation and Use of Personnel Selection Procedures* (SIOP, 2018; hereafter referred to as the *Principles*), which is published by the Society for Industrial and Organizational Psychology (SIOP) and updated periodically to reflect current scientific research and best practices in hiring and promotion. Although the *Principles* are applicable to *all personnel selection procedures*, the development and use of AI-based assessments presents unique challenges. Therefore, the focus here is on discussing the unique challenges and considerations that arise when using AI-based assessments and hiring procedures. Consequently, the content discussed in the *Principles* is not repeated here, except when doing so provides the necessary context for understanding particular recommendations for AI-based assessments.

Throughout this document, a key theme that emerges is that AI-based assessments used to make hiring and promotion decisions require the same level of scrutiny and should meet the same standards that traditional employment tests have been subjected to for decades. In other words, **AI-based assessments should still be required to meet traditional standards for hiring and assessment procedures, even if the way that those standards are evaluated and met varies slightly**. The aim here is to provide recommendations rather than mandates for the use of AI. Not all of the content discussed here will apply to all situations and, in some cases,

gathering the recommended information can take time. In these situations, the professional judgment of experts on hiring and assessment can provide essential additional guidance on the applicability and necessity of specific recommendations provided here. Nevertheless, adhering to the recommendations herein will strengthen the business-case for using a particular AI-based assessment and provide the evidence necessary to support its use for assessing and hiring talent.

This document is organized into five sections that cover the key considerations for the evaluation, use, and interpretation of AI-based assessments. The first three sections describe specific characteristics that AI-based assessments should have. These characteristics can be used as criteria for evaluating AI-based assessments and represent the minimal requirements necessary to justify the use of AI for hiring and promotion decisions. The fourth section discusses operational concerns with the administration and use of AI-based assessments. This section outlines specific factors that need to be considered when evaluating and implementing these assessments. The final section outlines the importance of documenting the development and scoring of AI-based hiring and assessment procedures for verification and auditing.

Section 1. AI-based Assessments Should Produce Scores that Predict Future Job Performance or Other Relevant Outcomes Accurately

In order to justify their use for assessing and hiring talent, scores produced by AI-based assessments should demonstrate relationships with future job performance or other job-relevant outcomes. In the assessment literature, this is known as establishing evidence for the validity of the use and interpretation of assessment scores. Validity evidence can be gathered using a variety of approaches. Nevertheless, not all sources of validity evidence that are relevant to traditional assessments will be applicable to AI-based assessments. Therefore, this section discusses the potential sources of evidence and their applicability to scores produced from AI-based assessments.

1.1 Sources of Validity Evidence

The *Principles* emphasize that the validity of hiring procedures will depend on the intended interpretation and use of their scores. Stated differently, the most appropriate sources of validity evidence will be determined by how well they inform the intended uses and interpretations of scores generated by the hiring procedures. There is no exception for scores generated from AI algorithms. Nevertheless, although the sources of validity evidence may be similar, their implementation is likely to vary for AI-based assessments.

1.1.1 Validity Evidence Based on Empirical Relationships between Scores on Predictors and Other Variables

One source of validity evidence described in the *Principles* is evidence based on the relationships between assessment scores and other variables. Here, different strategies can be used. First, *convergent validity* estimates examine the relationships between one or more assessments intended to measure the same characteristics. For example, scores from an AI-based assessment designed to measure personality traits should be highly correlated with other measures of personality assessing the same traits (e.g., job applicants' conscientiousness). This

form of validity evidence can help to empirically establish that the AI-based assessment is in fact measuring the characteristic it was intended to measure. A second related validation strategy involves establishing the *discriminant validity* of an assessment. This form of validity evidence is established by showing that an AI-based assessment is *less* related to measures of distinct characteristics that the assessment was not intended to measure. For example, an assessment of personality traits should be less highly correlated with a measure of cognitive ability (i.e., showing discriminant validity) than with another measure of the same personality traits (i.e., showing convergent validity).

Note that both convergent and discriminant validity evidence are examined whenever an AI-based assessment is intended to measure a specific attribute or characteristic. The focus of these sources of validity evidence is on demonstrating that the assessment is measuring the characteristic it is intended to measure (i.e., scores can be interpreted as indicators of an individual's standing on the measured characteristic). When AI-based assessments are not designed to measure a specific characteristic, an alternative validation approach is to focus on identifying a set of predictors (e.g., data extracted from a video interview or gamified assessment) that are used to predict relevant work-related behaviors or outcomes. This validation approach is known as *criterion-related validity* and is useful for demonstrating the job relevance of the AI-based assessment.

Regardless of which validation strategy is used to demonstrate relationships with other variables, the quality of the validity evidence that is obtained also depends on the validity of the outcome measure that is used to train the AI algorithm. This outcome measure is sometimes referred to as "ground truth" in the AI and machine learning literatures and the way that developers define "truth" will affect the algorithm. Therefore, the measures used to train the algorithm will need to be carefully selected. Similarly, the measures used to establish convergent or discriminant validity can also affect the validation results and should be carefully selected. Nevertheless, it is possible that AI-based assessments can reliably capture complex data that assess aspects of a characteristic (e.g., a personality trait) that are not currently captured with alternative measures, and this can artificially degrade estimates of convergent validity. However, a lack of convergent validity evidence does not automatically imply that this is the case and other sources of validity evidence would be required to establish that the assessment is measuring the characteristics it was intended to measure.

1.1.2 Content-Related Evidence

Another validation approach described by the *Principles* is *content validity*, reflecting evidence based on the content of the assessment. This type of validity evidence is obtained by asking subject matter experts to evaluate the match between the content of an assessment and the tasks performed on the job, the worker requirements, or the work products that are generated on the job (e.g., written reports, computer code). With traditional assessments, the content that is evaluated generally consists of the items or tasks in the assessment. In contrast, AI-based assessments may include a much broader range of content such as resumes, social media posts, transcripts from interviews, or past performance narratives, among other things, that can either supplement or replace the items or tasks that are common in traditional assessments. This broader range of content may make assessing content validity using traditional approaches more

difficult. In particular, it may not be feasible for subject matter experts to evaluate hundreds or thousands of data sources that are used as the content in an AI-based assessment. For example, although video interviews may ask applicants to respond to several job-related questions, the scores produced by an AI algorithm may be based on a broad range of content features that are not clearly related to the job such as word usage, voice pitch, facial expressions, and posture (Hickman et al., 2022). In this example, each of the features used by the AI algorithm need to be evaluated for content validity. Again, this may not be feasible if large numbers of predictors are used.

To facilitate evaluations of content-related validity, the data sources that are used in an AI-based assessment should be clearly stated and documented for evaluation. This documentation should describe the methods used to collect, sample, or extract the assessment content and the steps taken to ensure that these methods remain consistent during the training and implementation of the AI algorithm.

1.1.3 Evidence Based on Internal Structure

Another source of validity evidence described by the *Principles* is evidence based on the internal structure of an assessment. This form of evidence is typically gathered by examining the relationships between the items in an assessment and the extent to which they represent the characteristics that are being measured. Traditionally, this evidence comes from conducting various forms of factor analysis. However, because many AI-based assessments do not include multiple items designed to measure the same characteristic, evidence based on the internal structure of an assessment may be more challenging to collect. This is the case even when the assessment is designed to measure a specific characteristic (e.g., a personality trait) because AI-based assessments often do not include multiple predictors that are necessarily highly correlated with each other, which is an inherent focus of examining internal structure. Instead, AI-based assessments often include a broad range of predictors that only provide trace information about the job candidates' characteristics. In short, evidence of internal structure may not be possible to collect for all AI-based assessments. When possible, this form of validity evidence should be collected. When it is not possible, the other validation approaches described here may be more useful. Nevertheless, regardless of whether evidence of internal structure is collected for the AI-based assessments, it should still be collected for all scale scores that are used as predictors in the assessment as well as the alternative measures used to establish convergent or discriminant validity and the outcome measures used to train the AI algorithm.

1.1.4 Evidence Based on the Response Process

A final source of validity evidence advocated by the *Principles* is evidence based on the response process. This form of validity evidence is primarily examined when the assessment is intended to measure the process used by individuals to respond to test questions. The example used in the *Principles* is when a work sample measure is intended to assess whether applicants use the proper technique for performing the work. This form of validation evidence may not apply to most AI-based assessments, because they typically focus on predicting specific outcomes rather than measuring response processes. Nevertheless, as with other forms of validity evidence, evidence based on the response process can be collected for specific measures that are

used as input to an AI algorithm, even though it is not useful for the overall scores on the assessment.

