Psychometrika (2026), **00**, 1–22
doi:10.1017/psy.2026.10110

PSYCHOMETRIC SOCIETY

THEORY & METHODS - FOCUS ARTICLE

Fairness Issues and Evaluation in Psychometrics and AI/ML: What Can We Learn from Each Field?

Ying Cheng

Department of Psychology, University of Notre Dame, USA
Email: ycheng4@nd.edu

(Received 30 May 2025; revised 14 January 2026; accepted 23 March 2026)

Abstract

Fairness in psychological and educational measurement has been a long-standing issue. Meanwhile, fairness is an issue that has been receiving an increasing amount of attention in the machine learning (ML) and artificial intelligence (AI) community, particularly in the last decade. While there is some recognition by the AI/ML community of the early roots of fairness definitions and operational metrics in the testing context, especially the parallel between fairness in AI/ML and test fairness in a regression and prediction context, there has not been a systematic introduction about the rapidly developing research of AI/ML fairness and bias mitigation to the measurement professionals. There is an urgent need to build a shared foundation of fairness notions and criteria between the two fields, particularly given that AI/ML has started to play an ever increasingly important role in measurement and personnel selection. This article therefore draws parallel between the testing workflow and the AI/ML fairness paradigm, and explores the similarities and dissimilarities of fairness definitions and operational ways to evaluate fairness in both fields, as well as ways to address and mitigate bias. The exploration leads to a discussion of areas of future research, training, and collaborations.

Keywords: artificial intelligence (AI); bias; fairness; machine learning (ML); measurement

Introduction

Fairness in psychological and educational measurement has been a long-standing issue. The concept originated primarily in the context of standardized testing and aptitude testing, where concerns arose about whether tests were fair for individuals from different demographic backgrounds, including different cultures, races, socioeconomic statuses, and educational experiences. Critics often pointed to score differences between groups, oftentimes higher mean scores for individuals from privileged groups, while lower scores for those from disadvantaged groups, implicating inaccurate and unfair conclusions about people's abilities. In the early days of testing, especially with intelligence testing, there were concerns over the potential biases inherent in these assessments for respondents from different cultures and languages. This prompted research on culture-fair and culture-free testing (Cattell, 1940), which relies on geometric figures and spatial manipulation instead of language-based measures. Since the Civil Rights Act of 1964, the testing field has more actively responded to the need to assess test and item bias, especially in an employment testing context (Cole, 1993).

© The Author(s), 2026. Published by Cambridge University Press on behalf of Psychometric Society.
This is an Open Access article, distributed under the terms of the Creative Commons Attribution licence (https://creativecommons.org/licenses/by/4.0), which permits unrestricted re-use, distribution and reproduction, provided the original article is properly cited.

https://doi.org/10.1017/psy.2026.10110 Published online by Cambridge University Press

Fairness in psychological and educational measurement is an important issue because test results often have significant consequences or implications, from informing instruction or clinical diagnosis to influencing high-stakes decisions, such as college admissions, personnel selection, or even legal matters. If tests are not fair, they can perpetuate inequality, reinforce stereotypes, and hinder opportunities for certain groups of people. Ensuring fairness in testing in turn leads to more just societal outcomes. As such, the ongoing work to evaluate and improve fairness in measurement continues to be a key area of focus for researchers, educators, and policymakers. Nowadays, most large-scale educational testing programs have built in routines for fairness checks at the test and item level (Geisinger, 2022).

Meanwhile, fairness is an issue that has been receiving an increasing amount of attention in the machine learning (ML) and artificial intelligence (AI) community, particularly in the last decade. In that context, fairness refers to the principle of ensuring that the data and algorithms or models developed based on the data lead to decisions without bias, discrimination, or unfair treatment toward any individual or group. Initially, the focus of AI/ML research was on ensuring that AI/ML systems were technically effective, but as AI/ML began to be applied in more critical sectors, such as hiring, criminal justice, and healthcare, concerns about biased outcomes and unfair treatment arose. This led to a growing recognition of the need to address fairness in the design and deployment of ML/AI models. Today, fairness has become a key area of research, with the establishment of dedicated forums like the ACM Fairness, Accountability, and Transparency (FAccT) conference (https://facctconference.org/) in 2018.

There is some recognition by the AI/ML community of the early roots of fairness definitions and operational metrics in the testing context (Caton & Haas, 2024), especially the parallel between fairness in AI/ML and test fairness in a regression and prediction context (Hutchinson & Mitchell, 2019). Within the measurement community, there is also growing recognition of the connection on fairness between psychometrics and AI/ML. Existing work generally falls into two categories: (1) empirical research that applies or adapts psychometric fairness evaluation techniques to AI/ML systems (Suk & Han, 2024), or that draws on psychometric theory to inform fairness assessment in automated scoring (Johnson et al., 2022; Johnson & McCaffrey, 2023) and (2) broader discussion of responsible AI in education, encompassing issues, such as security, privacy, transparency, and so on (Burstein, 2023; Johnson, 2025). In both lines of work, however, there is limited space and attention that can be devoted to systematic treatment of fairness issues that illustrate the connections and differences between the two fields in their philosophical underpinning, statistical definitions, and operational framework to evaluate and protect fairness.

At the same time, the need for such a systematic discussion is more pressing than ever as AI/ML begins to make rapid advances in the measurement field, for example, in automated scoring of essays (Dikly, 2006; Ercikan & McCaffrey, 2022; Ke & Ng, 2019; Shermis, 2024; Shermis & Burstein, 2003, 2013), in AI-powered screening and recruitment of job candidates (Hunkenschroer & Luetge, 2023), and automated generation of questions (Circi et al., 2023; von Davier, 2018). Currently, *Psychometrika* has a special issue call for papers on leveraging AI to empower the development and application of diagnostic statistical models. There is an urgent need for researchers in both areas to understand the concept, framework, terminology, and methodology related to fairness in the other field. This focus article therefore first draws parallel between the testing workflow and the AI/ML fairness paradigm, then narrows down to the context of fair selection, in which it explores the similarities and uniqueness of fairness definitions and operational ways to evaluate fairness in both fields, further discusses ways to address and mitigate bias, and underscores fairness evaluation as a dynamic, long-term endeavor. The exploration will lead to a discussion of areas of future research, training, and collaborations.

1. Fairness in testing

In testing, some argue that there is no single objective, universally accepted definition of fairness, and that any fairness definition is inherently a value decision (Darlington, 1971) or ethical position (Hunter

https://doi.org/10.1017/psy.2026.10110 Published online by Cambridge University Press

& Schmidt, 1976). Nonetheless, different approaches have been developed to define and evaluate fairness in psychological and educational measurement over time. Historically, test fairness has deep roots in the predictive validity context, that is, how test scores predict future outcomes or criterion variables for different groups of people. Various definitions of test fairness have been proposed and debated in the literature when outcomes are regressed on test scores (Humphreys, 1952). According to Anastasi (1968), "Test bias refers to overprediction or under-prediction of criterion measures" (p. 559). More specifically, Cleary (1968) defined a test to be fair when the regression lines in different groups have the same variances of errors of prediction, slopes, and intercepts. Test bias is considered present when "there is a group difference in any one parameter" (Humphreys, 1986). This approach was later challenged by researchers from various philosophical and methodological perspectives (Cole, 1973; Flaughter, 1974; Linn & Werts, 1971; Thorndike, 1971) but *differential prediction* is still considered one of the classic definitions of test fairness and has far ranging impacts (Drasgow, 1987; Peterson & Novick, 1976). Interestingly, this is also the context in which most AI/ML fairness notions align with test fairness, which will be discussed in detail later.

Starting from the late 1970s and early 1980s, the focus of the test fairness discussion started to shift to measurement bias based on latent variable models, such as item response theory models (Lord, 1977) and factor analytic models (Geisinger, 2022). This shift from relational equivalency (i.e., equivalent relationship with an external variable) to measurement equivalency (internal evidence of non-bias) was noted in Drasgow (1987). A focus on the item-level bias instead of at the entire test level also began to emerge, with differential item functioning (DIF) gaining popularity in educational and psychological testing (Humphreys, 1986; Rogers & Swaminathan, 2016).

More recently, the test fairness consideration expanded to increasing accessibility of tests and maintaining validity of test scores for individuals with disabilities, such as providing accommodations (e.g., extending time, presenting written text through auditory means) and modifications (Sireci et al., 2005). Universal design elements are adopted for large-scale testing programs such as the National Assessment of Educational Progress (NAEP) (https://nces.ed.gov/nationsreportcard/about/accom_ude_descriptions.aspx). Cultural responsivity is another recent focus in test fairness research (Sinharay & Johnson, 2023) to ensure that test takers from all backgrounds and cultures are treated fairly.

The evolution of the scope, definition, and evaluation approaches of test fairness is reflected in the bible of testing, that is, *The Standards for Educational and Psychological Testing* (hereafter referred to as the *Standards*). In 1954, the American Psychological Association (APA) published the *Technical Recommendations for Psychological Tests and Diagnostic Techniques*. In the following year, the American Educational Research Association (AERA) and the National Council on Measurements Used in Education (NCMUE; now the National Council on Measurement in Education or NCME) co-released the *Technical Recommendations for Achievement Tests*. The APA, AERA, and NCME started to jointly create *Standards* in 1966, titled "*The Standards for Educational and Psychological Tests and Manuals*." In these early documents, "fairness" was rarely discussed, if mentioned at all. Since 1985, test fairness has started to take center stage in the *Standards*, first primarily in the predictive validity context, and later expanded through the 1999 and 2014 versions. In the 2014 *Standards*, fairness is defined as "responsiveness to individual characteristics and testing contexts so that test scores will yield valid interpretations for intended uses," which clearly indicates that fairness is not viewed only through the lens of the outcome, that is, test scores, or predictions and decisions rendered based on the test scores. Instead, fairness should be considered in the entire assessment process, from test design (e.g., universal design and cultural responsivity), test administration (e.g., testing in a private room with little distraction), test scoring, score reporting, and the interpretation and use of test scores. Further, it clarifies that a test that is fair needs to meet the following requirements:

1. Reflects the same construct(s) for all test takers, and scores from it have the same meaning for all individuals in the intended population.

https://doi.org/10.1017/psy.2026.10110 Published online by Cambridge University Press

2. Does not advantage or disadvantage some individuals because of characteristics irrelevant to the intended construct.
3. Considers to the degree possible, characteristics of all individuals in the intended test population throughout all stages of development, administration, scoring, interpretation, and uses.

In summary, within this paradigm, fairness predicates on adherence to the intended construct throughout the testing process for the intended population. Fairness therefore is an integral piece of ensuring test validity for all members of the intended population.

Given the familiarity of the Psychometrika audience with test fairness, in the following, we will focus more on explaining the definition and evaluation of fairness in AI/ML and explore connections to test fairness as we go along.

