AI and Ethics (2023) 3:845–868
https://doi.org/10.1007/s43681-022-00208-x

ORIGINAL RESEARCH

AI for hiring in context: a perspective on overcoming the unique challenges of employment research to mitigate disparate impact

Sara Kassir[1] · Lewis Baker[1] · Jackson Dolphin[1] · Frida Polli[1]

Received: 16 May 2022 / Accepted: 8 August 2022 / Published online: 22 September 2022

Abstract

Commentators interested in the societal implications of automated decision-making often overlook how decisions are made in the technology's absence. For example, the benefits of ML and big data are often summarized as efficiency, objectivity, and consistency; the risks, meanwhile, include replicating historical discrimination and oversimplifying nuanced situations. While this perspective tracks when technology replaces capricious human judgements, it is ill-suited to contexts where standardized assessments already exist. In spaces like employment selection, the relevant question is how an ML model compares to a manually built test. In this paper, we explain that since the Civil Rights Act, industrial and organizational (I/O) psychologists have struggled to produce assessments without disparate impact. By examining the utility of ML for conducting exploratory analyses, coupled with the back-testing capability offered by advances in data science, we explain modern technology's utility for hiring. We then empirically investigate a commercial hiring platform that applies several oft-cited benefits of ML to build custom job models for corporate employers. We focus on the disparate impact observed when models are deployed to evaluate real-world job candidates. Across a sample of 60 jobs built for 26 employers and used to evaluate approximately 400,00 candidates, minority-weighted impact ratios of 0.93 (Black–White), 0.97 (Hispanic–White), and 0.98 (Female–Male) are observed. We find similar results for candidates selecting disability-related accommodations within the platform versus unaccommodated users. We conclude by describing limitations, anticipating criticisms, and outlining further research.

Keywords Algorithmic fairness · Employment selection · AI hiring · Disparate impact · Test fairness · Employment discrimination

1 Introduction

It is widely acknowledged that progress on workforce diversity in the U.S. has been insufficient [1–3]. Various factors contribute to this stagnancy, but the procedures employers use to screen job candidates are clearly relevant to the problem. Since the passage of Title VII of the Civil Rights Act of 1964, employers have been prohibited from engaging in two forms of discrimination: disparate treatment (e.g., intentional exclusion of a person because of their identity) and disparate impact (e.g., unintentional disadvantage of a protected class via a facially neutral procedure) [4]. While the former has become less common over time, the latter remains very widespread [5], effectively barring diverse candidates from job opportunities.

Why have hiring procedures that disadvantage minority group members persisted since long after the civil rights era sought to eliminate them? In short, until fairly recently, there were very few equitable alternatives. Contrary to popular perception, the disparate impact provision of Title VII of the Civil Rights Act was intended as a form of pro-innovation regulation [6, 7]. In the 1960s, virtually, all hiring procedures were designed with white middle-class men in mind, and policymakers and testing experts recognized that new instruments needed to be created to facilitate equal access

✉ Sara Kassir
sara.kassir@pymetrics.com

Lewis Baker
lewisjbaker@gmail.com

Jackson Dolphin
jackson.dolphin@gmail.com

Frida Polli
frida.polli@pymetrics.com

[1] pymetrics, New York, USA

to opportunities [8].[1] Unfortunately, progress toward this goal was never fully realized, even as scientific research on human aptitudes expanded in adjacent fields [9]. Employers instead settled into compliance strategies that provided legal justification for use of biased selection tools [10].[2]

The emergence of so-called fairness-aware machine learning has brought renewed attention to quest for less-biased hiring procedures for the first time in decades.[3] Various authors have written about the potential for AI to overcome barriers to progress in standardized evaluations [11, 12], but these perspectives are limited in two important ways. First, contemporary investigations of AI are seldom framed in terms that resonate with the practical concerns of employers operating in the present day [13]. As technologists have rushed to comment on a future in which companies have already adopted novel hiring solutions en masse, little attention has been paid to the pros and cons of machine learning vis-à-vis incumbent selection methods. Exacerbating this disconnect is the tendency of machine learning researchers to frame empirical investigations of decision-making procedures as unprecedented when, in reality, employers have been conducting such analyses for over half a century. Second, data from employment procedures, whether algorithmic or paper-and-pencil, are seldom available to the public. Some AI experts have attempted to overcome this challenge with synthetic datasets [14, 15], but such tactics further dilute the research's immediate relevance for employers.