1.2 Collecting Validity Evidence

Providing support for each type of validity evidence described above requires designing and conducting studies that involve data collection, analysis, and interpretation. In general, the corresponding sections of the *Principles* (see pp. 6-8) still apply to validity studies conducted with AI-based assessments. However, the unique characteristics of AI-based assessments also create different challenges that need to be considered when designing and carrying out validation research. Therefore, readers are referred to the *Principles* for a more comprehensive discussion of the key characteristics of a validation study, but many of the unique aspects of these studies as applied to AI-based assessments are discussed below.

1.2.1 Job Relevance

As described above, one way to show that the scores on an AI-based assessment are job relevant is to demonstrate that they can predict job relevant outcomes (i.e., a criterion-related validity study). However, even with this form of validity evidence, it is still possible that the AI-based scores are deficient (i.e., they do not measure all job-relevant aspects of a characteristic) and/or contaminated (i.e., they are measuring aspects of a characteristic that are unrelated to the job). For example, incorporating organic data (e.g., social media posts, likes, pictures) into an AI-based assessment may unintentionally introduce many sources of job-irrelevant variance if these data sources are not pure indicators of the intended characteristics (e.g., Liou et al., 2022; Park et al., 2015). Overall, both deficiency and contamination can distort or degrade the overall validity of assessment scores. Thus, demonstrating the job-relevance of an AI-based assessment should be a key consideration while developing the assessment and designing the validation study.

In some cases, the outcome used to train an AI model is a particular characteristic that is traditionally used as a predictor in the hiring process (e.g., personality scores). By itself, this does not establish that the AI assessment scores are job-relevant. Instead, establishing job relevance would require additional validity evidence to demonstrate that the scores produced by the AI-based assessment are also related to a separate job-relevant outcome (e.g., job performance, turnover).

Regardless of which outcomes are used to train the AI model, the selection and measurement of the outcome is particularly important. Selection of the outcome should be based on a careful evaluation of the job requirements (i.e., a "job analysis"[1]) and the priorities of the organization (e.g., if the organization really wants to reduce turnover, turnover may be an appropriate criterion). Similarly, when organizations choose to combine scores from several criteria into a single composite to be predicted, there should be both a job-relevant and a statistical rationale to support the rules and weights used to combine these criteria. High-quality measures of the outcome(s) of interest are also required to ensure the accuracy and validity of the AI-based assessment. This includes measures that are reliable, representative of the construct of

[1] See p. 7 of the *Principles* for a more detailed discussion of a job analysis.

interest, uncontaminated by job-irrelevant variance, and unbiased against subgroups of individuals. If high-quality measures of the desired job-relevant outcomes are not available, it may not be feasible to conduct an appropriate criterion-related validation study, thereby limiting the available evidence supporting the AI-based assessment.

1.2.2 Validation Samples

As with other forms of assessment, the sample used to validate an AI-based assessment must be representative of the applicant pool (see *Principles*, p. 10). For AI-based assessments, obtaining a representative sample is not only important for the validation study, but also for the sample used to train the AI algorithm if the two samples are different. Any biases in the training data can propagate historical biases. Therefore, the training data should, for example, include a diverse sample of individuals that is carefully selected to be representative of the intended population for which the algorithm will be used. If the current workforce lacks diversity or has substantially restricted variance on the characteristics being assessed, then special care should be taken when using existing employees to train a machine learning algorithm.

The source of the training data can also influence whether the sample is representative of the broader candidate pool, as some individuals may be less likely to have data available from certain sources (e.g., social media). For example, some personality traits are associated with greater technology acceptance and social media use (Blackwell et al., 2017; Svendsen et al., 2013). It is also important to consider how subgroup differences in the availability of data can affect validity. If some groups (e.g., those protected under Title VII) generate different types of data or use the platforms that provide the data differently, then subgroup differences on the AI-based assessment might be influenced by factors other than the characteristics being assessed. These factors could also influence the observed pattern of relationships with other variables (e.g., patterns of convergent or criterion-related validities) and suggest a lack of validity for assessment scores.

Finally, statistical power may also be a concern for algorithms developed on small samples. Statistical power refers to the ability to detect a relevant predictor or combination of predictors incorporated in the algorithm during the training phase. When sample sizes are small (i.e., there are few people in the dataset used to train the algorithm), the overall validity of the model may be compromised, resulting in distorted, noisy, or otherwise inaccurate conclusions about the validity of an AI-based assessment. Even though machine learning algorithms contain built-in features to help ensure robust out-of-sample-prediction (e.g., regularization, cross-validation), when samples are too small, they can undermine the use of the model. This issue has to be considered carefully, such as when the majority group is large and minority subgroups are too small to inform the model well (such that even oversampling may not be helpful). These types of issues will continue to be a concern as AI algorithms are applied in new employment contexts.

1.3 Generalizing Validity Evidence

With regard to traditional selection procedures, the *Principles* note that existing accumulated validity evidence for an assessment can sometimes be sufficient to support the

implementation of the assessment in a new setting without conducting a local validation study. This may be appropriate when the characteristics being assessed have demonstrated generalizable validity across contexts and situations and when a compelling argument can be made that this validity evidence is also applicable to the new setting. In contrast, AI algorithms, and the big data on which they operate, run the risk of being less generalizable across contexts and situations. An algorithm that predicts an outcome in one setting may not be applicable to other settings if the AI-based assessment is not measuring generalizable characteristics or is using the same data across settings. Using an AI-based assessment tool with an algorithm developed and validated within another organizational context (e.g., purchasing an AI-based tool from a vendor) requires evidence that using the assessment in the current context results in valid inferences. In this situation, examining the cross-validity of an AI-based assessment and algorithm does not necessarily ensure that the predictive ability of its features apply in different settings or contexts.

In some situations, a validity generalization or transportability argument is presented to establish the validity of an AI-based tool. In such cases, the original validation study must be technically sound, and the new situation must be comparable to the original study in terms of job content, job context, job requirements, and applicant group. AI algorithms may be less generalizable across settings if the algorithm is not measuring readily identifiable characteristics (e.g., knowledge, skills, abilities), the algorithm is less explainable (e.g., a neural net), or the algorithm is learning and changing regularly.

1.4 Section Summary

In sum, many of the principles related to validation studies of traditional selection procedures are also applicable to AI-based assessments. However, validation studies of AI-assessments also face many unique challenges that should be considered when designing, conducting, and interpreting them. The ultimate results of a validation effort depend on numerous factors including, but not limited to, the selection and measurement of the outcome used to train the AI model (i.e., "ground truth"), the job relevance of both the predictors used in the AI model and the outcome it is designed to predict, the quality of any other measures used for the validation study (e.g., for convergent or discriminant validity), the quality of the training data, and the representativeness of both the training data and the validation sample. Despite these challenges, and in some ways because of them, establishing the validity of the AI-based assessment is critical for justifying its use and interpretation within an organizational context.

Section 2. AI-Based Assessments Should Produce Consistent Scores that Reflect Job-Related Characteristics (e.g., upon re-assessment).

In addition to demonstrating the validity of AI-based assessments, demonstrating that the scores produced from these assessments are consistent across replications of the assessment can also help to justify their use for assessing and hiring talent. Consistency can be assessed in a number of ways, including by aggregating across relevant items or samples of applicant behavior, across different algorithmic predictions within an ensemble, across different raters/observers, or across repeated assessments. In the assessment literature, this consistency is known as the *reliability* of an assessment. Demonstrating reliability is important for ensuring that

applicants' scores are an adequate representation of their standing on the characteristics that are being assessed, rather than a result of random error. Conversely, if an assessment provides inconsistent scores, then it will be unclear which (if any) applicant scores are accurate, resulting in weaker convergent and discriminant validity, weaker relationships with job-relevant outcomes (e.g., job performance, turnover), and less accurate decisions based on these scores. Therefore, estimating the reliability of AI-based assessment scores is necessary both to engender confidence in the scores that are used for decision-making and to obtain accurate estimates of their validity.