2. Fairness in machine learning

2.1. Definition of fairness in machine learning

A typical *supervised* ML application involves building a model by training it on a dataset, which contains labeled examples that represent the patterns or relationships the model needs to learn. During the training phase, the model adjusts its internal parameters to minimize errors or improve its predictive accuracy based on the data it receives. Once the model is trained, it is validated using new, unseen data, or a hold-out sample (a portion of the data that was not used during training). This validation process helps assess how well the model generalizes to real-world situations, ensuring it performs reliably when exposed to data it hasn't encountered before. If the model's performance is satisfactory, it can then be deployed to make predictions or decisions in a production environment.

In ML, a model is deemed fair if it behaves or treats people "without discrimination or bias" against particular individuals or groups, especially concerning "sensitive characteristics or protected attributes such as age, sex, disability, ethnic or racial origin, religion or belief, or sexual orientation." (Weerts, 2021). Over the years, researchers in ML have realized that fairness is a nuanced concept and can be defined and evaluated in different ways.

Two central approaches to fairness are *group fairness* and *individual fairness* (Tang et al., 2023; Weerts, 2021). *Group fairness* ensures that outcomes from an ML model do not disproportionately favor or disadvantage any specific group, with pre-determined group membership. Groups are typically defined based on sensitive or protected attributes like race, gender, or age. The goal is to ensure that the model's prediction or classification performance is equally good across these groups. For example, in the context of college admissions, group fairness might mean an equal percentage of applicants admitted from different groups.

Individual fairness, on the other hand, emphasizes that similar individuals should be treated similarly. It focuses on the idea that the model should not make different decisions for individuals who are similar in terms of relevant features, even if those individuals belong to different demographic groups. The goal is to ensure that the model treats each person fairly based on their unique qualities, rather than making decisions based on sensitive or protected attributes (Barocas et al., 2019; Dwork et al., 2011; Hardt et al., 2016). While the idea of individual fairness is general enough to be applicable in various practical scenarios, the specification of the similarity metric is a key point in these applications and is "not often straightforward" (Tang et al., 2023). For ways to learn the metric from data, please see Mukherjee et al. (2020) and Ruoss et al. (2020).

Other important fairness notions include causal and counterfactual fairness (Kusner et al., 2017; Plečko & Bareinboim, 2024), which deems a prediction fair if it would be the same in a *counterfactual situation* where the individual's protected attribute (e.g., gender and race) were different; and rich subgroup fairness (Kearns et al., 2018) and intersectional fairness (Gohar & Cheng, 2023), the first requiring a statistical fairness constraint to hold over an exponentially or infinitely large collection of

https://doi.org/10.1017/psy.2026.10110 Published online by Cambridge University Press

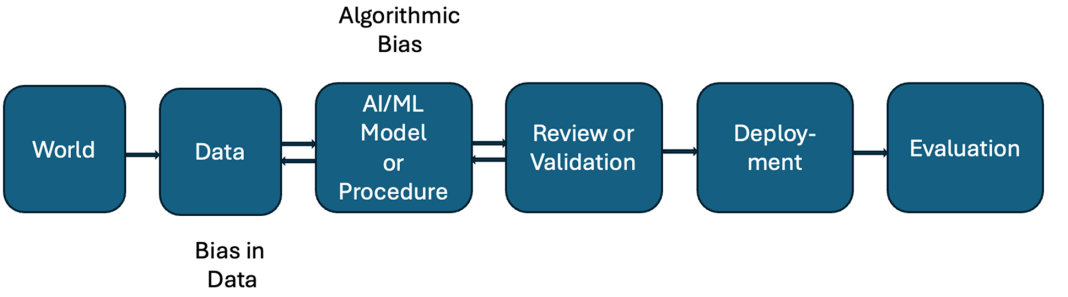


Figure 1. The flowchart of AI/ML applications.

subgroups instead of predefined ones, while the latter focusing on how bias emerges at the intersections of multiple protected attributes (e.g., race + gender) rather than along single dimensions.

On the other hand, fairness can be viewed from a temporal perspective, which looks at not just fairness at one moment, but how fairness evolves (or deteriorates) over time.

Traditionally, fairness is evaluated in a single-shot setting, typically when a model is deployed. But many real-world systems are often *feedback-driven* or *sequential*: decisions now affect future data, features, and populations. This means fairness-over-time is an important and much more complex issue than fairness at a given moment. Liu et al. (2018) showed that even simple one-step feedback models can lead to outcomes where applying a fairness constraint now doesn't necessarily lead to long-term fairness for disadvantaged groups—and in some cases can harm them. For example, a bank guided by the fairness principle tries to approve loans equally across groups. But because one group has lower starting credit scores (and thus different default probabilities), giving more loans to that group may increase default rates, their future credit scores may drop, and the group can become worse off over time. In some situations, there may not be a feedback loop but long-term fairness can be affected by sequential decisions or actions (Zhang & Liu, 2020). For example, in an educational context, a decision to admit (or not) certain students may affect their future outcomes (e.g., skills and networks) and thus can impact future datasets on labor market trends. If fairness is only checked at admission time, we might miss long-run unfairness in career outcomes. Later in this focus article, we will explain how the testing process itself is made up with sequential decisions and actions and potentially feedback loops. Hence, the notions of long-term fairness and fairness drift deserve critical attention there.

2.2. Sources of bias

Since fairness was defined as the absence of discrimination or bias, naturally, the types of biases and sources of them are central to the fairness discussion. Even though fairness in ML is often evaluated based on the eventual outcomes or decisions, various types of bias can enter at any phase. See Figure 1 for a workflow of ML applications. If the training data is not a true and fair representation of the phenomenon or process that it is trying to capture, the resulting model would be misled. Bias in the data could be due to historical reasons, for example, if historically fewer women were hired into a position than men, the resulting ML model for personnel selection could predict women to be hired at a much lower rate than men. Such a model, if deployed in hiring, would lead to fewer women in the workplace and create further biased data. This will produce a feedback loop discussed earlier that over time can exacerbate fairness concerns. Other biases that could enter the data include measurement bias (a familiar concept to the psychometric community), temporal bias, omitted variable bias, and so on.

Even if the data are a true and fair representation of the world, the process of building and training the ML model could introduce bias. For example, the model may have missed features that are critical in reaching accurate prediction or classification decisions for certain groups, but not for others. In medical

https://doi.org/10.1017/psy.2026.10110 Published online by Cambridge University Press

applications, risk factors for a certain disease may differ between males and females. A model that considers only primarily risk factors for males may lead to much less accurate predictions for females. Biased models can produce biased data, which enter the training data for future models and create feedback loops that perpetuate and amplify the bias.

Once a model has been built, humans who are in the position to review its performance and make refinements may introduce additional bias (note the bidirectional relationships in Figure 1). Implicit bias that exists in human associations and judgment has been well documented (Gopal et al., 2021; Greenwald et al., 1998) and can be perpetuated in AI systems (Bai et al., 2025). Bias could also enter the deployment phase, when the model is misused. For example, a model that is primarily built based on the general population may be used on the clinical population, which can cause fairness concerns.

The end outcome can be biased due to data containing bias (data bias) or AI/ML model having bias (algorithmic bias), or bias introduced in the human review or implementation process, or a combination of them. In this article, we will focus on data bias and the algorithm/model bias in Figure 1, the two most explored biases that cause fairness concerns in AI/ML applications, and the feedback loop between them. Such a flow emphasizes the interplay between data and the trained model, and how bias can be amplified throughout this process. As shown in the next section, the testing process can be viewed as a chain of data and algorithms/actions. Therefore, such an interplay and compounded bias can be expected in the testing process.

3. Connections between fairness research in testing and AI/ML

3.1. Fairness in testing viewed through the lens of the AI/ML flowchart

To have a common framework for discussion of fairness in testing versus ML, let's consider a situation where a test was developed and administered for a sample of subjects, which generated testing data for them in the form of *responses, selected choices, ratings (e.g., teacher report or parent report), response times*, etc. These testing data, following certain rules or algorithms, get translated to scores (e.g., item scores, subscale scores, total scale scores, weighted sum scores, latent trait estimates, and so on). The scores, then following certain procedures or rules, will lead to decisions, such as pass or failure (e.g., in the context of licensure or certification), hiring or rejection (e.g., in the context of personnel selection), and admission or declination (e.g., in the context of college admission, or admission into a gifted education program). See Figure 2 for the testing process cast as a chain of actions and resulting data.

When comparing Figure 2 against Figure 1, it is straightforward for one to draw the analogy that scores are the data, the standard setting process is analogous to the AI/ML model, and review and implementation will follow (last segment in a red circle). But the analogy actually goes beyond scores and decisions based on them. In fact, the test itself is data. A test can be considered sample data generated by sampling (an action/decision) from a vast population of items that tap into the construct. This *domain sampling* framework can be used in the development and validation of tests. It posits that a test is designed by sampling a representative subset of the larger universe of content or skills that define the construct being measured. This approach assumes that no single test item or small set of items can "exhaust" the population of items for the construct; rather, a comprehensive test should randomly sample from the full "domain" of knowledge, behaviors, or abilities that the test is designed to assess, so that the test results can generalize to other similar situations or contexts. In classical test theory (CTT), one way to conceptualize reliability is as an indicator of consistent measurement across different samples of items (Lord et al., 1968). Consulting Figure 1, the population of items is the "world," and developing a test with unbalanced sampling could introduce "representation bias" in the test itself. In other words, (algorithmic) bias in domain sampling will lead to bias in the resulting test, which subsequently embeds bias in all steps that follow.

Bias could enter the system through test administration as well. The *Standards* talk extensively about universal design and testing accommodations, which are essential to ensure that students in need receive support that is "responsive to individual characteristics" or needs. Failing to account for individual needs

https://doi.org/10.1017/psy.2026.10110 Published online by Cambridge University Press

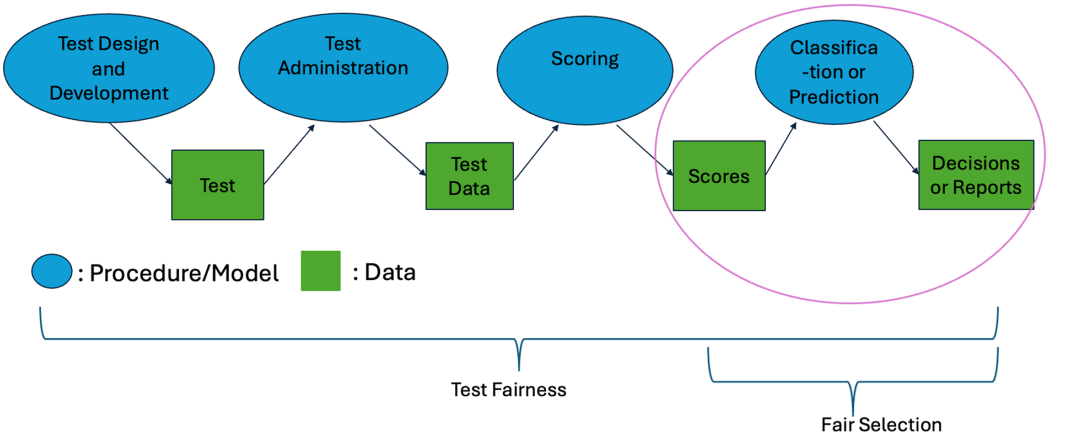


Figure 2. Testing process cast as a chain of actions and data.

would mean that the resulting testing data, such as responses and response times, may not reflect "the same construct(s) for all test takers," and may "advantage or disadvantage some individuals because of characteristics irrelevant to the intended construct." For example, some individuals require more time to answer certain questions, possibly due to the digital divide or dyslexia. Without proper accommodation, they may experience time pressure later in the test, leading to speeded responses that are not reflective of their true ability. It could also mean that scores may be missing for certain individuals, creating an unbalanced sample of the population. Such biases in the testing data will lurk in the scores, even if the scoring procedure itself does not introduce additional bias.