In this paper, we deviate from work that has positioned AI for hiring in a futuristic vacuum, instead evaluating the technology for its potential to overcome familiar challenges in employment selection.[4] Since the passage of Title VII of the Civil Rights Act, the industrial–organizational (I/O) psychology literature has struggled with the development of hiring procedures that simultaneously demonstrate validity and avoid significant racial impact [16]. The field's general consensus that such assessments are not achievable has dramatically influenced how employers approach compliance with anti-discrimination regulations [17]. Namely, if an organization is considering a novel hiring procedure designed with less disparate impact, the assumption is that it lacks the validity evidence necessary to respond to litigation. Because machine learning models can be trained with both aspects of Title VII compliance in mind, we argue that a critical advantage offered by the technology is an unprecedented capacity to disrupt this paradigm, known as the "diversity–validity dilemma."[5]

Our investigation further stands out from previous investigations of AI for hiring using data sourced directly from a commercial talent platform to empirically test our theory. The platform is an example of a system that uses fairness-aware machine learning, with the purpose being to provide large corporate employers with models to evaluate the alignment of candidates' "soft skills" to the needs of a particular role. Beyond simply explaining how these models are trained and tested prior to deployment, we examine the magnitude of disparate impact ratios observed when they are used to screen real-world job applicants. By benchmarking these results against the typical impact observed with traditional hiring assessments, such as cognitive ability tests, we

[1] According to a 1969 report to the U.S. Commission on Civil Rights, "The reason why minority groups do not perform on the average as well as members of the majority group on written tests are many and complex…In simplest form the problem can be stated as follows: most written examinations were developed by white middle class individuals to be administered by white middle class individuals." See Hanna, J., Freeman, F., Garcia, H., Hesburgh, T., Mitchell, M., Rankin, R., Glickstein, H.: For all the people by all the people: A Report on Equal Opportunity in State and Local Government Employment. U.S. Government Printing Office, Washington, D.C. (1969).

[2] According to legal scholar Linda Lye, "Unlike disparate treatment defendants, disparate impact defendants cannot defend by disavowing discriminatory intent. Instead, a disparate impact defendant must establish that a challenged practice is justified by a business necessity—i.e., that it constitutes a 'demonstrably…reasonable measure of job performance." See Lye, L.: Title VII's Tangled Tale: The Erosion and Confusion of Disparate Impact and the Business Necessity Defense. Berkeley Journal of Employment and Labor Law (1998). https://doi.org/10.15779/Z38NS76.

[3] Fairness-aware machine learning has also been referred to as "discrimination-aware classification." According to computer scientists Toon Calders and Sicco Verwer, the field "is motivated by the observation that often, training data contains unwanted dependencies between the attributes. Given a labeled dataset and a sensitive attribute; e.g., ethnicity, the goal of our research is to learn a classifier for predicting the class label that does not discriminate with respect to the sensitive attribute; e.g., for every ethnic group the probability of being in the positive class should roughly be the same." See Calders, T., Vervew, S.: Three naïve Bayes approaches for discrimination free classification. Data Mining and Knowledge Discovery (2010). https://doi.org/10.1007/s10618-010-0190-x.

[4] Regarding terminology, the discourse on AI in employment selection often refers to the technology simply as "AI." While different tools exist to facilitate the development of AI, or the ability of a computer system to simulate reasoning and render decisions, the most common technique used in HR applications is machine learning. Machine learning describes a process by which computers use mathematical and statistical models to infer patterns in data and make predictions without explicit programming.