Factors critical to interpreting the reliability of an assessment include the accuracy of the responses that are provided and the conditions that might affect the accuracy of the scores that are generated, such as the context in which the data are generated (e.g., incumbents might not respond to AI-based assessments in the same way as highly motivated job applicants) or the consistency of the applicant data over time. Additional critical factors to understand include identifying the relevant conditions under which the predictor scores might vary (e.g., some measured characteristics might vary across situations or over time) and adopting study designs that allow for appropriate statistical analyses (e.g., reliability, validity, subgroup differences). If it is not possible to gather data about these critical factors during the development of the assessment or other data collection efforts, then results regarding the reliability of predictor scores should be qualified accordingly.

Choosing the most appropriate estimates of reliability depends on the intended interpretation of the AI-based assessment scores (e.g., whether scores are intended to reflect a particular characteristic or show consistency over time) and how scores from the assessment will be used (e.g., for rank ordering applicants, or for making pass-fail or hire-no hire decisions) (Schmidt & Hunter, 1996; Putka & Sackett, 2010). Whether or not one is able to evaluate the most appropriate reliability estimate for their situation also depends on whether appropriate data are available or can be reasonably collected given the predictors used in the AI assessment (e.g., whether one can partition text data into multiple samples for evaluating consistency, whether one has multiple samples of text data across time for respondents). Any reliability estimates that are reported should be accompanied by a clear description of the study design, the data collection effort, and the score interpretations that are supported by the reliability estimates (e.g., whether scores can be interpreted consistently across administrations or over time).

Section 3. AI-Based Assessments Should Produce Scores that are Considered Fair and Unbiased

Demonstrating that AI-based assessments produce scores that are considered fair and unbiased can also support their use for assessing and hiring talent. Because there are many definitions of bias and fairness, one must communicate the specific definitions used, the analyses that operationalize these definitions, and the study design used to evaluate these features of an assessment.

3.1 Fairness

The *Principles* defines fairness as a "*social rather than a psychometric concept*" (p. 22; italics added) meaning that this assessment characteristic is conceptual rather than statistical in

nature. Thus, establishing a shared understanding of fairness, if not a single definition, is a necessary first step for discussing the "fairness" of any specific assessment, whether that assessment involves AI or not (Landers & Behrend, 2022). Importantly, fairness as described in the *Principles*, is considered from the perspective of adherence to professional test development standards, not in relation to individual perceptions of fairness. Keeping this in mind, the *Principles* describe four major meanings of fairness to consider.

First, the *Principles* explicitly reject a popular meaning of fairness requiring equality of group outcomes, such as hiring rates or performance evaluations. Although this meaning is common, it reflects an incomplete view that does not consider the potential for differences between groups to be fully or partially related to job-related characteristics. For example, if Group A on average has less of a job-related skill than Group B, this meaning of fairness would suggest that a hiring system would only be fair if Group A and Group B were still hired at equal rates, regardless of their differences in job-related skills. From the perspective outlined in the *Principles*, such a system is less fair than an alternative system in which individuals are hired based on their job-related skills.

The *Principles* provides a second meaning of fairness as "equitable treatment of all examinees during the selection process…in terms of testing conditions, access to practice materials, performance feedback, retest opportunities, and other features of test administration, including providing reasonable accommodation for test takers with disabilities" (p. 22). With this definition, many common implementations of AI-based assessment are likely to raise fairness concerns. For example, algorithms that are being constantly updated with new data suggest that incoming job applicants may not be competing with previous applicants on the same standards. Applicant retesting, where each score may be contingent on a different algorithm and dataset, makes this problem even more complex. For AI-based assessments that rely upon historical data, such as digital trace records, spatiotemporal workplace data, and social media data, the highly unstable nature of those data sources may also create fairness concerns under this definition of fairness. For example, differences between specific technical implementations (e.g., Facebook vs. Instagram as a data source) suggest that different quality and formats of data will be available for different people. Further, changes in those platforms over time make comparisons of data from people using those platforms in the past with people actively using them challenging. These potential concerns suggest that the analysis of these data sources may be inherently unfair when considered in the context of this definition of fairness. Outside of trace data analysis, problems related to inequality of access are likely more common. For example, an assessment that incorporates video recordings may be less accessible to those with low-speed internet access or low-quality recording equipment. Similarly, an assessment involving AI-based scoring of text responses to written prompts under time pressure may be less accessible to those with reduced fine motor control due to physical disability. Importantly, these differences, as well as other differences that may arise in AI-based assessments, may pertain to legally protected classes of applicants (e.g., women, racial-ethnic minorities, individuals with disabilities).

The third meaning of fairness given in the *Principles* requires "*comparable access to the constructs measured by the selection procedure*" (p. 22). Although this is closely related to the second meaning of fairness, it is distinct in its focus on the rights of job applicants to have equal opportunity to express their competitiveness for the job. For example, in hiring based upon social

media traces, not everyone has the time or interest in maintaining active accounts on social media or necessarily engages with those platforms in the same way, let alone in a way that is intended to be job-relevant. Examinees will, by definition, have different numbers or types of data points or may have none at all. Further, consider a wide range of job-irrelevant factors on which applicants differ: e.g., age, race, ethnicity, gender, socio-economic status, and cultural background. These are all characteristics that could differentially accentuate, attenuate, distort, or restrict access to both job-relevant experiences (e.g., going to coding camp, displaying leadership-related behaviors) and job-irrelevant experiences that machine learning algorithms might pick up on (e.g., states visited, enjoyment of curly fries, sports played). Even outside of the context of social media and other trace data, AI-based modeling may vary in accuracy according to group membership. For example, some computer vision algorithms have been found to assess facial expressions less accurately for darker skin colors, which would systematically prevent or harm the assessment quality of emotional expression for minority subgroups.

The *Principles* also provides a fourth meaning of fairness as "*a lack of bias*" (p. 22). Unfortunately, as with the term fairness, there is no consensus either within or between disciplines as to the definition of bias. As with fairness, reaching any shared understanding of bias requires that all parties agree on the definition of bias they are referencing (e.g., predictive bias is a mathematical form of bias discussed below). Despite this challenge, lack of bias remains one of the most central concerns and certainly one of the most discussed problems in the development of fair AI. Yet an evaluator of the fairness of an AI-based assessment, using one of the alternative meanings above, may or may not consider bias to be relevant to their fairness determination. Thus, it is critical to define both fairness and bias before beginning to investigate and make any claims about a specific assessment on these terms. Therefore, we discuss the definitions of bias used in the *Principles* in more detail below.

3.2 Bias

Although fairness is considered a social concept, bias is a measurement concept that can be evaluated using widely researched psychometric and statistical methods. The *Principles* discusses two primary forms of bias that can be evaluated: predictive bias and measurement bias. These forms of bias are closely related but typically evaluated separately using distinct techniques. Both forms of bias should be considered when evaluating an AI-based assessment.

3.2.1 Predictive Bias

The *Principles* defines *predictive bias* as "when, for a given subgroup, systematic non zero errors of prediction are made for members of the subgroup" (p. 23). By this definition, subgroups (race/ethnicity, gender, etc.) can still show observed mean differences on employment tests and/or organizational outcomes to be predicted, so long as *errors* in prediction are unrelated to the subgroup to which a person belongs. When a linear regression model is used to make predictions, then a lack of predictive bias is best supported by the *Cleary model* (Cleary, 1968), which was proposed in the late 1960s to determine the presence or absence of predictive bias in educational data across various races/ethnicities. With this model, the conclusion that an assessment shows predictive bias is reached when a common regression line estimated in the full sample does not apply equally well to all relevant subgroups of individuals and shows systematic

and meaningful levels of overprediction or underprediction (i.e., different intercepts and/or different slopes). For example, predictive bias occurs when a common regression line estimated in the full sample underpredicts performance for a particular group relative to the group-specific regression line. In contrast, overprediction reflects predictive bias favoring a specific subgroup, such as when use of a common regression line overpredicts performance for that group relative to a group-specific regression line.

The definition of predictive bias provided in the *Principles* still applies to many AI-based assessments, meaning that predictions from these assessments should not differ meaningfully across subgroups of individuals. That said, AI-based assessments are often much more complex than traditional linear regression. For example, AI-based assessments can incorporate hundreds (or thousands) of predictor variables. In this case, a regression-based predictive bias analysis can still be conducted by treating the AI-based assessment score as the sole predictor of interest in the Cleary approach to determine if slope and intercept differences exist across groups (Sackett, Laczo, & Lippe, 2003). This approach sheds light on *overall* bias in prediction but not necessarily on the set of underlying predictor variables within an AI assessment.