Scoring plays a critical role in determining the properties of the resulting scores. For example, for the same scale, the reliability of the unweighted total score across items, of the weighted total score across items, and of the maximum-likelihood factor score are all different (McDonald, 1999). The scoring algorithm can be considered a model translating raw testing data, such as item responses, into scores. A prime example of how the scoring model affects test fairness is the formula scoring. The SAT used to employ formula scoring, a method where test-takers would lose a fraction of a point for each incorrect answer, with the intention of encouraging students to leave questions unanswered if they were unsure, rather than guessing randomly on multiple-choice questions. This practice was eventually discontinued in 2016 in a SAT redesign (Manhattan Review, 2022). The College Board's significant decision to remove formula scoring from the SAT was related, at least in part, to the recognition that certain groups of students—particularly those from disadvantaged backgrounds—were more risk-averse when it came to guessing on questions. This behavior was influenced by a variety of factors, including test anxiety, cultural differences, and socioeconomic status. As a result, the scores no longer "reflect the same construct(s) for all test takers," and may "advantage or disadvantage some individuals because of characteristics irrelevant to the intended construct." Thus, the scoring algorithm itself may introduce bias to the scores, which affect subsequent decision-making.

Once the scores are produced, classification decisions can be made (or sometimes no decision is made but the scores get reported) following certain standard procedures. For example, the procedure may involve setting a pre-set cut score or a set of cut scores (a process that is systematically and rigorously carried out, sometimes referred to as "standard setting") (Reckase, 2022) and comparing one's score against the cut(s), or having an ML model trained with past data of test scores and job performance and applying it on new data to classify candidates. Biases can certainly be introduced through unfair classification algorithms (e.g., those learned from biased historical data as described earlier).

Highlighting the fact that the testing process involves a series of actions and data produced along the way, Figure 2 helps relate the fairness considerations in testing to the bias propagation framework

https://doi.org/10.1017/psy.2026.10110 Published online by Cambridge University Press

in AI/ML (Figure 1). Here, in Figure 2, we distinguish between actions following certain procedures, protocols, and/or rules as "model" or "algorithm" (denoted by ovals), and "data" generated as an outcome of the actions (denoted by rectangles). For example, test development is an action following an algorithm of domain sampling from a population of eligible items. It takes input data such as a test blueprint and/or an existing item pool and generates new data (i.e., a test) as a result. Test administration is an action following standard protocols that takes the test as input and generates testing data (item responses, response times, process data, etc.[1]) as new data. Scoring is an action following a scoring rule that takes raw response data and outputs scores as new data. The scores may be the end outcome in itself if no further action or decision is taken. In many situations, the scores (possibly in combination with other types of data) may be the input data for a classification or prediction action, which leads to the eventual decisions.[2]

From an ML perspective, all these actions/algorithms/models could have algorithmic bias. All the input and output data could have data bias. Bias could propagate and accumulate over the chain of sequential actions and data, or generate feedback loops through which bias gets amplified. There can be a close alignment between the test fairness and the ML fairness, when the testing process is cast in the framework of a chain of actions and data. The advantage of such a framework is increased awareness of the propagation of bias throughout the testing process, the importance of fairness in every step, possible feedback loops between steps, and an emphasis on long-term fairness rather than a one-shot evaluation. For example, if the scoring algorithm is biased, the resulting test scores contain bias, which in turn might affect the classification algorithm trained on the testing scores, and ultimately bias the decisions. The decisions later may be used to update the classification algorithm, creating a feedback loop that reinforces and exacerbates the bias.

Meanwhile, mapping out the testing process as a chain of actions and data similarly to an AI/ML application also helps illustrate how AI/ML can be used to support each step, with proper guidance of measurement theory and quality control. For instance, automated item generation (AIG; Gierl et al., 2021; Lathrop & Cheng, 2017) in test development, AI-assisted online proctoring (Belzak et al., 2025) in test administration, and automated scoring (Shermis, 2024) in scoring, feature selection (Mislevy et al., 2012), knowledge tracing (Lu et al., 2024), and so on. Johnson et al. (2022) and Johnson and McCaffrey (2023) are prime examples where measurement theory is infused in the fairness evaluation of automated scoring. More research beyond the context of automated scoring is urgently needed.

3.2. Operational ways to evaluate fairness: Narrow down to fair selection

Much of the discussion in AI/ML on how to evaluate algorithmic fairness is linked to prediction or classification decisions. This is also where the AI/ML community recognizes as most of the parallels exist between the two fields (Hutchinson & Mitchell, 2019). Hence, we will introduce a variety of commonly used fairness evaluation metrics in ML and show their connections to fairness evaluation in testing, from the *differential prediction/classification* perspective. Please note that the concept of test fairness has a much broader scope than *differential prediction/classification* as discussed earlier (and Figure 2 clearly contrasts test fairness vs. fair selection). In the fairness literature in psychological and educational measurement, researchers have made a distinction between tests that are fair and selection procedures that are fair (Cronbach, 1976; Crocker & Algina, 2008), which represents "data" and "procedure/model" fairness, respectively. What we will present next are various definitions of fair selection and ways to evaluate it, which is the last segment of Figure 2, that is, the selection algorithm that translates scores to decisions.

[1] Some scoring mechanisms explicitly take into account response times, problem-solving processes, etc. (see Duan & Cheng, 2024).

[2] The idea of the testing process passing "messages" (including data objects and data structures) on was described in Mislevy et al. (2012), in the context of evidence-centered design.

https://doi.org/10.1017/psy.2026.10110 Published online by Cambridge University Press

To simplify the discussion, let's assume that there is no bias in the scores. This could mean that there is no bias in any of the actions or data before it in the chain, or somehow miraculously the bias canceled out. There is only one step between the scores and the ultimate decision, that is, the classification algorithm. Suppose the decision is to hire or not to hire a job candidate. Let's also assume that the decision is made solely based on the test scores; there is no reliance on any other types of data.

In this context of personnel selection, based on the ML *group fairness* definition, individuals belonging to different (demographic) groups should have equal probability to be hired. Based on the definition of *individual fairness*, individuals who are similar should have equal probability to be hired. If we restrict our discussion to *group fairness*, there are (at least) three widely discussed statistical criteria in the ML literature to evaluate *group fairness:*

- *Independence*: Candidates from different groups have equal probability to be hired.
- *Separation*: Those who actually qualify (or not) (in true status) have equal probability to be hired (or not), regardless of group membership.
- *Sufficiency*: Those who are predicted to qualify (and thus hired) have equal probability to be qualified for the job, regardless of group membership.

Mathematically, independence suggests that the classification decision R is independent of group membership (G). Separation suggests that the R and G are independent conditioning on the true status Y. Sufficiency, on the other hand, indicates that Y and G are independent, conditioning on R. Please note that in the AI/ML context, true status Y is often a variable that is observable instead of a latent variable. For example, in student retention, whether the student drops out of college or not is their true status. If an AI/ML system is used to make predictions of student drop out, its prediction could be consistent or inconsistent with the true status.

3.2.1. Independence

In the context of testing, the independence criterion implies that the selection ratio p is equal across groups, regardless of true status: $P(R = 1) = P(R = 1|G)$, where $R = 1$ indicates a candidate being hired, and G is a categorical variable indicative of group membership. In ML, this is often referred to as "*demographic parity*," which turns out to have quite significant lineage in test fairness. While requiring the selection ratio to be strictly equal seems unrealistic, the 0.8 rule in personnel selection is often used to determine if a selection procedure is fair. According to the *Uniform Guidelines on Employee Selection Procedures (1978*), the selection ratio (or hiring rate) of the minority group needs to be at least 80% of that for the group with the highest selection rate. Otherwise, the test or selection instrument is considered to have "adverse impact" (Burgoyne et al., 2021).

Demographic parity was the central point of argument of *The Golden Rule lawsuit*, a landmark case involving Educational Testing Service (ETS) and the Golden Rule Insurance Company in 1984 (Lee & Zhang, 2010). The case revolved around allegations that the test used to license insurance agents discriminated against minority groups. The insurance company claimed that ETS's test, which was used for insurance agent licensing, was unfair because it had a disparate impact on minority candidates, particularly African Americans, and resulted in lower pass rates for them. The lawsuit argued that the test violated Title VII of the Civil Rights Act, which prohibits discrimination in employment practices.

Researchers have argued that avoidance of such "adverse impact" is in line with the concept of test equity, but not necessarily test fairness (Burgoyne et al., 2021). In fact, some researchers argued many years ago against the assumption that "different mean scores" for different groups of individuals would "by themselves, prove test bias" (Bray & Moses, 1972). It is also argued that meeting the independence criterion may not have any bearing on fairness, because randomly selecting candidates with the same selection ratio p from each group would meet the independence criterion, though it is not a meaningful selection procedure at all.

https://doi.org/10.1017/psy.2026.10110 Published online by Cambridge University Press

3.2.2. Separation

In the context of testing, the separation criterion implies that for someone truly qualifies or does not qualify, the hiring probability or selection ratio should be the same regardless of this individual's group membership. This suggests that the true positive rate (also known as sensitivity or recall in the ML literature) and the false positive rate are equal across groups: $P(R = 1|Y = 1) = P(R = 1|Y = 1, G)$, and $P(R = 1|Y = 0) = P(R = 1|Y = 0, G)$, where $R = 1$ or 0 indicates a candidate being hired or not, $Y = 1$ or 0 indicates a candidate qualifies for the job or not, and G is group membership. In ML literature, meeting the requirements of equal true positive rate and false positive rate is often referred to as "*equalized odds*." If only the false positive rate is required to be equal, it is called "*equal opportunity.*" The major distinction between independence and separation is that the latter is equality conditioning on the true status Y.

It turns out that the idea of conditional independence is not a foreign concept in test fairness. The idea could be traced back to at least as early as the 1970s, when personnel psychologists were examining closely the question of selection bias. For example, in 1973, Cole discussed the "conditional probability model," which requires that "for both minority and majority groups whose members can achieve a satisfactory criterion score… there should be the same probability of selection regardless of group membership." Here, achieving a satisfactory criterion score means the individual truly qualifies for the job. So the "conditional probability model" is consistent with the separation criterion in AI/ML.