[5] One I–O psychology publication summarizes the diversity–validity dilemma as follows: "Due to racioethnic and sex subgroup differences on predictor scores in many selection procedures, it is difficult for organizations to simultaneously maximize the validity of their selection procedures and hire a diverse workforce." See Pyburn, K. M., Jr., Ployhart, R. E., Kravitz, D. A.: The diversity-validity dilemma: Overview and legal context. Personnel Psychology (2008). https://doi.org/10.1111/j.1744-6570.2008.00108.x.

provide a practical interpretation of how machine learning may influence employment selection.

The paper is organized in three parts, with the first unpacking the status quo of the hiring technology industry, the second focusing on the opportunity for progress with AI, and the third presenting empirical results from a modern hiring platform utilized in high-volume screening contexts.[6] In part one, we describe how the limitations of science and technology in the post-civil rights era led employers to assume that they could only achieve compliance with Title VII through the use of highly traditional assessments and rigid validation methods. Over time, despite the law's clear mandate for employers to consider *both* the fairness and validity of selection procedures, a consensus emerged that only the latter could be practically addressed. In part two, we explain how machine learning allows for a shift away from employment selection's hyper-focus on traditional psychometric validity, which has perpetuated the so-called "diversity–validity dilemma." For example, since algorithms can be used to efficiently build hundreds of versions of an assessment, the technology facilitates identifying the one that simultaneously meets standards for group-level statistical parity (e.g., disparate impact) and cross-validated accuracy (e.g., concurrent criterion validity). In part three, we conduct an empirical investigation of the utility of ML for mitigating using data directly sourced from a commercial AI platform. We conclude by responding to anticipated concerns and providing directions for further research.

2 How Title VII of the Civil Rights Act tried (and failed) to disrupt ineffective employment testing

Lack of progress on disparate impact is often attributed to the failure of regulations to meaningfully influence employers' behavior. We offer a more-nuanced reality: while employers have always been highly motivated to avoid discrimination liability, their options for doing so were historically constrained by the state of assessment science. In this section, we provide some important historical background on the use of standardized selection tests and the rationale for their inclusion in the Civil Rights Act. We explain how the enactment of Title VII effectively offered two pathways for employers to comply with the law: develop new hiring procedures without disparate impact or justify the bias in legacy assessments with evidence of job-relevance. As validation of traditional tests took off and research on fairer alternatives stagnated, the marketplace of employment selection technology was frozen in the biased science of the mid-twentieth century. To demonstrate the persistence of this problem in the present day, we summarize the available empirical literature on the disparate impact of various common hiring assessments. This overview of the status quo frames our explanation of ML's advantages in the next section.

2.1 Some background on why employment selection needed regulation

Most people are unaware of the fact that standardized assessments are developed and sold as commercial products, but the industry dates back to the early twentieth century [18]. During World War I and II, the U.S. military experimented with the use of psychometric tests to classify recruits into appropriate roles [19, 20]. By the 1940s, American businesses had begun to develop formal personnel management functions, and supervisors "were no longer fully responsible for hiring workers, and they were not expected to have technical knowledge of all the jobs of their subordinates" [21]. Employment assessments became highly popular during this time, with some 3000 new products flooding the market between WWII and the civil rights movement; by the 1960s, some 80% of employers were using standardized selection tools [22].

While the quantity of products available in the testing industry grew rapidly in the mid-twentieth century, improvements in quality were not comparable. In 1963, one sociologist wrote that "virtually nothing" was known about the administration of most standardized assessments, let alone their efficacy in predicting important outcomes [23]. As employers adopted the new technology with "unchecked enthusiasm," test publishers faced minimal incentives to engage with proper scientific methodology [24]. In the years leading up to the Civil Rights Act, professional psychologists began articulating concerns about the integrity of common hiring practices, emphasizing that profiteering consultants were selling organizations inappropriate products [25]. Additionally, commentators expressed a general suspicion that many selection tools on the market were facilitating covert discrimination against Black candidates under the auspices of objectivity [26]. When hiring procedures ultimately became a legislative issue under Title VII of the Civil Rights Act of 1964, lawmakers were concerned both with the immorality of discrimination *and* with the economic inefficiency of needlessly biased tests [27].