Investigations of predictive bias, as with investigations of validity discussed earlier, should involve samples that are as representative of the population of interest as possible (e.g., with similar demographic characteristics, job types, tenure on the job, and location of work). Researchers have focused on how range restriction impacts evaluations of predictive bias with traditional assessments, but the applicability of this work to AI-based assessments is as yet unknown. Additionally, the *Principles* states that "[i]n domains where relevant research exists, generalized evidence can be appropriate for examining predictive bias" (p. 24). As discussed in the section on validity (see Section 1), this statement may not apply to all AI-based assessments, particularly when they are not designed to measure specific generalizable constructs. Technical work on the generalizability of AI-based assessment scores has only recently been published (e.g., Landers et al., 2022; Speer, 2021; Yuan et al., 2021). Scores from AI-based assessments will only generalize to other jobs, organizations, or contexts when they amplify job-relevant information and minimize sources of *deficiency* (i.e., when the predictors are not measuring all job-relevant aspects of a characteristic) and *contamination* (i.e., when the predictors are measuring aspects of a characteristic that are completely unrelated to the job).

3.2.2 Measurement Bias

As noted in the *Principles*, "measurement bias refers to sources of irrelevant variance that result in systematically higher or lower scores for members of particular groups" (p. 24). Note that this is different from predictive bias because there is no criterion being predicted. Instead, measurement bias helps to ensure that the assessment is measuring the characteristic(s) it was designed to measure in the same way across all relevant subgroups. As defined here, measurement bias can also influence AI-based assessments if irrelevant sources of variance affect scores for some groups of individuals more than others (e.g., see Sections 1.2.1 and 3.2.2 for additional discussions of what is considered relevant for an AI-based assessment). Therefore, it is important to examine how measurement bias can be detected and understood in AI-based assessments, and how these techniques compare to the traditional methods of identifying measurement bias as discussed in the *Principles*.

Traditional methods of identifying measurement bias, as found in the *Principles*, are typically focused on examining the content or the response process for individual *items* in an assessment. For example, item sensitivity reviews (Golubovich, Grand, Ryan, & Schmitt, 2014) have traditionally been used and involve asking diverse groups of individuals to review each item in an assessment and determine whether the meaning or interpretation of the item might differ across subgroups. Measurement bias can also be evaluated psychometrically using differential item functioning (DIF) analysis, which statistically examines whether each item is measuring the underlying attribute in the same way for different subgroups of individuals. Ideally, individuals with the same level of the measured attribute will respond to each item in the same way, regardless of the subgroup they come from. If this is not the case, then DIF is inferred for that item and the item should be revised accordingly or removed from the assessment.

Given their focus on individual items, these traditional methods of identifying measurement bias may be less applicable to AI-based assessments, which often involve complex models of predictors rather than a set of items that are all intended to measure the same characteristic(s). Unless the predictors that are included in the AI algorithm can all be considered indicators of the same underlying characteristic, DIF analyses cannot be readily applied to AI-based assessments. However, analogous concepts may serve the same goals in the context of AI-based assessments. For example, ensuring a training dataset that is representative of individuals from diverse demographic groups, and with different backgrounds and experiences, could ensure that the underlying algorithm (machine learning or otherwise) incorporates diverse perspectives and content. Similarly, conducting a sensitivity review of the AI assessment—including the features incorporated into the AI algorithm and the outcome that the algorithm is intended to predict—could also identify aspects of the content and perhaps the algorithm that might have different meanings, interpretations, or limited access for members of some subgroups. To be effective, this type of sensitivity review also requires people with diverse perspectives, backgrounds, and experiences to evaluate the algorithm and the outcome used to train it. Nevertheless, an important limitation of conducting a sensitivity review in this context is that AI algorithms sometimes incorporate thousands of features to predict an outcome, and it may not be practical or feasible to conduct a comprehensive and thorough review of all of them. In addition, even if individual features pass the sensitivity review, it could be that a combination of such features does not. Note that if the AI-based assessment does include individual measures comprised of multiple items that assess the same characteristics (e.g., a measure of conscientiousness), then DIF analyses, as described in the *Principles*, may still be useful to apply to *those specific measures*.

Although traditional methods of examining measurement bias (e.g., DIF analyses) may not apply to AI-based assessments in all cases, measurement bias still needs to be considered. Any form of irrelevant variance that affects the relationship between a feature in the algorithm and the outcome it is intended to predict can be a source of measurement bias. Examples of irrelevant variance include the extraneous information provided by some data sources (e.g., social media) that is not related to the job or individual attribute being predicted, the increased media richness (e.g., graphics, videos, animations, sound) inherent in some AI-based assessments that can influence how individuals from various subgroups respond to or interact with the assessment (Ryan & Nye, 2022), factors that affect data collection and/or feature

extraction differently for some groups (e.g., differences in microphones, video cameras, internet connectivity, cell phones), or subgroup differences in familiarity with the technology used to gather data for the assessment (e.g., computers, audio-visual equipment). These and other sources of irrelevant variance degrade the AI-assessment's ability to reflect the intended job-relevant characteristics. In this way, these sources of irrelevant variance are a form of measurement bias (subgroup differences in the measurement of the intended characteristic) that can be detected using the methods described in the previous section on predictive bias.

It is important to note that the definition of measurement bias used here is fundamentally different from the forms of bias discussed in many other disciplines involved in the development of AI-based assessments. Given the definition provided above, measurement bias focuses on *irrelevant* factors that affect individuals' scores on the assessment. In contrast, other definitions of bias in the AI research literature tend to consider score differences caused by *any* attribute being measured, whether these attributes are relevant to the characteristics being measured or not. From this perspective, any differences on an AI-based assessment that are observed across subgroups of individuals (e.g., groups defined by race/ethnicity or gender) could be perceived as bias. However, from the perspective outlined in the *Principles*, it is possible to observe mean differences in scores across subgroups even if there is no bias in the assessment. As a result, bias mitigation strategies that involve eliminating subgroup differences in an algorithm are likely to leave the problem of measurement bias unresolved and may also be illegal (e.g., race norming explicitly violates the Civil Rights Act of 1991). By extension, other approaches to bias mitigation can also violate employment law if job applicants from some subgroups are required to meet different standards in order to be hired (e.g., when features of an AI-assessment are implicitly or explicitly scored differently across subgroups).

Adverse impact is a legal concept in employment testing that is distinct from but related to measurement bias. Adverse impact can occur when subgroup differences on an assessment result in members of one group being disproportionately selected over members of another group. Although measurement bias can cause the subgroup differences that result in adverse impact (Nye & Drasgow, 2011), adverse impact can also result from subgroup differences in the work-relevant attributes being measured. Thus, test developers need to assess and attempt to remedy any measurement bias that is found, in addition to finding the most accurate measures of job-relevant characteristics.

3.3 Section Summary

In sum, fairness is a social rather than a psychometric concept that requires establishing a shared definition before it can be addressed effectively in AI-based assessments. In contrast, bias, as described here, is a psychometric term and is one of the most important concerns for AI-based assessments. Although some traditional forms of detecting measurement and predictive bias may still apply to AI-based assessments, other forms may be less relevant or will require different approaches. In addition, new forms of bias will also need to be considered (e.g., bias in the training data). Nevertheless, the focus of all these bias detection methods should be on detecting sources of variance that are *irrelevant* to the job or the attribute being measured. Subgroup mean score differences and adverse impact in hiring rates are also very important

concepts and information that is relevant to employment testing but should not be relied upon to indicate psychometric bias.

Section 4. Operational Considerations and Appropriate Use of AI-based Assessments for Hiring

The sections above describe specific characteristics that AI-based assessments should have to support their use for making hiring and promotion decisions. The process of evaluating these characteristics and collecting the evidence necessary to support the intended use of the assessment takes time and requires several steps. This section outlines the most important considerations that can arise throughout this process, including issues related to collecting validity evidence and analyzing the data from the assessment. Once the decision to use an AI-based assessment has been made, there are also factors that may affect the administration of the assessment. These issues are also covered in this section.

As noted in the introduction to these guidelines, many of these issues are discussed in more detail in the *Principles*. Nevertheless, the sections below specifically describe issues that are unique to AI-based assessments.