Another key development in test fairness—DIF—is also based on the idea of conditional independence. An item j is free of DIF if $(X_j = 1|\ \theta)$ equals $(X_j = 1|\theta, G)$, where $X_j = 1$ indicates a correct response to item j. Here, θ is the true ability. Conditioning on θ, the probability of answering an item correctly or not is independent of group membership G. Given a binary response, the absence of DIF could also be defined as $\mathrm{E}(X_j = 1|\ \theta) = \mathrm{E}(X_j = 1|\theta, G)$. The difference between DIF and the separation criterion in ML is that the true ability θ is a continuous latent variable which needs to be estimated based on observed responses and an assumed IRT model, instead of an observable dichotomy like Y. Operationally, some DIF evaluation methods condition on the total score instead of θ: $P(X_j = 1|\ X) = P(X_j = 1|X, G)$, such as the logistic regression approach (Swaminathan & Rogers, 1990), which makes it more closely aligned with the separation criterion, or specifically, the equal opportunity criterion in ML. Yet another key difference is that X_j is not a decision that is made without knowledge of the true status Y. Instead, it is an indicator of θ, or it contributes directly to the matching variable X. So to be precise, DIF is not about prediction or classification; it is about internal bias instead of relational equivalency. For these reasons, we cannot draw a strict parallel between the separation or equal opportunity criterion in ML and DIF, though the idea of conditional independence underlies both.

3.2.3. Sufficiency

Based on the sufficiency criterion, in a personnel selection context, the proportion of correct positive predictions (i.e., out of those who are hired, the percentage of them who can successfully perform on the job) should be the same for different groups. Statistically, this means $P(Y = 1|R = 1) = P(Y = 1|R = 1, G)$ (or equivalently $P(Y = 0|R = 1) = P(Y = 0|R = 1, G)$). It is still a conditional concept, but it is not conditional on the true status; instead, it is conditioning on the positive decision. In the ML literature, this is often referred to as "equal precision," or positive predictive value (PPV) parity (Fraenkel, 2020).

In the testing literature, PPV is a statistic that was explicitly explored in the Taylor–Russell Table (Taylor & Russell, 1939). It tabulated the PPV rate (i.e., out of those who are selected, the percentage of people who are actually able to do the job), with a given base rate (the percentage of people capable of a job—true status) and a given classifier (i.e., a test with a certain predictive validity r & a selection ratio p). Please note that the base rate is essentially $P(Y = 1)$ and the selection ratio is $P(R = 1)$. The Taylor–Russell Table was produced assuming that the test scores and job performance are bivariate normal, and candidates with test scores above a certain threshold will be hired. The model of bivariate normality and comparing test scores against a predetermined threshold for all groups essentially constitute the classification algorithm in Figure 2.

https://doi.org/10.1017/psy.2026.10110 Published online by Cambridge University Press

In applications of the Taylor–Russell Table, much emphasis was placed on maximizing the PPV rate when the base rate, predictive validity r, and selection ratio p vary. At a given base rate and predictive validity r, the smaller the selection ratio p, the higher the PPV. At a given base rate and selection ratio p, the higher the predictive validity r, the higher the PPV. In order to reach high PPV for personnel selection, an employer should pick instruments of high predictive validity and be very selective. From a test fairness perspective, some researchers proposed to reach PPV parity (e.g., Linn, 1973).

In the context of fair selection, another related concept is the equal risk model (Einhorn & Bass, 1971), which requires that both majority and minority groups have the same *maximum* tolerable risk that a hired candidate will fail based on the regression equations. Sufficiency implies that the probability of failure for hired candidates should be the same between groups, that is, $P(Y = 0|R = 1) = P(Y = 0|R = 1, G)$. Therefore, though related, sufficiency is not the same as the equal risk model.

Notably, separation and sufficiency are both conditional criteria. The major distinction between them is that the former is conditioned on the true status Y, and the latter is conditioned on the decision or prediction. Interestingly, in economics, there is a history of both criteria. In examining salary discrimination, economists first regressed earnings on qualifications and gender ("forward regression"), and looked for gender differences in earning (an employer decision) conditioning on qualifications (true status). Later, they turned the question around and used "reverse regression," which examined gender differences in qualifications, conditioning on the same earnings (Goldberger, 1984; Green & Ferber, 1984). They can be viewed as the parallels of separation and sufficiency criteria in salary discrimination.

3.2.4. Other criteria

It has become clear that for the three most widely discussed statistical criteria for group fairness in AI/ML, independence is mapped to adverse or disparate impact, separation is mapped to the Cole model, and sufficiency was discussed in Linn (1973). For an example illustrating the application of these fairness evaluation criteria in a testing context, please see the Appendix.

In the ML literature, there are many other definitions of fairness and operational criteria, such as Calibration, which refers to ensuring that predicted probabilities match actual probabilities (i.e., base rate) within each group. In other words, $P(R = 1|G) = P(Y = 1|G)$. If the base rate is different between groups, the selection ratios will differ accordingly. This may require different decision rules for different groups. In test fairness literature, this is referred to as the *constant ratio* model by Thorndike (1971), who said that "the qualifying scores on a test should be set at levels that will qualify applicants in the two groups in proportion to the fraction of the two groups reaching a specified level of criterion performance." In other words, the selection ratio should be proportional to the base rate. However, this approach has similar problems to the independence criterion, that is, it does not use validity information at all. In other words, selection could be done randomly (without test score information) but still meet the constant ratio criterion.

There are likely many definitions and metrics of AI/ML fairness that are translatable to test fairness, or more specifically, fair selection, such as the overall equality of accuracy, treatment equality, equalizing disincentives (Caton & Haas, 2024; Jung et al., 2020; Zafar et al., 2017), and so on and so forth. As a field, we should research these definitions and metrics carefully and critically, and identify and possibly adapt ones that align well with our goals and contexts.

It is important to note that these various criteria are unlikely, or even impossible to be met simultaneously, except under special conditions (Baumann et al., 2022). For example, having equal p across groups meets the independence criterion but does not guarantee PPV parity, because the base rate and predictive validity r may differ across groups. Conversely, when PPV parity is met, it does not mean the selection ratio p or the predictive validity r is the same across groups. In the Taylor–Russell Table, we can find many instances where different combinations of p and r lead to the same PPV. For example, at a base rate of 0.2, $r = 0.3$ & $p = 0.05$ and $r = 0.4$ & $p = 0.15$ result in the same PPV of 0.41. It is therefore important to recognize that meeting all these criteria at the same time should not be the

https://doi.org/10.1017/psy.2026.10110 Published online by Cambridge University Press

goal. It may be necessary to choose the most relevant or appropriate metric, or balance between multiple metrics. The choice of the metric can be very context dependent and has legal implications.

3.3. Bias mitigation

If the chosen metric is not met, or in other words, there is statistical evidence of bias, bias mitigation strategies can be used. There is a rich body of literature in ML on bias mitigation, mainly categorized as preprocessing, in processing, and postprocessing strategies. Please note that the goal of these approaches is not only reducing bias, but also preserving the performance of the AI/ML models. This is an area of research in AI/ML that psychology as a field may not be very familiar with (except for some small pockets in I/O psychology). As psychometricians, this line of research in AI/ML fairness can be rather informative to us.

3.3.1. Preprocessing techniques (data-level mitigation)

Preprocessing methods involve modifying the training data to remove or balance out bias before training the model. This may involve adjusting the weights of the training examples or resampling the data to create a more balanced dataset. For example, in cases of class imbalance in the training data (e.g., if there were few women in the sample), you might oversample the underrepresented group or undersample the overrepresented one to balance the dataset (Kamiran & Calders, 2012). Or one could transform the data such that the transformed variables are independent of the sensitive attributes. For example, removing features from the data that are associated with strong group differences. In this case, the downstream model learns to make predictions without depending on these sensitive attributes. Another example is to remove the sensitive attributes themselves from the data. The principle is sometimes referred to as "fairness through unawareness" (Baker et al., 2023), where the model is designed to ignore sensitive attributes. The idea is that by not using sensitive attributes directly in the model, it can avoid discrimination or bias that might occur from these attributes.

3.3.2. In-processing techniques (model-level mitigation)

In-processing methods adjust the model during the training phase to minimize bias while still maintaining model accuracy. For example, models can be trained with explicit fairness goals ("constraints"), such as demographic parity, equalized odds, or equal opportunity, as we discussed earlier. Or this can be done through regularization approaches. For example, the objective function in training the ML model can include a penalty term for discrepancies between groups in selection ratio, PPV, or odds. Another approach is called adversarial debiasing. This approach tries to optimize the main model's prediction or classification decision while reducing the ability of an adversarial model to determine the protected attribute of an individual from the classification decision for him or her. In other words, the main model is trained to work well in its prediction or classification task while simultaneously maximizing the adversarial model's error or "fooling" the adversary, making it harder for the adversarial model to predict the sensitive attribute (Zhang et al., 2018).

3.3.3. Postprocessing (output-level mitigation)

Postprocessing techniques modify the model's output to ensure fairness after the model has been trained (Hasanzadeh et al., 2025). This may involve adjusting the model's output to ensure that the selection ratios are the same for different groups, by applying group-specific decision thresholds or boundaries. Notably, the thresholding may not be deterministic like what is assumed in the Taylor–Russell Table. Instead, some postprocessing techniques could involve solving a linear program to find probabilities with which to change decision labels (e.g., hiring or not) using probabilistic thresholding (Hort et al., 2023). Deterministically or probabilistically, such a practice violates the "fairness through unawareness" principle mentioned earlier.

https://doi.org/10.1017/psy.2026.10110 Published online by Cambridge University Press

3.4. Balance between performance and fairness

In the discussion of bias mitigation, it is evident that there can be tension between optimizing the model performance in prediction or classification and ensuring fairness. Striking a balance is complex and requires careful consideration of both objectives, often depending on the specific context and ethical implications of the application. Most bias mitigation strategies seek to reduce bias while preserving the prediction or classification performance. For a demonstration of how bias mitigation strategies work in an educational assessment context, please see Cheng, Pei and Liu (2025). In this example, the authors illustrated how various bias mitigation strategies can be used in conjunction with a variety of ML algorithms to balance the reduction of bias and the preservation of accuracy in predicting low-performance students early in a semester.

In measurement, such a tradeoff or balance has received explicit attention as well. For example, in personnel selection, the diversity–validity dilemma is a well-known issue (Parker & Dearing, 2009; Pyburn et al., 2008). Sackett et al. (2022) contrasted common selection procedures in their scores' validity (predictive validity in r, i.e., the correlation between test scores and job performance) and diversity (score differences between black and white candidates, measured by Cohen's d). The scatterplot of the two metrics shows that they are generally positively correlated, that is, procedures with higher r often come with higher d. In other words, procedures that are more predictive of the job performance tend to be associated with larger divergence in the average scores between groups, which may result in large discrepancies in selection ratios. Given this, Rottman et al. (2023) proposed to iteratively remove predictors that have large group difference–predictive impact ratio. Predictors with larger group difference–predictive impact ratio contribute substantially to average score difference but little to the prediction of the outcome, and are thus undesirable. This is considered a preprocessing method.