[6] Throughout this paper, we focus on the use of hiring procedures by large employers in need of a standardized means of evaluating a large volume of job candidates. While it may be possible for the benefits of machine learning to be made relevant for small employers, such employers are significantly less likely to use pre-hire tests to "filter out" significant proportions of an applicant pool. Additionally, Title VII of the Civil Rights Act generally only applies to employers with at least 15 employees.

The historical context of the early assessment industry is important, because it disputes a common narrative regarding anti-discrimination regulations. Critics of disparate impact doctrine have long claimed that this statute impedes an organization's ability to make optimal hiring decisions [28, 29], but when Title VII was enacted, it was hardly the case that employers were forced to do away with carefully designed procedures. Instead, the concern for lawmakers was that most hiring tools were extremely inefficient and responsible for significant "economic waste" in the labor market, as evidenced by the high unemployment rate among Black jobseekers [30]. By repeatedly emphasizing that Title VII would in no way restrict use of "bona fide qualification tests," Congress was clear that the goal of disparate impact doctrine was to restrict only those hiring practices that were failing businesses and society [27]. Therefore, as legal scholar Steven Greenberger summarizes, disparate impact was intended "as a form of governmental regulation intended to enhance the nation's labor productivity by fostering the creation and implementation of personnel practices which will insure that business accurately evaluates its applicants and employees" [6].

In practical terms, once the Civil Rights Act acknowledged disparate impact as a form of workplace discrimination, the task for employers became considering whether their hiring procedures presented liability risks, in accordance with the three-pronged analysis outlined by the disparate impact provision of Title VII.[7] First, an employer should evaluate whether a hiring procedure disproportionately excludes members of a protected class. Second, they should ensure that the procedure is designed to measure job-relevant criteria. Third, the employer should consider whether the procedure could be replaced by an equally valid alternative with less disparate impact [4]. To reiterate, while the *threat* of discrimination litigation certainly created the impetus to ask these questions, the intent of regulators has never been to rely on lawsuits to effect change [31]. On the contrary, as summarized in *Albemarle Paper Co. v. Moody*, Title VII's preference for *voluntary compliance* instructs organizations "to self-examine and to self-evaluate their employment practices and to endeavor to eliminate, so far as possible, the last vestiges" of discrimination [32].

[7] As Trindel and co-authors summarize: "According to Title VII, in the event of disparate impact litigation, a three-pronged analysis will be applied to evaluate the lawfulness of an employment procedure.… From the employer's perspective then, voluntary compliance [with Title VII] is being able to answer these three questions in the affirmative." See Ref. [123].

2.2 Title VII: a strong focus on voluntary compliance and self-reflection from employers

Once the Civil Rights Act was enacted and employers were required to take a closer look at their hiring procedures, it became clear that "almost all contemporary employment testing" was characterized by the "ubiquitous design defect of failing to account for the nationwide impacts of segregation" [33]. Assessments were so likely to disproportionately filter out minority candidates that one commentator described finding evidence of disparate impact as "no more difficult than picking up stones from a gravel road" [24]. Adding to this insult was the fact that "employers, for the most part, had never tried to articulate their job performance goals in a systematic fashion, to develop selection devices carefully targeted to serve those goals, or to measure the success of such devices by validity studies" [34]. Since virtually all hiring procedure filtered out diverse candidates, the new regulatory regime offered two options for achieving voluntary compliance: identify selection tools with less disparate impact or establish evidence of job-relevance.