4.1 Initiating a Validation Effort

4.1.1 Defining the Organization's Needs, Objectives, and Constraints

As with other selection procedures, users of AI-based assessments should identify relevant stakeholders, determine their goals and objectives, and discuss and minimize any conflicts among them that arise. Here, the cost of building and implementing an AI-based assessment can be a significant concern. When this is the case, a cost-benefit analysis will help to determine the value of the AI-based selection procedure relative to its costs. However, users should be cautioned to avoid prioritizing one objective (e.g., lower costs, speed, efficiency) at the expense of others (e.g., obtaining valid assessment scores) that are also important to the employer.

4.1.2 Climate and Culture

The business environment, demands on the organization, and the employer's labor and legal history can all affect the extent to which an AI-based selection procedure is likely to be effective in the hiring context and accepted by the various stakeholders. In addition, the culture and climate of the organization may inform the type of validation study that is feasible (e.g., an organization with a data-driven culture may seek greater amounts of empirical evidence prior to adoption of a tool).

4.1.3 Sample Size and Availability

The validity of most AI-based selection procedures is based on a complex combination of factors that contributed to the development of the AI algorithm (e.g., sample demographics, data

used for input, choice of algorithm, and its parameters). Each of these factors should be investigated for their contributions to validity. In some cases, the number of respondents available to develop or validate the algorithm will be relatively small, compared with the number of predictors that can be used. As noted earlier, the use of regularization methods that are common in many AI models may help to deal with sample sizes that are less than ideal for traditional modeling methods. Nevertheless, use of small samples generally yields unstable estimates of validity and imprecise predictions. Although AI models are often better than ordinary least squares regression at detecting unexpected interactions and higher-order terms, these are not common in psychological measurement, and there can be cases where ordinary least squares regression performs better in cross-validation than alternative machine learning models developed on the same data.

In some cases, the validation of models for AI-based assessments requires little or no participation from employees or their managers. For example, data can be scraped from social media and a single manager can designate who the "good" employees are or vendors can develop the scoring algorithm for an AI-based assessment based on data collected from their previous clients. These approaches are risky if they are not clearly job relevant (e.g., based on a well-conducted job analysis), accompanied by evidence that the validity of the scores generalizes to the new setting (see the section on "Generalizing Validity Evidence" in Section 1.3), or based on objective data sources.

For all validation studies, participants must be willing to participate in the study and provide honest and accurate information. Labor organizations must cooperate and ideally should encourage their members to participate. To facilitate this cooperation, data collection of both predictors and criteria should be done at times that are convenient for both businesses and employees. Differences between incumbents and the applicant pool (e.g., differences in capabilities, work experiences, education, demographic characteristics, etc.) should also be taken into account. Organizations seeking to upgrade the capability of their workforce should also consider the impact of an AI-based selection procedure that is based on the characteristics of lower-skilled incumbents.

4.1.4 Sources of Information

Caution is advised when using data that the applicant did not provide for the purposes of evaluation (e.g., social media or other online data) as input to an AI-based assessment because these data may reflect job-irrelevant content. Input to AI-based tools may come from a variety of sources, such as, workers or job applicants, their managers, and archival records of training and performance. Input from applicants or incumbents may come directly and/or indirectly. For example, an applicant may provide a response to a question (i.e., a direct input) while their reaction time and cursor location are being captured (i.e., an indirect input) simultaneously. In addition, some of these tools rely on information created for purposes other than employment, such as social media posts, that may be dated, inaccurate, or missing for some applicants. Users should be aware that in some jurisdictions in the U.S., users must inform applicants when this kind of information is used for employee selection. As a general rule, applicants should be informed about the kinds of data that are used in evaluating their job suitability.

4.1.5 Communicating the Validation Plan

Managers and workers should understand the purpose of the validation research, their roles in it, and the impact of their participation. In particular, the level of confidentiality associated with predictors and criteria should be stated. Communications regarding the validation plan should be clearly communicated in language that is easily understood by the intended audience.

4.2 Understanding Work and Worker Requirements

To conduct a successful validation study and ensure that an AI-based assessment is relevant to the job for which individuals are applying, it is critical to first understand the work that will be performed and the worker characteristics that are required to perform this work. In the context of AI-based assessments, this information can be used to identify appropriate outcomes for training the AI algorithm or for a criterion-related validation study (see Section 1.1.1). In addition, understanding the work and worker requirements can also help developers and users to identify *irrelevant* sources of variance that may contribute to measurement and predictive bias (see Section 3.2). Although no specific methodology for evaluating work or worker requirements is prescribed in the *Principles*, testing professionals are expected to analyze the work domain and define worker requirements. Sampling strategies should take into account factors such as the variations in the work, work locations, demographic characteristics, etc. The methodology, data collection methods, analyses, results, and implications of the work analysis for the validation effort should be documented.

4.3 Selecting Assessment Procedures for the Validation Effort

When selecting assessments, organizations may use strategies such as a test plan to determine the extent to which work-related characteristics will be measured, using the test plan to match such characteristics to the methods or features that will be used to assess them. It can be difficult to create a test plan for AI-based assessments in cases where some of the characteristics being measured – for example, facial expressions, voice quality, response time – may not be well-defined in advance and may not be directly matched to characteristics of the work. When it is unclear what is being measured, it is not possible to determine its relevance to job performance. Whether evaluating an AI-based or traditional assessment for use, connections to the work and to job-relevant outcomes should be considered prior to adoption.

4.3.1 Review of the Research Literature and the Organization's Objectives

Testing professionals should leverage the research literature to inform their choices about the assessment procedures and validation strategies to be used. Given the recency of the development of many AI-based assessments, the applicability of findings from research on other forms of assessments (e.g., personality-based assessments or video interviews not scored via AI) requires careful consideration.

4.3.2 Scoring Considerations

Testing professionals must ensure that AI-based tools can be administered and scored accurately and consistently across candidates. When algorithms are frequently updated, scoring is not consistent over time and candidates and changes must be adequately documented so erroneous inferences by end users (e.g., comparisons of candidates scored via different algorithms) do not occur.

4.3.3 Format and Medium

Testing professionals should be aware that format and medium (e.g., video versus written question stimuli; audio versus written responses) can affect mean score differences among subgroups. Likewise, scores from AI-based tools may not be comparable if administered using different formats or media. For example, an algorithm developed to score video interviews using one transcription tool may not be comparable with an AI assessment deployed using a different transcription tool.

4.3.4 Acceptability to the Applicants

As with traditional selection tools, users want selection procedures that are predictive of job performance (or other valued criteria such as turnover), administratively easy, legally defensible, acceptable to organizational stakeholders, and viewed positively by applicants. However, what is acceptable to applicants can vary widely depending on their perceptions of fairness. AI-based assessments may not be well-regarded by applicants if they are not viewed as being fair or relevant to the job. For example, some applicants may feel uncomfortable with an algorithm (e.g., rather than human beings) determining their scores, consider games to be trivial and not related to the job, or judge certain information pulled from social media to be inappropriate for making hiring decisions. Others may be offended because they were unaware how they were being evaluated, or more specifically, they do not know how predictors are used and weighted. In some jurisdictions, users of AI-based assessments are expected to inform applicants about the information that is used and the weights applied to each predictor in the model, although it may be difficult to communicate this information in ways that are clear to a lay person. Although face validity (i.e., the appearance that the measure is relevant to the job) is not an acceptable substitute for other forms of validity evidence, the lack of face validity for some features used in AI-based assessments may pose challenges to their acceptability. Research on technology acceptance, perceptions of AI-based tools, explainable AI, privacy concerns, and the like continue to emerge and should be considered in developing appropriate communication to candidates and other stakeholders.

Users must also consider the potential for AI-enabled assessments to be less accessible for job seekers with disabilities. Employers should share meaningful information about the data collected from and the operation of assessments so job seekers with disabilities can determine if they need to seek a reasonable accommodation (e.g., an alternative to an AI-based gamified assessment if the applicant is vision impaired or has slower reaction times than what is demanded of the game due to their disabilities).