Or naturally, employers may want to use a combination of selection instruments and methods to help balance diversity and validity. In educational settings, researchers have advocated for holistic reviews (Hossler et al., 2019) and the use of multiple types of assessments, such as cognitive testing and structured and unstructured interviews. One might ask how one can best select the instruments and methods. Researchers have explored systematic strategies to combine various predictors, creating an optimal composite of scores coming from various assessments. For example, the Pareto-optimal weighting approach proposed by De Corte et al. (2007) generates a set of hiring solutions that balance the trade-off between diversity and job performance, creating a Pareto trade-off curve. This was deemed an effective way to achieve diversity improvements with minimal loss of validity (De Corte et al., 2008; Song et al., 2023). More recently, researchers have also examined the use of multi-penalty optimization to balance performance and diversity, combining regression of job performance on predictor scores, a regularization term, and group difference constraint (Rottman et al., 2023), which is very similar in nature to some of the in-processing strategies in ML discussed earlier.

4. To-Do's

4.1. Technical aspects

The measurement field has been contending with the fairness issue for decades. Over time, the scope of fairness notions and evaluation metrics has greatly expanded. That said, in light of the rapid development of fairness research in AI/ML, our field can benefit from keeping up with new developments there. This article, and similarly Hutchinson and Mitchell (2019), both focusing on fairness in supervised learning and group fairness, only serves as a starting point. Fairness research in AI/ML has a much broader scope than those. More needs to be researched in terms of the connections between psychometrics and AI/ML in the context of unsupervised learning, individual fairness, and other types of fairness notions. In particular, we believe the following topics may deserve special attention for measurement professionals from a technical perspective, as alluded to in the previous sections:

https://doi.org/10.1017/psy.2026.10110 Published online by Cambridge University Press

- Exploring fairness notion and evaluation outside of the group fairness regime, such as individual fairness, causal fairness, and temporal fairness.
- Exploring fairness notion and criteria in unsupervised learning, such as fair principal component analysis (fair PCA) and fair clustering (Chierichetti et al., 2017; Ghadiri et al., 2021).
- Understanding and quantifying bias propagation through the chain of actions and data, perhaps in a similar fashion to how we understand and quantify the uncertainty and measurement error propagation (Cheng & Yuan, 2010).
- Exploring bias mitigation strategies that are applicable in testing and are legally compliant.
- Developing methods that balance performance and fairness.
- The use of AI/ML to improve "responsiveness to the individual characteristics," for example, automating the development of culturally responsive assessments.

Meanwhile, the AI/ML community would benefit from greater engagement with the distinctive measurement perspectives on fairness, particularly those that extend beyond fairness defined in predictive or regression contexts. Most, if not all, the parallels of fairness criteria that have been drawn are limited to the last segment of Figure 2. Indeed, the well-cited Hutchinson and Mitchell (2019) surveyed 50 years of research on test fairness and introduced it to the AI/ML community, but it was limited to drawing parallels in the last segment of Figure 2. It considered test items to be analogous to model features, and "item responses analogous to specific activations of those features. Scoring a test is typically a simple linear model which produces a (possibly weighted) sum of the item scores." Hence, their conclusion was that test fairness and modern day ML fairness "trivially map on" to each other. There was a dramatic simplification of the scoring process in this characterization for sure. More importantly, it overlooked the multiple steps that exist before (and potentially after) the last segment of Figure 2. Because of that, it missed the entire point of construct validity, measurement invariance, and also missed the recent conversations on accessibility, accommodations, universal design, and cultural responsiveness, which are all an integral part of the big picture of test fairness. For a fictitious example that illustrates how measurement non-equivalency could translate to unfair selection in ML, please see the Appendix.

It is also important for the AI/ML community to recognize that measurement is an integral component of almost every AI/ML system. Hence, fairness issues in measurement exist in the entire chain of actions and data in an AI/ML system. For example, in the statistical criteria of separation and sufficiency, the idea of true status is critical. However, the true status may be a latent variable that is not directly observable or measurable, but needs to be estimated from a set of indicator variables. Besides, when an AI/ML system is trying to predict a certain outcome variable, even if the criterion is directly measurable, how to ensure it is measured accurately and free of bias? If the criterion variable or the predictor variables are not bias free, nothing in the AI/ML algorithm can remedy that; or it will come up with "artificial" solutions (e.g., switching labels through postprocessing) that does not address the root cause at all.

In addition, fairness in measurement places a great emphasis on adherence to the "intended construct" and responsiveness to the individual characteristics, which may or may not be related to a sensitive attribute. In the case of formula scoring for SAT, if risk tolerance was an intended construct of SAT, then individual difference or group difference on risk tolerance does not necessarily evoke fairness concerns. If it is an unintended construct, however, individual differences on this dimension will pose a threat to test fairness. This is true regardless of whether there is group difference on risk tolerance. Various groups could have the same level of risk aversion, but formula scoring would still be a problem from a test fairness perspective because it could "advantage or disadvantage some individuals" based on their individual difference on risk aversion. This is also true regardless of whether the individual difference on risk tolerance translates to differential prediction of outcome. Irrespective of how risk tolerance predicts first-year college GPA, its influence on SAT scores would cause fairness concerns by itself if it is an unintended construct. For example, measurement invariance studies could reveal that for one group, two factors underlie the test scores, including risk tolerance, whereas the risk tolerance factor

https://doi.org/10.1017/psy.2026.10110 Published online by Cambridge University Press

did not emerge for the other group. This violation of measurement equivalence is internal evidence of bias and will trigger fairness concerns.

The above example shows clearly that test fairness predicates on a well-defined construct that the test is purported to measure. This has strong implications for ML/AI applications. In some applications that rely on Big Data, a great number of variables are collected. There may not be much effort taken to understand what constructs they collectively measure or whether they measure what they are supposed to measure, as long as the prediction or classification meets the performance and fairness criteria discussed earlier. This lack of well-defined constructs also raises a major issue of explainability. Since explainable AI is a big, separate topic, we are not going to dive deep here, but it highlights the need of AI/ML research to consider setting out with clearly defined constructs or identifying constructs that underlie the vast number of variables involved in complex AI/ML systems (Lu et al., 2022).

Clearly, the AI/ML systems need distinctive measurement perspectives on fairness. At the same time, this places a responsibility on the measurement community to actively contribute to this dialogue—through outreach, education, and collaboration—by articulating how concepts such as construct representation, measurement equivalence, and validity inform fairness prior to prediction or decision-making.

One caveat that we would like to bring up here is that when we talk about testing in this article, we are referring to a systematic process of data collection using instruments that are developed and deployed to derive quantities that are "intended to support inferences about an attribute or property of an object, event, or phenomenon" (Ackerman et al., 2024). That said, we do not intend to preclude "unobtrusive assessment" of subjects' "digital footprints" (Behrens et al., 2019). For example, in our discussion of Figure 2, we explicitly said that response data can include the traditional item responses, as well as response times, process data, and texts, as long as they are collected systematically to support inferences on the intended constructs. However, our discussion is still limited to the development, deployment, and scoring of instruments following systematic procedures for intended purposes. Hence, information beyond the instruments, such as historical or contextual information (e.g., past courses taken, GPA, the school currently attending, etc.) is not considered in this discussion (except for demographic information that may be used for fairness evaluation), though those variables may play a large role in AI/ML applications and may have measurement bias and error that need to be carefully examined too.

4.2. Laws, policies, and guidelines

The legal implications involved in applying AI/ML to testing need to be very carefully considered by professionals in both fields. Some researchers in psychology had cautioned that the AI/ML field may not be fully aware of the legal implications when group-specific preprocessing is conducted in the testing context (Rottman et al., 2023), such as calculating z-scores for each group using group-specific means and standard deviations. The Civil Rights Act of 1964 outlawed disparate treatment of groups of individuals. Data preprocessing through group-specific transformation can be considered a violation. Furthermore, with the passage of the Civil Rights Act of 1991, "race-norming and any other means of changing or modifying employment related tests on the basis of race or ethnicity became illegal" (Greenlaw & Jensen, 1996). Race norming refers to a practice used in the context of standardized testing, particularly in employment and educational settings, where scores are adjusted or "normalized" based on the race or ethnicity of the individual being tested. This adjustment could involve setting different "cutoffs" for different racial groups. From this perspective, many postprocessing strategies may have legal consequences if applied in employment or educational testing.

Currently, there exist specific laws and guidelines related to using AI/ML algorithms in hiring. The first law regulating the use of AI in employment was issued by the City of New York (Local Law 144), which requires employers to conduct yearly third-party bias audits and post the results on the company website if they use automated employment decision tools to "substantially assist or replace" human judgment in hiring or promotion decisions (Morris, 2023). The U.S. Equal Employment Opportunity Commission (EEOC) also offered guidance to ensure compliance with federal civil rights laws, such

https://doi.org/10.1017/psy.2026.10110 Published online by Cambridge University Press

as the Americans with Disabilities Act and the Civil Rights Act, regarding the use of AI/ML and other emerging technologies in hiring and other employment decisions. The EEOC's ADA AI Guidance (The Americans with Disabilities Act and the Use of Software, Algorithms, and Artificial Intelligence to Assess Job Applicants and Employees) released in 2022 clearly indicates that employers are held responsible for ensuring fairness, for example, providing reasonable accommodations. In 2023, the EEOC issued the AI Disparate Impact Guidance (Assessing Adverse Impact in Software, Algorithms, and Artificial Intelligence Used in Employment Selection Procedures Under Title VII of the Civil Rights Act of 1964), providing employers with guidance regarding the application of federal non-discrimination laws when they use AI/ML algorithms or tools in making employment decisions. It specifies that decisions made by algorithms are subject to the same requirements as any other selection procedure, and employers are responsible for systems created by external vendors.

Beyond employment testing, the measurement field should devote concerted efforts on guiding and regulating AI/ML uses in educational testing and psychological assessments. There have been some efforts made within pockets of the community, for example, Duolingo English Test released Responsible AI Standards for public comment in 2023 and updated them in 2024, which included discussions on issues including fairness, transparency, privacy, and security (Burstein, 2023). This is a great starting point, which also calls for broader discussions and more organized efforts to encompass the diverse perspectives and use cases in measurement, for example, admission, selection, licensure, certification, and so on. It was recently reported that the State Bar of California's exam in February 2025 had at least a handful of the multiple-choice questions made with AI's assistance, which had struck the candidates "as bizarrely worded or legally unsound" (Mayorquín, 2025). Such controversy may surface more frequently as AI plays an increasingly more prominent role in the chain of actions in testing. As the revision of the 2014 Edition of the *Standards* is expected to be released in 2025, it will be an exciting time to see how its guidelines on fairness respond to the recent technological developments and challenges.