2.2.1 Compliance option 1: mitigating disparate impact

The years immediately following the passage of the Civil Rights Act saw testing professionals launched into optimistic investigations of "test fairness," spurred by the belief that scientific research could promote social equity. Prominent I/O psychologist Philip Ash described the mindset in a 1965 statement before the American Psychological Association: "Psychologists face, in the matter of civil rights, not a threat to their instruments but a challenge to their talents: to serve as resource people and advisers, to organize and administer programs, to make effective use of our tools, and to do research to clarify the problem" [35]. Efforts to develop "culture free" and "culture fair" tests were undertaken [36], and psychometricians "hastened to provide definitions of bias in terms of objective criteria, to develop rigorous and precise methods for studying bias, and to conduct empirical investigations of test bias" [37].

While assessment experts in the post-civil rights era demonstrated an openness to innovation, they were also limited by the state of science and the availability of data. Theoretically, strategies to reduce disparate impact could either focus on refining existing assessments or on measuring novel constructs [38]. Regarding the former option, researchers could hardly rely on historical datasets, because employers did not generally track candidates' protected class information until it was required by law [39]. Additionally, even after demographic data became readily available, employment contexts were often extremely homogenous, making group-level comparisons unviable. Proposals that may have seemed promising for reducing disparate impact, like using

non-verbal intelligence assays and ensuring work samples were properly scoped to focus on relevant tasks, were difficult to explore in depth [40, 41]. Regarding the latter option, new means of identifying and measuring psychological constructs were in their infancy in the 1960s and 1970s. The development of psychological instruments that did not rely on self-report would only occur after dramatic advancements in emerging fields, like cognitive psychology, cognitive neuroscience, and behavioral economics. Not until the dawn of the computer age would scientific progress in these disciplines translate to behavioral assessment technology [9].

The assessment industry's initial enthusiasm to find real-world solutions to combat employment discrimination did not last long enough to bring about meaningful change. By the mid-1970s, measurement experts were no longer exploring questions like "how can we better assess worker capabilities?" and were instead engaged in an esoteric debate about the statistical definitions of bias, led by prominent psychometricians like Cleary, Thorndike, and Darlington [42–44]. The shift was understandably frustrating for organizations awaiting improved assessments, as one commentator working for Educational Testing Services in the late 1970s noted: "I find disturbing…the behavior of many…social scientists who…retreat from all the controversies over testing and evaluation by retiring into cozy littler coteries where they write beautiful essays to one another that are so heavily laced with mathematical equations that it is a rare person…who can understand what they are talking about" [45]. As machine learning researchers, Hutchinson and Mitchell summarize: "The fascination with determining fairness ultimately died out as the work became less tied to the practical needs of society, politics and the law, and more tied to unambiguously identifying fairness" [46].

The erudite debates of measurement experts may have distracted from the goal of improving hiring tests, but academic scholarship did not singlehandedly derail progress on fair testing. Beginning in the 1970s, fairness-centered innovation was also deprioritized by a shift in demand from employers. The civil rights era had been motivated by a broad acknowledgement of the realities of systemic discrimination, but this consensus dissipated quickly. An alternative perspective emerged that appealed to the majority class, claiming that "racial subordination was largely past and…social inequalities, if any, reflected the cultural failings of minorities themselves" [47]. According to this narrative, efforts to remediate racial disparities were equivalent to "reverse discrimination" against deserving White people [48]. This narrative was bolstered by the re-emergence of hereditarian theories of intelligence by the likes of Arthur Jensen [49]. Unsurprisingly, as the popularity of this position increased, employers lost interest in voluntarily exploring tests with less disparate impact. In the words of legal scholar Ian López, "The window for fundamental change opened just slightly before blowing shut again in the face of a quickly gathering backlash" [47].