4.3.5 Selecting Criterion Measures

When the source of the validity evidence is the relationship between predictor and criterion, the criterion chosen should be related to the proposed use of the selection procedure. Nevertheless, all criteria used should represent important aspects of work performance or relevant organizational expectations. Depending upon the purpose of the assessment, criteria related to organizational outcomes other than job performance may be appropriate (e.g., absenteeism, turnover). Criteria directly related to job performance (e.g., supervisor ratings, sales, error rates, productivity indices) are especially appropriate when the source of validity evidence is the relationship between predictor and criterion. When selecting a criterion, testing professionals should examine the psychometric characteristics of criterion measures and avoid criteria that show bias against demographic groups. Just as with applicant assessments, all criterion measures should also demonstrate adequate levels of reliability. Some AI algorithms are intended to differentiate groups, so the criterion is group membership rather than an indicator of work performance. When the criterion is group membership, the process and standards used to identify members of the group and the job or organizational relevance of these groups should be documented.

4.4 Other Circumstances Affecting the Validation Effort

4.4.1 Influence of Changes in Organizational Demands

As suggested in the *Principles*, testing professionals should examine the impact of organizational changes on the validation and use of AI-based assessment tools. Changes in organizational functioning and work requirements may necessitate modifications to existing AI-based assessments, adjustments to scoring, or the introduction of new assessments to ensure the continued validity of inferences about work performance or organizational expectations.

4.4.2 Review of Validation and Need for Updating the Validation Effort

Users of AI-based assessments should anticipate changes (e.g., in work, worker requirements or work settings) that could impact the continued validity of inferences made from an algorithm. They should also anticipate any changes in the sources of data that could impact the validity of inferences, such as changes in a social media platform's policies (e.g., privacy, available data fields) or in policies regarding work products used as input (e.g., retention of records, required elements of applications). An advantage of AI is that algorithms can learn and adapt over time, thus, AI-based assessments are more likely to be updated regularly than traditional assessments. However, continually changing algorithms, such that the basis for candidate scores vary over time, pose administrative challenges to ensuring the validity and fairness of inferences based on such scores. Therefore, organizations need to determine and document how often and under what conditions AI-based assessments will change and whether change occurs continuously or at a predefined time after sufficient validation data have been obtained. Regardless of how frequently AI-based selection procedures change, they should be monitored periodically to ensure they are performing as expected and are free from bias, such as predictive bias and measurement bias, as discussed in Sections 3.2.1 and 3.2.2.

The technical documentation for the assessment should also be updated if changes to the algorithm or the training data affect the interpretation or use of scores. Continually changing algorithms pose a specific challenge with regard to maintaining up-to-date technical

documentation. When changes affect the interpretation or use of scores, documentation must specify which versions of the algorithm produce comparable scores and, hence, can be used in the same way to make employment decisions.

4.5 Data Analysis

4.5.1 Data Accuracy and Management

As with traditional assessments, complete documentation should be created describing the precise types and sources of data captured for model training, validation, and implementation. Because some AI-driven approaches incorporate much more complex data than traditional measures, such as by integrating unstructured text, video, or audio data, the documentation necessary for such an approach must be commensurately complex to be considered complete. Such documentation should also fully present the transformation pathway from raw data to predicted values, including data capture, pre-processing, feature generation, model choice, hyperparameter tuning, model training, cross-validation, and the generation and use of any predicted values or categorizations. Documentation should be sufficiently detailed to enable a meaningful external audit of all steps of data processing along this pathway.

If third party data processors are used, documentation should also be created to explain what data are transmitted, what is done with those data by the external processor, and how details or data can be obtained from that processor for later analysis. Special attention should be paid to documentation of any potential transmission across national borders or data storage in countries other than the country in which such data were collected, as laws protecting data privacy, use, and storage may differ across countries.

4.5.2 Missing Data and Outliers

Some AI systems, such as those relying on social media data, may collect and analyze data that do not exist for all applicants. For example, a question in a video interview may be skipped or information from a LinkedIn profile may not have been entered. Testing professionals should examine these missing data carefully to determine if traditional statistical strategies for handling missing data are sufficient to ensure validity and fairness. Because data may be captured from multiple dissimilar sources, the strategies used to deal with missing data may differ across sources despite feeding into a single model. Technical documentation should describe how missing data are handled.

4.5.3 Descriptive Statistics

Descriptive statistics, including measures of central tendency, variability, and distribution, should be made available for all raw data and all features used as input in AI algorithms. Such statistics should further be provided for both the total group and relevant subgroups if large enough to yield reliable estimates, especially as related to class memberships protected by federal or state law, such as by race, color, sex, gender, national origin, religion, disability, and age.

4.5.4 Appropriate Analyses

The choice of AI algorithm, such as between ordinary least squares regression, random forests, or neural networks, should be explained and justified. For example, model developers commonly base their choice of algorithm upon existing published research, common rules of thumb, or by analyzing collected data. Because there is a wide array of algorithms, usually with no single definitive choice to be made, the choice of algorithm should be stated explicitly and justified given the nature of the data. If multiple algorithms are appropriate given the nature of the data, the impact of the choice of algorithms on validity, fairness, and other outcomes should be documented.

The ability to interpret and explain an algorithm differs depending upon the transparency of the algorithm's calculations and model output. Whereas some models produce coefficients that may be interpreted directly, such as those based upon the general linear model, other models are uninterpretable and require special techniques to make them explainable. Thus, a description of how the final algorithm can be interpreted and explained should be provided. For example, is it clear how changing an algorithm's input will result in changes to the output, or are such changes hidden such that they are only known when put into practice?

4.5.5 Combining Selection Procedures into a Selection System

The distinction between an AI selection procedure and AI selection system can be fuzzy, as not all AI selection procedures purport to measure psychological characteristics. As such, the distinction between algorithms intended to estimate an individual's standing on a particular characteristic and algorithms intended to combine scores should be considered carefully when determining where a selection "system" begins and ends. It may be that information from various algorithms (and potentially from other traditional assessments) is then combined into a final prediction or such information is provided to decision-makers to use holistically. Research has indicated the former approach is preferable in terms of maximizing validity, but organizations may adopt the latter for reasons of stakeholder acceptability and other feasibility concerns.

4.5.6 The Use of Person-Centric Modeling and Classification

The goal of some AI-based applications is to determine a classification to which the data for a specific person fits best (e.g., personality type). To be clear, because these approaches have been called into question (e.g., when there are no objective groups, when there is no organizational criterion to be measured and incorporated into the analysis), the theoretical and practical justification for such types or classes should be carefully specified with attributions to the appropriate research literature.

As classifications may be probabilistic in nature, individuals may belong to all groups to some degree, and therefore one should specify the decision rules that ultimately classify individuals into a discrete group, including descriptive statistics and plots of the posterior probabilities by class or type, and how multiple-class membership is handled and used in subsequent demonstrations of cross-validation and testing. Methods used to determine validity should be specified, including but not limited to application of signal detection theory (SDT) statistics such as "precision" and "recall" that require specification of cut-scores. Indices of effect size should be reported, such as *d*, area under the curve (AUC), or any index that conveys practical value alongside statistical significance. When such methods are used, complete

information regarding the underlying distributional form should be detailed (e.g., the shape of the receiver operating characteristic or ROC curve).

4.6 Communicating the Effectiveness of Selection Procedures

4.6.1 Expectancies and Practical Value

Understanding the relationship between AI systems and practical outcomes is essential for acceptance of their use. Research on transparency and the ability to explain AI results has pointed to the challenge of translating complex statistical analyses and approaches to lay audiences. However, it is incumbent on users of any assessment tool (AI-based or traditional) to ensure that all relevant stakeholders have sufficient information to judge the value of the tool. Both research on communicating the value of selection procedures (e.g., how to explain the concept of a correlation or to illustrate expected performance of those selected at various cut scores) as well as research on the ability to explain AI results (e.g., on the framing and depth of explanations) can help inform the best ways to provide the necessary information for organizational users to understand the extent to which an AI assessment is effective.

4.6.2 Utility

Projected gains (e.g., in productivity, in reducing turnover) from using a selection procedure are based on utility analyses, which consider costs, increments in validity, and other factors. The long-term costs and risks associated with the use of AI must be additionally considered beyond the costs normally identified in utility analysis. For example, some developers recommend continuously updating source datasets, which necessitates a long-term commitment to pay for the expenses associated with data collection and model updating. The use of third-party AI tools to support such systems, such as the use of third-party speech-to-text AI algorithms, may incur both long-term and per-applicant costs that should be considered in the context of utility analyses.

Importantly, the legal risks associated with the use of AI systems should be explored when considering the usefulness of an assessment, especially as the regulatory landscape for AI continues to evolve. Frequent revision of models, for example, reduces the value of past validation studies and documentation, which may harm the legal defensibility of an otherwise well-developed assessment. As laws change in various countries, the transmission and storage of data across international borders may incur additional practical and legal risks.