4.3. Ethics and values

As mentioned earlier in this commentary, any fairness definition in testing is inherently a value encoding (Linn, 1976), ethical position (Hunter & Schmidt, 1976), or a political and philosophical pursuit (Cronbach, 1976; Zwick & Dorans, 2016). The topic has always engendered varied positions, controversies, and contentious debates. The conversation is as much social, political, and philosophical as it is technical and statistical, as "outcomes should be evaluated not only in terms of the educational/occupational participants involved in selection but also their effects on society at large" (Lubinski, 2025). In fact, the issue of fairness was at the heart of much hostility in the society toward standardized testing. While many advocate for evidence-based and data-driven approaches, intense and longstanding resistance to these positions exist.

Even within Psychology, in different areas, the prevailing way to look at fairness varies. In personnel selection, disparate impact or adverse impact (as evaluated by the "four-fifths rule") is considered evidence of discrimination, and can be the basis to prohibit the use of AI/ML decision-making tools. In educational testing, however, many have diverted from such marginal "independence" criterion of fairness; instead, the field has embraced conditional independence criteria, such as DIF.

Such a struggle is not unique to the measurement community. In AI/ML research (and in economics, as mentioned earlier), obviously, a great variety of viewpoints and approaches exist in relation to fairness. Some researchers attempt to draw on traditional philosophical theories of fairness to guide the valuation of AI systems (Loi & Heitz, 2022). Leben (2025) offered a theory of algorithmic justice, inspired by the theory of justice proposed by American political philosopher John Rawls. As Linn (1976) powerfully argued (and echoed by the AI/ML researchers), the values encoded in the statistical criteria of fairness "should be made explicit" (Huchinson & Mitchell, 2019). It is not our intention to promote a certain test fairness definition or criterion, or a particular ethical proposition, but to raise awareness and stimulate conversations about values encoded in various definitions and criteria. We would also like to add that ethical issues related to testing have a much broader scope than what is covered in this article,

https://doi.org/10.1017/psy.2026.10110 Published online by Cambridge University Press

encompassing academic integrity, transparency, privacy, security (Burstein, 2023), and appropriate use or misuse of test scores (Benjamin Jr., 2009), and so on. By working together and mastering the language of the other discipline, professionals in measurement and AI/ML will not only enrich their technical and statistical toolbox to drive more evidence-based decisions, but also enrich each other's understanding in philosophical, moral, and ethical dimensions of the problem at hand.

5. Conclusions and discussion

The rapid expansion of AI/ML applications in educational, psychological, and employment testing presents the measurement community with both unprecedented opportunities and profound challenges. AIG through generative AI, AI-based fraud detection, automated scoring, and predictive analytics are already reshaping how assessments are created, administered, scored, and interpreted. Psychometricians need to make conscious efforts to engage in this transformative movement; otherwise, the field risks being marginalized in conversations it is uniquely qualified to shape. These efforts can be understood as operating at four increasingly expansive levels.

The first level focuses on understanding what AI/ML systems are doing within measurement contexts, ensuring that their use in assessment is responsible, valid, and fair. At this level, we place appropriate guardrails around our own products—assessment systems that increasingly rely on AI/ML components. The two frameworks each have their own unique aspects and advantages. Psychometric methods are essential for establishing measurement validity and equivalence, including evidence of measurement invariance and fair construct representation. These steps are prerequisite to any consequential use of assessment scores. AI/ML-based fairness metrics and bias mitigation techniques can be very useful at the decision level, such as classification or ranking. Accordingly, fairness evaluation could be staged: psychometric analyses to ensure valid measurement, followed by AI/ML methods to evaluate and mitigate inequities in downstream algorithmic decisions. This combined approach allows construct validity and algorithmic fairness to be addressed coherently, while accounting for ethical and legal considerations at each stage.

At the second level, we focus on adopting and adapting AI/ML approaches to fairness into measurement research. Johnson et al. (2022) provide a great example, which addresses the question of how to evaluate whether automated scoring algorithms are fair by integrating measurement principles with the widely adopted AI/ML fair prediction framework. Suk and Lyu (2025) adapted the idea of causal fairness to DIF to create causal DIF, which is defined as the difference in item functioning when individuals are assigned to one group versus a different group. A recent paper by Maeda and Lu (2025) tried to predict DIF using large language models (LLMs) and explainable AI, yet another example. Our immediate objective of this article is to promote more integration between psychometric and AI/ML approaches to fairness evaluation. We encourage measurement professionals to engage with how AI/ML researchers operationalize and assess fairness, and to consider how these frameworks can be integrated instead of treating them as alternatives. Some areas of research have been suggested in the To-do's section. For example, when the testing process is viewed as a series of data and actions or models, how can any of the bias mitigation approaches be meaningfully applied to educational and psychological testing (it has already been applied to employment testing in some sense)? What are the legal and/or ethical complications in doing so? Can the techniques of bias mitigation be adapted to measurement contexts beyond classification or prediction, for example, measurement invariance? What AI/ML techniques can we leverage to improve or innovate our fairness evaluation process?

The third level moves beyond AI/ML applications *within* measurement or testing to recognize that many AI/ML systems, regardless of domain, embed implicit measurement processes. These systems classify, score, rank, or infer latent attributes, often without acknowledging or validating their measurement assumptions. Engagement at this level means intentionally extending psychometric principles and validity frameworks to all AI/ML systems that have inherent measurement components. We would argue that any AI/ML systems that use "Big Data" inherently raise the question of what construct(s) the large number of variables are measuring and how they are measured. Core measurement concerns—

https://doi.org/10.1017/psy.2026.10110 Published online by Cambridge University Press

construct definition, reliability, validity, bias, and consequences of misuse and/or misinterpretation—naturally apply. Hence, measurement professionals ought to play a more prominent role in the development and evaluation of the AI/ML applications, regardless of whether the applications are in the field of assessment or testing or not. To do so, the psychometrics community should make a concerted effort in outreach, that is, raising awareness in the AI/ML community of the inherent measurement issues and the consequences of ignoring them. Specifically regarding fairness, we should promote recognition of the broader scope of fairness beyond relational equivalency with external variables (e.g., fair prediction or selection). Rather, it is critical to establish evidence of measurement equivalency beforehand—that is, fair definition and representation of the construct.

The fourth and most transformative level involves using psychometric thinking to actively shape how AI/ML systems are designed and evaluated. Here, measurement principles do not merely constrain or audit AI/ML after the fact; they influence how constructs are defined, how validity evidence is accumulated and organized, and how performance is interpreted from the outset. This level represents a shift from oversight to leadership—reframing how AI/ML developers think about model development, evaluation, and claims about capability. At this level, the expansion of AI/ML presents a unique opportunity for psychometricians to push the boundary of measurement research. For example, there has been research on identifying strong subsets of items for benchmarking LLM performance (tinyBenchmarks; Polo et al., 2024), similar to creating a psychometrically strong short form, and using adaptive testing to greatly reduce the number of items or tasks involved in LLM benchmarking (ATLAS; Li et al., 2025). Specific to the topic of fairness, some researchers have made innovative uses of psychometric approaches to fairness evaluation in AI/ML, such as Suk and Han (2024), which translates and adapts the well-known test fairness techniques such as DIF to evaluation of algorithmic fairness, and Jacobs and Wallach (2021), which discusses a general framework of fairness evaluation in AI/ML by treating fairness as a latent construct with various metrics as indicators.

In summary, psychometrics as a field has a great opportunity to make innovative uses of existing AI/ML methods, to apply measurement theory and frameworks to general AI/ML systems, and to inform, influence, and transform AI/ML research. At this point of time, one thing is clear: whether measurement practitioners *choose* to embrace AI/ML is no longer an open question. Adoption is already happening. The more pressing question is whether the measurement community is prepared and how to prepare for this transformation and even leverage it.

In 2023, the National Council for Measurement in Education (NCME) released a report on Foundational Competencies in Educational Measurement, which specified three domains of foundational competencies, including Domain 1: Communication and Collaboration, Domain 2: Technical, Statistical, and Computational Competencies, and Domain 3: Educational Measurement Competencies (Ackerman et al., 2024). In this report, AI and natural language processing appeared multiple times in discussions pertaining to Domain 2 and 3. In a commentary, Cheng et al. (2024) analyzed 80 job posts between July 1, 2023 and January 25, 2024 in psychometrics, with respect to their required competencies. They found that between 4% and 5% of the jobs explicitly require skills, such as ML, NLP, and AI. Though the percentage looked small, as the report indicated, AI and natural language models "seem poised to transform" the field (Ackerman et al., 2024).

However, many measurement practitioners have limited formal training in AI/ML methods. Without sufficient expertise, practitioners may feel compelled either to defer to technical systems they do not fully understand or to reject them wholesale, forfeiting the opportunity to influence their development. *This calls for systematic, coordinated investment in training and upskilling.* In recent years, we have seen numerous workshops and sessions at conferences, which are invaluable but still insufficient. The measurement community needs sustained pathways for professional development. This includes graduate curricula that integrate psychometrics, data science, and AI, continuing education opportunities for current practitioners, and cross-disciplinary training that brings together expertise in statistics, computer science, ethics, and law. The goal is not to turn psychometricians into engineers, but to equip them with the conceptual and technical literacy needed to evaluate, guide, and govern AI-based

https://doi.org/10.1017/psy.2026.10110 Published online by Cambridge University Press

assessment systems. Further, with proper training, psychometricians may exert influence far beyond our own field or assessment products, that is, to general AI/ML systems.

Supplementary material. The supplementary material for this article can be found at https://doi.org/10.1017/psy.2026.10110.

Data availability statement. Not applicable.

Acknowledgements. Parts of the article were presented at the 2024 International Meeting of Psychometric Society (IMPS) and the 2025 American Educational Research Association (AERA) annual meeting. The author would like to thank Drs. John Behrens, David Lubinski, and Scott Maxwell for their feedback on conceptualization and drafts of the article.

Funding statement. The research is supported by NSF OAC No. 2321054.

Competing interests. The author declares none.

References

Ackerman, T. A., Bandalos, D. L., Briggs, D. C., Everson, H. T., Ho, A. D., Lottridge, S. M., Madison, M. J., Sinharay, S., Rodriguez, M. C., Russell, M., von Davier, A. A., & Wind, S. A. (2024). Foundational competencies in educational measurement. *Educational Measurement: Issues and Practices*, *43*(3), 7–17. https://doi.org/10.1111/emip.12581

Anastasi, A. (1968). *Psychological testing*. (3rd ed.) Macmillan.

Bai, X., Wang, A., Sucholutsky, I., & Griffiths, T. L. (2025). Explicitly unbiased large language models still form biased associations. *PNAS*, *122*(8), e2416228122. https://doi.org/10.1073/pnas.2416228122

Baker, R. S., Esbenshade, L., Vitale, J., & Karumbaiah, S. (2023). Using demographic data as predictor variables: A questionable choice. *Journal of Educational Data Mining*. https://doi.org/10.5281/zenodo.7702628

Barocas, S., Hardt, M., & Narayanan, A. (2019). *Fairness and machine learning*. http://fairmlbook.org/

Baumann, J., Hannák, A., & Heitz, C. (2022). *Enforcing group fairness in algorithmic decision making: Utility maximization under sufficiency*. Preprint. https://arxiv.org/abs/2206.02237

Behrens, J., DiCerbo, K., & Foltz, P. (2019). Assessment of complex performances in digital environments. *The Annals of the American Academy of Political and Social Science*, *683*(1), 217–232. https://doi.org/10.1177/0002716219846850

Belzak, W., Burstein, J., & von Daver, A. A. (2025). *Evaluating fairness in AI-assisted remote proctoring*. In *Proceedings of machine learning research* (Vol. *273*, pp. 125–132).

Benjamin Jr., L. T . (2009). The birth of American intelligence testing. *Monitor on Psychology*, *40*(1), 20.

Bray, D. W., & Moses, J. L. (1972). Personnel selection. *Annual Review of Psychology*, 545–576. https://doi.org/10.1146/annurev.ps.23.020172.002553

Burgoyne, A. P., Mashburn, C. A., & Engle, R. W. (2021). Reducing adverse impact in high-stakes testing. *Intelligence*, *87*, 1–7. https://doi.org/10.1016/j.intell.2021.101561

Burstein, J. (2023). *The Duolingo English test responsible AI standards*. Updated March 29, 2024. https://go.duolingo.com/ResponsibleAI

Caton, S., & Haas, C. (2024). Fairness in machine learning: A survey. *ACM Computing Surveys*, *56*(7), 1–38. https://doi.org/10.1145/3616865

Cattell, R. B. (1940). A culture-free intelligence test. *Journal of Educational Psychology*, *31*, 161–179.

Cheng, Y., Pei, B., Filonczuk, A., & Le, A. T. (2024). Commentary: A data-driven analysis of recent job posts to evaluate the foundational competencies. *Educational Measurement: Issues and Practice*, *43*(3), 39–44. https://doi.org/10.1111/emip.12630

Cheng, Y., Pei, B., & Liu, C. (2025). Balancing Act: Early, Fair, and Accurate Identification of At-Risk Students. *Journal of Learning Analytics*, *12*(3), 47–65. https://doi.org/10.18608/jla.2025.8761

Cheng, Y., & Yuan, K. (2010). The impact of fallible item parameter estimates on latent trait recovery. *Psychometrika*, *75*(2), 280–291. https://doi.org/10.1007/s11336-009-9144-x

Chierichetti, F., Kumar, R., Lattanzi, S., & Vassilvitskii, S. (2017). Fair clustering through fairlets. In *Proceedings of the 31st international conference on neural information processing systems (Long Beach, California, USA) (NIPS'17)* (pp. 5036–5044). Curran Associates Inc.

Circi, R., Hicks, J., & Sikali, E. (2023). Automatic item generation: Foundations and machine learning-based approaches for assessments. *Frontiers in Education*, *8*, 858273. https://doi.org/10.3389/feduc.2023.858273

Cleary, T. A. (1968). Test bias: Prediction of grades of negro and white students in integrated colleges. *Journal of Educational Measurement*, *5*, 115–124.

Cole, N. S. (1973). Bias in selection. *Journal of Educational Measurement*, *10*(4), 237–255.

Cole, N. S. (1993). History and development of DIF. In P. W. Holland, & H. Wainer (Eds.), *Differential item functioning* (pp. 25–30). Lawrence Erlbaum.

Crocker, L., & Algina, J. (2008). *Introduction to classical and modern test theory*. Cengage Learning.

Cronbach, L. J. (1976). Equity in selection: Where psychometrics and political philosophy meet. *Journal of Educational Measurement*, *13*(1), 31–42.

Darlington, R. B. (1971). Another look at "cultural fairness." *Journal of Educational Measurement*, *8*(2), 71–82. https://doi.org/10.1111/j.1745-3984.1971.tb00908.x

De Corte, W., Lievens, F., & Sackett, P. R. (2007). Combining predictors to achieve optimal trade-offs between selection quality and adverse impact. *Journal of Applied Psychology*, *92*(5), 1380–1393. https://doi.org/10.1037/0021-9010.92.5.1380

De Corte, W., Lievens, F., & Sackett, P. R. (2008). Validity and adverse impact potential of predictor composite formation. *International Journal of Selection and Assessment*, *16*(3), 183–194. https://doi.org/10.1111/j.1468-2389.2008.00423.x

Dikly, S. (2006). An overview of automated scoring of essays. *The Journal of Technology, Learning, and Assessment*, *5*(1). https://ejournals.bc.edu/index.php/jtla/article/view/1640/1489

Drasgow, F. (1987). Study of the measurement bias of two standardized psychological tests. *Journal of Applied Psychology*, *72*(1), 19–29. https://doi.org/10.1037/0021-9010.72.1.19

Duan, Q., & Cheng, Y. (2024). Detecting differential item functioning using response time. *Educational and Psychological Measurement*, *85*(2). https://doi.org/10.1177/00131644241280400

Dwork, C., Hardt, M., Pitassi, T., Reingold, O., & Zemel, R. (2011). *Fairness through awareness*. Preprint. https://arxiv.org/abs/1104.3913

Einhorn, H. J., & Bass, A. R. (1971). Methodology considerations relevant to discrimination in employment testing. *Psychological Bulletin*, *75*, 261–269.

Ercikan, K., & McCaffrey, D. F. (2022). Optimizing implementation of artificial-intelligence-based automated scoring: An evidence centered design approach for designing assessments for AI-based scoring. *Journal of Educational Measurement*, *59*(3), 272–287. https://doi.org/10.1111/jedm.12332

Flaughter, R. L. (1974). The new definitions of test fairness in selection: Developments and implications. *Educational Researcher*, *3*(9), 13–16.

Fraenkel, A. (2020). Fairness and algorithmic decision making. https://afraenkel.github.io/fairness-book/content/05-parity-measures.html

Geisinger, K. F. (2022). The history of bias assessment in educational and psychological testing. In J. L. Jonson, & K. F. Geisinger (Eds.), *Fairness in educational and psychological testing: Examining theoretical, research, practice and policy implications of the 2014 standards*, 13–32. American Educational Research Association. https://doi.org/10.3102/9780935302967_1

Ghadiri, M., Samadi, S., & Vempala, S. (2021). Socially fair K-means clustering. In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency (virtual event, Canada) (FAccT '21)* (pp. 438–448). Association for Computing Machinery. https://doi.org/10.1145/3442188.3445906

Gierl, M. J., Lai, H., & Tanygin, V. (2021). *Advanced methods in automatic item generation*. Routledge. https://doi.org/10.4324/9781003025634

Gohar, U., & Cheng, L. (2023). *A survey on intersectional fairness in machine learning: Notions, mitigation, and challenges*. In *Proceedings of the thirty-second international joint conference on artificial intelligence (IJCAI-23)* (pp. 6619–6627). https://doi.org/10.24963/ijcai.2023/742

Goldberger, A. S. (1984). Reverse regression and salary discrimination. *The Journal of Human Resources*, *19*(3), 293–318.

Gopal, D. P., Chetty, U., O'Donnell, P., Gajria, C., & Balckadder-Weinstein, J. (2021). Implicit bias in healthcare: Clinical practice, research and decision making. *Future Healthcare Journal*, *8*(1), 40–48.

Green, C. A., & Ferber, M. A. (1984). Employment discrimination: An empirical test of forward versus reverse regression. *The Journal of Human Resources*, *19*(4), 557–569.

Greenlaw, P. S., & Jensen, S. S. (1996). Race-norming and the civil rights act of 1991. *Public Personnel Management*, *25*(1), 13–24. https://doi.org/10.1177/009102609602500102

Greenwald, A. G., McGhee, D. E., & Schwartz, J. L. K. (1998). Measuring individual differences in implicit cognition: The implicit association test. *Journal of Personality and Social Psychology*, *74*(6), 1464–1480. https://doi.org/10.1037//0022-3514.74.6.1464

Hardt, M., Price, E., & Srebro, N. (2016). Equality of opportunity in supervised learning. In *Proceedings of the 30th international conference on neural information processing systems (NeurIPS 2016)*.

Hasanzadeh, F., Josephson, C. B., Waters, G., Adedinsewo, D., Azizi, Z., & White, J. A. (2025). Bias recognition and mitigation strategies in artificial intelligence healthcare applications. *NPJ Digital Medicine*, *8*(1), 154. https://doi.org/10.1038/s41746-025-01503-7

Hossler, D., Chung, E., Kwon, J., Lucido, J., Bowman, N., & Bastedo, M. (2019). A study of the use of nonacademic factors in holistic undergraduate admissions reviews. *The Journal of Higher Education*, *90*(6), 833–859. https://doi.org/10.1080/00221546.2019.1574694

Humphreys, L. G. (1952). Individual differences. *Annual Review of Psychology*, *3*, 131–150.

Humphreys, L. G. (1986). An analysis and evaluation of test and item bias in the prediction context. *Journal of Applied Psychology*, *71*, 327–333.

Hunkenschroer, A. L., & Luetge, C. (2023). The impact of AI on recruitment and selection processes: Analysing the role of AI in automating and enhancing recruitment and selection procedures. *International Journal for Global Academic & Scientific Research*, *2*(2), 78–93. https://doi.org/10.55938/ijgasr.v2i2.50

https://doi.org/10.1017/psy.2026.10110 Published online by Cambridge University Press

Hunter, J. E., & Schmidt, F. L. (1976). Critical analysis of the statistical and ethical implications of various definitions of test bias. *Psychological Bulletin*, *83*(6), 1053–1071.

Hutchinson, B., & Mitchell, M. (2019). 50 years of test (unfairness): Lessons for machine learning. In *Proceedings of the conference on fairness, accountability, and transparency* (Vol. *2019*, pp. 49–58).

Jacobs, A. Z., & Wallach, H. (2021). Measurement and fairness. In *Conference on fairness, accountability, and transparency (FAccT '21), March 3–10, 2021, virtual event, Canada* (p. 11). ACM. https://doi.org/10.1145/3442188.3445901

Johnson, M. S. (2025). Responsible AI for measurement and learning: Principles and practices(Research Report No. RR-25-03). ETS. https://www.ets.org/Media/Research/pdf/RR-25-03.pdf

Johnson, M. S., Liu, X., & McCaffrey, D. F. (2022). Psychometric methods to evaluate measurement and algorithmic bias in automated scoring. *Journal of Educational Measurement*, *59*(3), 338–361.

Johnson, M. S., & McCaffrey, D. F. (2023). Evaluating fairness of automated scoring in educational measurement. In *Advancing natural language processing in educational assessment* (pp. 142–164).

Jung, C., Kannan, S., Lee, C., Pai, M. M., Roth, A., & Vohra, R. (2020). *Fair prediction with endogenous behavior*. Preprint. https://arxiv.org/pdf/2002.07147

Kamiran, F., & Calders, T. (2012). Data preprocessing techniques for classification without discrimination. *Knowledge and Information Systems*, *33*(1), 1.

Ke, Z., & Ng, V. (2019). Automated essay scoring: A survey of the state of the art. In *Proceedings of the twenty-eighth international joint conference on artificial intelligence (IJCAI-19) IJCAI* (Vol. *19*, pp. 6300–6308).