4.7 Administration of AI-Based Assessments

As with other assessments, AI-based assessments require guidelines for administration that include information about handling irregularities, scoring the assessment, and interpreting results. Administrative guidelines for AI-based instruments should outline the procedures to follow when a person declines to provide the information necessary for the evaluation or to be recorded. Many AI-based assessments are scored by the vendor who has responsibility for ensuring the accuracy of the scores. Guidelines for interpreting these scores that are consistent with the research evidence should be provided to the user.

4.7.1 Applicability

The description of the AI-based procedures should indicate for whom the procedure is applicable, state any exceptions to being assessed (i.e., qualifications exceptions, missing data allowances), provide information on the applicability of the assessment to individuals with disabilities and those from different cultural and linguistic backgrounds, and state any rules about when administration and retesting occur. When an AI-based assessment changes, the description should be updated as appropriate.

4.7.2 Administration Responsibilities

Many AI-based assessments are self-administered and do not require those taking the assessment to follow a specific protocol. Others, such as video-based interviews, require candidates to follow specific procedures. Regardless of the nature of the administration requirements, applicants should be provided with instructions for taking the assessment and opportunities for practicing for the assessment when appropriate.

Administrators should be trained on administration policies (e.g., test-retest), and data should be reviewed to detect compromises to the assessment's security. Although quality control mechanisms to ensure accurate and consistent administration as well as periodic reviews of pass rates should be built in, as with any other assessment, users of AI-based assessments may want to build in control checks related to drift (e.g., algorithmic drift if not a fixed algorithm or the changing nature of features in the population over time).

4.7.3 Information Provided to Candidates

As with other forms of testing, users of AI-based assessments should consider what information needs to be provided to those taking the assessment, with careful attention to what is legally mandated. The requirements regarding what must be shared with candidates are rapidly changing with the passage of international, national, state, and local laws. Users should be aware of relevant legal requirements and incorporate them as appropriate into their communications.

4.7.4 Guidelines for Administration

There should be documentation for AI-based assessments to convey information regarding instructions and practice, as well as information for those taking the assessment related to specific environmental conditions (e.g., noise, browser, monitor size) that may impact their performance. In addition, security procedures and consequences for not adhering to them should be clearly stated.

As noted above, most AI-based assessments have complicated scoring procedures that are executed by the vendor. Nevertheless, users should obtain sufficiently detailed interpretative guidelines from the vendor so as to be able to understand at a general level what types of features are considered and whether the algorithm is fixed or updated on some basis (and what that basis is). That is, the user should obtain sufficient information to understand the scoring process at a high level.

AI-based assessments should retain raw feature data as well as algorithmically derived scores for sufficient time periods to support periodic audits. Users should be aware that some jurisdictions require data and assessment results be destroyed while some regulatory agencies require retention of the same information. Additionally, for algorithms that are updated periodically, a means of linking raw data to the algorithm used at that point in time should be maintained in score databases. Information regarding how scores are reported, who has access to them, and appropriate and inappropriate uses of the scores should be included in user documentation.

As with other assessments, nonstandard implementation is a potential issue for AI-based assessments. For example, interruptions during administration may threaten the validity of an assessment. Although there are some AI-based assessments for which an interruption does not affect the score (e.g., assessments based on data scraped from social media), many assessments (e.g., games, video-recorded interviews) can be affected. Guidelines for administration should also indicate the procedures for documenting requests for reassessment and the process for doing so, including informing those taking the assessment of the conditions under which they can ask for a reassessment, the required process, and required timing of requests.

It is important to communicate, exercise, and enforce practices that protect the security of AI-based assessment documents (e.g., verification codes for access, content) as well as the procedures for scoring. Awareness of requirements (legal, company policy) regarding confidentiality, data privacy, and the persons to whom scores may be released are as important for AI-based assessments as for any other assessment.

Section 5. All Steps and Decisions Relating to the Development and Scoring of AI-Based Assessments Should be Documented for Verification and Auditing

Once an AI-based assessment has been developed and evidence to support its use and interpretation has been gathered, it is also critical to document all steps in the process so that they can be verified and reviewed by potential end-users. This step is critical for ensuring that stakeholders have sufficient information to evaluate appropriate uses and interpretations of the assessment. It is also increasingly important as auditing of AI-based systems becomes a more established practice and/or legally required. In this section, the key areas that need to be covered in this documentation are discussed.

5.1.1 Data Sources

As noted above, complete documentation should be created describing the precise types and sources of data captured for model training, validation, and implementation. Because some AI-driven approaches incorporate much more complex data than in traditional psychometric measure development, such as by integrating unstructured text, video, or audio data, the documentation required for such an approach may also have to be complex in order to be considered complete. A complete description of all sources of data should be provided including the rationale for their use, the relationship between the information that is collected and actual

job content, and/or other evaluations of job relevance. Documentation should be sufficiently detailed to enable a meaningful external audit of all steps of the data processing.

5.1.2 Validity Evidence

The technical report should provide a description of the validation studies conducted such that another testing professional could reproduce the analyses and results. The report should also describe the methods used by the testing professional to determine that the selection procedure is statistically and practically related to a criterion measure and/or representative of a job content domain. The documentation for criterion-related validation studies, when conducted, should report the following in detail: a description of the criterion measures; the rationale for their use; the data collection procedures; and a discussion of the measures' relevance, reliability, possible deficiencies, possible sources of contamination, and freedom from or control of biasing sources of variance. If the testing professional developed the criterion measure, then the report should include the rationale and steps taken to develop it, so it can be well understood and, if needed, replicated in future validation studies.

5.1.3 Characteristics of the Development and Validation Samples

The sampling procedure and the characteristics of the validation sample relative to the appropriate interpretation of the results should be described. The description should include a definition of the population that the sample is designed to represent, the demographic characteristics of the sample with respect to legally protected subgroups of individuals, sampling biases that may detract from the representativeness of the sample, the significance of any deviations from representativeness for the interpretation of results, and any statistical power analysis results (see Section 1.2.2 for a discussion of statistical power). Data informing the potential restriction in the range of scores on predictors or criterion measures are especially important. When a transportability study is conducted to support the use of a particular AI-based assessment in a new setting or context, the relationship between the original validation research sample and the population for which the use of the assessment is proposed should be included in the technical report. Test developers should make clear whether the psychometric characteristics described in the technical report refer to candidates or incumbents, and results for concurrent validation studies (i.e., studies in which the predictor and criterion data are collected at the same time point) should not be represented as results for predictive validation studies (i.e., when the criterion data are collected after the predictor data).

AI-based assessments are often developed in an iterative manner, where choices are made regarding the data sources to include or exclude based on cross-validated results, refinement of the assessment, and subsequent testing and generation of further results. It is important to fully describe the development process, including information about the decisions that are made about the data and the samples from whom the data were collected. When samples are repeatedly used in subsequent stages of any exploratory analytic procedure, the chronological sequence of how those samples were used should be specified.

5.1.4 Details of the AI Algorithm

Unlike the limited methods for conducting a traditional linear regression analysis, there are unlimited variations on AI algorithms that can be used to identify patterns in predictor data and their relationships to criteria of interest. Therefore, the specific algorithm (or combinations of algorithms in the case of ensembles) and related software package(s) used to conduct these analyses should be identified clearly, along with any modified or default settings. The information provided should be sufficient for the reader to replicate the analyses performed. If the algorithm is considered confidential for proprietary reasons, a more general description may suffice if information is also provided specifying how access to the complete details of the algorithm can be obtained.

5.1.5 Evidence of Data Sufficiency

Methods used to test the boundary conditions for data sufficiency in application of any algorithm should be detailed. Of critical importance is how the minimum amount of information needed for an appropriate analysis was determined, as well as the nature or range of conditions under which that was tested. For example, there may be a minimum amount of text that is required for training the AI algorithm.

5.1.6 Technological Requirements

If the assessment has specific electronic or technological requirements, the researcher should document these requirements and any accommodations that can be provided by the administrator for test takers with disabilities.

5.1.7 References

There should be complete references for all published literature and technical reports cited in the report. Technical reports completed for private organizations are often considered proprietary and confidential, and the testing professional cannot violate the limitations imposed by the organization. Consequently, some technical reports that may have been used by the testing professional may not be available for inclusion in the list of references.

References

Blackwell, D., Leaman, C., Tramposch, R., Osborne, C., & Liss, M. (2017). Extraversion, neuroticism, attachment style and fear of missing out as predictors of social media use and addition. *Personality and Individual Differences*, *116*, 69-72. https://doi.org/10.1016/j.paid.2017.04.039

Civil Rights Act of 1991 § 109, 42 U.S.C. § 2000e et seq (1991).

Cleary, T. A. (1968). Test bias: Prediction of grades of Negro and white students in integrated colleges. *Journal of Educational Measurement, 5*, 115–124. https://doi.org/10.1111/j.1745-3984.1968.tb00613.x

Hickman, L., Bosch, N., Ng, V., Saef, R., Tay, L., & Woo, S. E. (2022). Automated video interview personality assessments: Reliability, validity, and generalizability investigations. *Journal of Applied Psychology*, *107*(8), 1323-1351. https://doi.org/10.1037/apl0000695

Landers, R. N., Armstrong, M. B., Collmus, A. B., Mujcic, S., & Blaik, J. (2022). Theory-driven game-based assessment of general cognitive ability: Design theory, measurement, prediction of performance, and test fairness. *Journal of Applied Psychology, 107*(10), 1655-1677. https://doi.org/10.1037/apl0000954

Landers, R. N., & Behrend, T. S. (2022). Auditing the AI auditors: A framework for evaluating fairness and bias in high stakes AI predictive models. *American Psychologist.* Advance online publication. https://doi.org/10.1037/amp0000972

Liou, G., Bonner, C. V., & Tay, L. (2022). A psychometric view of technology-based assessments. *International Journal of Testing*, *22*(3-4), 216-242. https://doi.org/10.1080/15305058.2022.2070757

Morgeson, F. P., Brannick, M. T., & Levine, E. L. (2019). *Job and work analysis: Methods, research, and applications for human resource management*. Sage Publications. https://doi.org/10.4135/9781071872536

Nye, C. D., & Drasgow, F. (2011). Effect size indices for analyses of measurement equivalence: Understanding the practical importance of differences between groups. *Journal of Applied Psychology*, *96*, 966-980. https://doi.org/10.1037/a0022955

Park, G., Schwartz, H. A., Eichstaedt, J. C., Kern, M. L., Kosinski, M., Stillwell, D. J., ... & Seligman, M. E. (2015). Automatic personality assessment through social media language. *Journal of Personality and Social Psychology*, *108*(6), 934-932.

Putka, D. J., & Sackett, P. R. (2010). Reliability and validity. In J. L. Farr & N. T. Tippins (Eds.), *Handbook of employee selection* (pp. 9–49). Routledge/Taylor & Francis Group.

Ryan, A. M., & Nye, C. D. (2022). Fairness in technology-enhanced selection assessments: Promises and Challenges. In K. Geisinger & J. L. Jonson (Eds.), *Fairness in Educational and Psychological Testing: Examining theoretical, research, practice, and policy implications of the 2014 Standards* (pp. 187-210). American Educational Research Association.

Sackett, P. R., Laczo, R. M., & Lippe, Z. P. (2003). Differential prediction and the use of multiple predictors: The omitted variables problem. *Journal of Applied Psychology*, *88*(6), 1046–1056. https://doi.org/10.1037/0021-9010.88.6.1046

Schmidt, F. L., & Hunter, J. E. (1996). Measurement error in psychological research: Lessons from 26 research scenarios. *Psychological Methods*, *1*(2), 199-223. https://doi.org/10.1037/1082-989X.1.2.199

Society for Industrial and Organizational Psychology (2018). *Principles for the Validation and Use of Personnel Selection Procedures* (5th ed.). Bowling Green, OH: Author.

Speer, A. B. (2021). Scoring dimension-level job performance from narrative comments: Validity and generalizability when using natural language processing. *Organizational Research Methods*, *24*(3), 572-594. https://doi.org/10.1177/1094428120930815

Spisak, B. R., van der Laken, P. A., & Doornenbal, B. M. (2019). Finding the right fuel for the analytical engine: Expanding the leader trait paradigm through machine learning? *The Leadership Quarterly*, *30*(4), 417-426. https://doi.org/10.1016/j.leaqua.2019.05.005

Svendsen, G. B., Johnsen, J.-A. K. Almas-Sorensen, L., & Vitterso, J. (2013). Personality and technology acceptance: The influence of personality factors on the core constructs of the Technology Acceptance Model. *Behaviour & Information Technology*, *32*(4), 323-334. https://doi.org/10.1080/0144929X.2011.553740

Tippins, N. T., Oswald, F. L., & McPhail, S. M. (2021). Scientific, legal, and ethical concerns about AI-based personnel selection tools: A call to action. Personnel Assessment and Decisions, 7(2). https://doi.org/10.25035/pad.2021.02.001

Yuan, S., Kroon, B., & Kramer, A. (2021). Building prediction models with grouped data: A case study on the prediction of turnover intention. *Human Resource Management Journal*. https://doi.org/10.1111/1748-8583.12396

Glossary of Terms

Adverse Impact
A legal concept that occurs when subgroup differences on an assessment result in members of one group being disproportionately selected over members of another group.

Artificial Intelligence
A broad range of technologies and statistical techniques that have the potential to identify patterns of behavior and predict outcomes.

Convergent Validity
Validity evidence that is obtained by demonstrating that the scores produced from an assessment are related to scores from one or more assessments intended to measure the same characteristic(s).

Criterion-Related Validity
Validity evidence that is obtained by demonstrating a relationship between a set of predictors and work-related behaviors or outcomes.

Content-Related Validity
Validity evidence that is typically obtained by asking subject matter experts to evaluate the match between the content of an assessment and the tasks performed on the job, the worker requirements, or the work products that are generated on the job.

Differential Item Functioning (DIF)
A statistical analysis that is used to test for measurement bias in assessments with multiple items that are designed to measure the same underlying characteristic.

Discriminant Validity
Validity evidence that is obtained by showing that the scores produced by an assessment are less related to measures of distinct characteristics that it was *not* designed to measure than to scores from other measures that assess similar characteristics.

Factor Analysis
A statistical technique that models the relationships between observed variables (e.g., items in a scale) and the underlying characteristics that they are supposed to measure.

Fairness
A social concept that has many potential definitions. It is critical to develop a shared understanding of fairness before discussing fairness for a specific assessment.

Ground Truth
A term used in the AI and machine learning literature to refer to the outcome measure that is used to train the AI algorithm.

Hyperparameter

A term used in the machine learning literature to refer to characteristics of the model that can be adjusted or "tuned" by the developer to place constraints on the model or otherwise control the algorithm. Examples include placing constraints on the number of iterations used to estimate a model or on the size of a neural network.

Job Relevance
The inference that scores on an assessment are related to some aspect of the job such as job performance, other job outcomes (e.g., turnover), tasks performed on the job, worker requirements, or the work products that are generated. Job relevance can be demonstrated using criterion-related validity evidence or content-related validity evidence.

Measurement Bias
As defined in the *Principles*, measurement bias occurs when sources of irrelevant variance influence scores for subgroups of individuals such that members of the subgroup score systematically higher or lower than a reference group.

Model Training
The phase in the development of an AI algorithm in which the model is actually estimated.

Predictive Bias
As defined in the *Principles*, predictive bias occurs when a common regression line estimated in the full sample consistently over- or under-predicts an outcome for subgroups of individuals within the sample.

Pre-Processing
Refers to the steps taken to identify and prepare the data prior to analyses.

Psychometrics
A scientific discipline that focuses on the objective measurement of psychological characteristics as well as developing, refining, and improving assessments of these characteristics.

Reliability
Evidence that the scores produced from an assessment are consistent across replications of the assessment. These replications can occur over time, across relevant samples of behavior, across algorithmic predictions within an ensemble, or across raters/observers.

Sensitivity Reviews
Involves asking diverse groups of individuals to review the content of an assessment to determine whether the meaning or interpretation of the content might differ across subgroups.

Statistical Power
The ability to detect a relevant predictor among the features incorporated in the algorithm during the training phase.

Training Data
Data that is used to develop the AI algorithm.

Transportability
A strategy for generalizing validity evidence that was accumulated in one work setting or context to another. This strategy consists of identifying important similarities between the previous settings or contexts and the new setting or context the validity evidence will be applied to.