Kearns, M., Neel, S., Roth, A., & Wu, Z. S. (2018). *Preventing fairness gerrymandering: Auditing and learning for subgroup fairness*. In *Proceedings of the 35th international conference on machine learning (ICML 2018)* (pp. 2564–2572).

Kusner, M. J., Loftus, J. R., Russell, C., & Silva, R. (2017). *Counterfactual fairness*. In *Proceedings of the 31st international conference on neural information processing systems (NeurIPS 2017)* (pp. 4066–4076).

Lathrop, Q. N., & Cheng, Y. (2017). Item cloning variation and the impact on the parameters of response models. *Psychometrika*, *82*, 245–263. https://doi.org/10.1007/s11336-016-9513-1

Leben, D. (2025). *AI fairness: Designing equal opportunity algorithms*. The MIT Press.

Lee, Y.-H., & Zhang, J. (2010). Differential item functioning: Its consequences. ETS RR–10-01. https://files.eric.ed.gov/fulltext/EJ1111171.pdf

Li, P., Tang, X., Chen, S., Cheng, Y., Metoyer, R., Hua, T., & Chawla, N. V. (2025). *Adaptive testing for LLM evaluation: A psychometric alternative to static benchmarks*. Preprint. https://arxiv.org/abs/2511.04689

Linn, R. L. (1973). Fair test use in selection. *Review of Educational Research*, *43*(2), 139–161.

Linn, R. L. (1976). In search of fair selection procedures. *Journal of Educational Measurement*, *13*(1), 53–58.

Linn, R. L., & Werts, C. E. (1971). Considerations for studies of test bias. *Journal of Educational Measurement*, *8*, 1–4.

Liu, L. T., Dean, S., Rolf, E., Simchowitz, M., & Hardt, M. (2018). *Delayed impact of fair machine learning*. In *Proceedings of the 35th international conference on machine learning (PMLR 80)* (pp. 3150–3158).

Loi, M., & Heitz, C. (2022). Is calibration a fairness requirement? In *2022 ACM conference on fairness, accountability, and transparency (FAcctT'22), June 21–24, 2022, Seoul, Republic of Korea* (p. 9). ACM. https://doi.org/10.1145/3411764.3445195

Lord, F. M. (1977). A study of item bias, using item characteristic curve theory. In Y. H. Poortinga (Ed.), *Basic problems in cross-cultural psychology* (pp. 19–29). Swets & Zeitlinger.

Lord, F. M., Novick, M. R., & Birnbaum, A. (1968). *Statistical theories of mental test scores*. Addison-Wesley.

Lu, Y., Ober, T. M., Liu, C., & Cheng, Y. (2022). Application of neighborhood components analysis to process and survey data to predict student learning of statistics. In *Conference proceeding of 2022 international conference on advanced learning technologies (IEEE ICALT)*. https://doi.org/10.1109/ICALT55010.2022.00051

Lu, Y., Tong, L., & Cheng, Y. (2024). Advanced knowledge tracing for intelligent tutoring systems: Incorporating process data and curricula information via an attention-based framework. *Journal of Educational Data Mining*. https://jedm.educationaldatamining.org/index.php/JEDM/article/view/689

Lubinski, D. (2025). Education, intelligence, placement, and selection: A discussion of paradoxes and fairness. *Intelligence*, *108*. https://doi.org/10.1016/j.intell.2024.101881

Maeda, H., & Lu, Y. (2025). Finding words associated with DIF: Predicting differential item functioning using LLMs and explainable AI. *Journal of Educational Measurement*, *62*(4), 883–906. https://doi.org/10.1111/jedm.70017

Manhattan Review. (2022). Changes to the SAT in 2022. https://www.manhattanreview.com/sat-changes/#:~:text=The%202016%20SAT:%20More%20Higher,when%20preparing%20for%20the%20SAT

Mayorquín, O. (2025). On California's State bar exam, more questions than answers. *New York Times*, April 30, 2025. https://www.nytimes.com/2025/04/30/us/california-state-bar-exam.html

McDonald, R. P. (1999). *Test theory: A unified treatment*. Lawrence Erlbaum.

Mislevy, R. J., Behrens, J. T., Dicerbo, K. E., & Levy, R. (2012). Design and discovery in educational assessment: Evidence-Centered design. *Psychometrics, and Educational Data Mining. Journal of Educational Data Mining*, *4*(1), 11–48. https://doi.org/10.5281/zenodo.3554641

Morris, S. (2023). *Bias audits for artificial intelligence assessments*. In *Ideas in testing conference*.

Mukherjee, D., Yurochkin, Y., Banerjee, Y., & Sun, Y. (2020). *Two simple ways to learn individual fairness metrics from data*. In *International conference on machine learning* (pp. 7097–7107). PMLR.

https://doi.org/10.1017/psy.2026.10110 Published online by Cambridge University Press

Parker, C. P., & Dearing, R. L. (2009). Mitigating the diversity-validity dilemma in personnel selection: A review of fairness-enhancing interventions. *Journal of Organizational Behavior*, *30*(7), 1151–1166.

Peterson, N. S., & Novick, M. R. (1976). An evaluation of some models for culture-fair selection. *Journal of Educational Measurement*, *13*, 3–29.

Plečko, D., & Bareinboim, E. (2022). *Causal fairness analysis*. Preprint. https://arxiv.org/pdf/2207.11385

Plečko, D., & Bareinboim, E. (2024). Causal fairness analysis: A causal toolkit for fair machine learning. *Foundations and Trends® in Machine Learning*, *17*(3), 304–589. https://doi.org/10.1561/2200000106

Polo, M. F., Weber, L., Choshen, L., Sun, Y., Xu, G., & Yurochkin, M. (2024). tinyBenchmarks: Evaluating LLMs with fewer examples. In *Proceedings of the 41st international conference on machine learning* (Vol. *235*, pp. 34303–34326). PMLR.

Pyburn, K. M., Ployhart, R. E., & Kravitz, D. A. (2008). The diversity-validity dilemma: Overview and legal context. *Personnel Psychology*, *61*(1), 143–151. https://doi.org/10.1111/j.1744-6570.2008.00108.x

Reckase, M. D. (2022). *The psychometrics of standard setting connecting policy and test scores*. Chapman & Hall.

Rogers, H. J., & Swaminathan, H. (2016). Concepts and methods in research and differential item functioning of test items: Past, present and future. In C. S. Wells, & M. Faulkner-Bond (Eds.), *Educational measurement: From foundations to future* (pp. 126–142). Guilford.

Rottman, C., Gardner, C., Liff, J., Mondragon, N., & Zuloaga, L. (2023). New strategies for addressing the diversity–validity dilemma with big data. *Journal of Applied Psychology*, *108*(9), 1425–1444. https://doi.org/10.1037/apl0001084

Ruoss, A., Balunovic, M., Fischer, M., & Vechev, M. (2020). Learning certified individually fair representations. In *Advances in neural information processing systems* (Vol. *33*, pp. 7584–7596).

Sackett, P. R., Zhang, C., Berry, C. M., & Lievens, F. (2022). Revisiting meta-analytic estimates of validity in personnel selection: Addressing systematic overcorrection for restriction of range. *Journal of Applied Psychology*, *107*, 2040–2068.

Shermis, M. D. (2024). State-of-the-art automated essay scoring: Competition, results, and future directions from a United States demonstration. *Assessment Writing*, *20*, 53–76. https://doi.org/10.1016/j.asw.2013.04.001

Shermis, M. D., & Burstein, J. (2013). *Handbook of automated essay evaluation*. Routledge.

Shermis, M. D., & Burstein, J. C. (2003). *Automated essay scoring: A cross-disciplinary perspective*. Routledge.

Sinharay, S., & Johnson, M. (2023). Computation and accuracy evaluation of comparable scores on culturally responsive assessments. *Journal of Educational Measurement*, *61*(1), 5–46. https://doi.org/10.1111/jedm.12381

Sireci, S. G., Scarpati, S. E., & Li, S. (2005). Test accommodations for students with disabilities: An analysis of the interaction hypothesis. *Review of Educational Research*, *75*(4), 457–490.

Song, Q. C., Tang, C., Newman, D. A., & Wee, S. (2023). Adverse impact reduction and job performance optimization via pareto-optimal weighting: A shrinkage formula and regularization technique using machine learning. *Journal of Applied Psychology*, *108*(9), 1461–1485. https://doi.org/10.1037/apl0001085

Suk, Y., & Han, K. T. (2024). A psychometric framework for evaluating fairness in algorithmic decision making: Differential algorithmic functioning. *Journal of Educational and Behavioral Statistics*, *49*(2), 151–172. https://doi.org/10.3102/10769986231171711

Suk, Y., & Lyu, W. (2025). Rethinking item fairness using single world intervention graphs. https://osf.io/preprints/psyarxiv/bue62_v1

Swaminathan, H., & Rogers, J. (1990). Detecting differential item functioning using logistic regression procedures. *Journal of Educational Measurement*, *27*(4), 361–370.

Tang, Z., Zhang, J., & Zhang, K. (2023). *What-is and how-to for fairness in machine learning: A survey, reflection, and perspective*. Preprint. https://arxiv.org/pdf/2206.04101

Taylor, H. C., & Russell, J. T. (1939). The relationship of validity coefficients to the practical effectiveness of tests in selection: Discussion and tables. *Journal of Applied Psychology*, *23*(5), 565–578. https://doi.org/10.1037/h0057079

Thorndike, R. L. (1971). Concepts of culture-fairness. *Journal of Educational Measurement*, *8*, 63–70.

von Davier, M. (2018). Automated item generation with recurrent neural networks. *Psychometrika*, *83*(4), 847–857. https://doi.org/10.1007/s11336-018-9608-y

Weerts, H. J. P. (2021). *An introduction to algorithmic fairness*. Preprint. https://arxiv.org/abs/2105.05595

Zafar, M. B., Valera, I., Rodriguez, M. G., & Gummadi, K. P. (2017). Fairness beyond disparate treatment & disparate impact: Learning classification without disparate mistreatment. In *Proceedings of the 26th international conference on world wide web* (pp. 1171–1180).

Zhang, B. H., Lemonine, B., & Mitchell, M. (2018). Mitigating unwanted biases with adversarial learning. In *AAAI*. https://arxiv.org/pdf/1801.07593

Zhang, X., & Liu, M. (2020). *Fairness in learning-based sequential decision algorithms: A survey*. Preprint, arXiv:2001.04861.

Zwick, R., & Dorans, N. J. (2016). Philosophical perspectives on fairness in educational assessment. In *Fairness in educational assessment and measurement* (pp. 267–281). Routledge. https://doi.org/10.4324/9781315774527-17

Cite this article: Cheng, Y. (2026). Fairness Issues and Evaluation in Psychometrics and AI/ML: What Can We Learn from Each Field?. *Psychometrika*, 1–22. https://doi.org/10.1017/psy.2026.10110

https://doi.org/10.1017/psy.2026.10110 Published online by Cambridge University Press