2.2.2 Compliance option 2: establishing validity evidence

The search for new hiring procedures may have faltered in the post-civil rights era, but efforts to validate existing assessments took off in full force. As one former president of the Society for Industrial and Organizational Psychology (SIOP) summarizes, "Before [the Civil Rights Act of 1964], I/O psychologists were interested in test validity, but their interest was a scientific one, not a legal one. The CRA began a tidal wave of work on test validation…Once we realized how important it was to be able to validate tests, the race was on to discover factors that led to lower than desired validities, and to validate tests more efficiently" [50]. Particularly after the landmark case *Albemarle Paper Co. v. Moody*, in which the Supreme Court interrogated the technical soundness of the defendant's business-necessity evidence, employers realized that "defenses based on apparent commonsense and rationality" were insufficient to shield against disparate impact liability [34].

Lawmakers may have hoped that increased validation research would improve the quality of selection tests, but in reality, the most pressing concern for employers was to avoid litigation. As sociologist Robin Stryker and co-authors observe, because the cumulative logic of legal precedent is "backward looking," this created a tension with the "forward looking" logic of scientific progress [17]. Psychologists working to arm employers with a legally sound validity defense unsurprisingly felt more secure in relying on large datasets and well-established psychometric constructs than in betting on the frontiers of research. Influential psychometrician Robert Guion lamented his field's stagnancy after serving as president of SIOP, noting that many researchers "[threw] out good hypotheses about predictors because of inadequate sample sizes" and continued to rely on problematic supervisor ratings as criteria simply because they were ubiquitous [51].

Over the years, psychometric pedagogy has dovetailed with legal incentives to make the dominance of established hiring assessments inevitable. Assessment experts Cole and Zieky explain: "Traditional validity studies focus on existing measures. They do not seek out alternative measures that may measure the same construct in different ways, nor do they seek out other constructs of likely utility. Furthermore, changes that do not add to validity are not sought after, even if such changes might allow some individuals to display strengths that might otherwise remain hidden" [52]. In light of these constraints, it is perhaps unsurprising that deference to *g*-based theories of intelligence continues to dominate employment research, with many contemporary I/O psychologists believing "most critical questions

regarding intelligence that are pertinent to personnel selection have been answered" [53]. Legal scholar Micheal Selmi summarizes the implications: "Most written examinations today continue to have substantial disparate impact; what has changed is that the tests are better constructed, in the sense that they are harder to challenge in court because they have been properly validated, but not better in the sense of being better predictors of performance" [7].

2.3 Today: understanding the status quo for hiring assessments

Today, several decades after Title VII was enacted, the two pieces of guidance employers receive regarding selection procedures remains very aligned with the conclusions of post-civil-rights-era psychologists: (1) the task of drastically reducing disparate impact and maintaining validity is largely futile and (2) the best evidence of job-relevance is existing psychometric literature. Acceptance of these tenets is so widespread in the testing industry that they are frequently combined and stylized as "the diversity–validity dilemma" [16], or the idea that selection tools with less impact necessarily "limit the capability of the workforce" [54]. Sales collateral from companies that develop standardized assessments further reinforce this position. For example, one 2017 marketing document from the cognitive test publisher Wonderlic reviews "the multitude of commonly used hiring tools…which typically exhibit disparate impact, along with how employers can justify their use" [55].

While industry rhetoric may suggest that disparate impact is an inevitable feature of effective hiring tests, it is important to note that the magnitude of the problem is rarely discussed in specific terms.[8] The assessment industry's reticence on disparate impact metrics was likely only further cemented by the passage of the Civil Rights Act of 1991, which placed the initial burden of proof for disparate impact litigation on plaintiffs. Stated plainly: without publicly available evidence of a procedure's biased outcomes, job candidates cannot even begin the process of filing a charge with regulators [56].

But the fact that most hiring procedures inhibit workforce diversity is not mere conjecture, since an extensive body of research exists on the topic. As organizational consultant Nancy Tippins states, "Regardless of where an employer stands on the topic of adverse impact, it must be measured" [54]. Over the course of a century, test publishers and employment experts have developed fairly standardized means of summarizing group disparities in assessment performance. I/O psychologists particularly reference the standardized mean difference *d* (or *Cohen's d*) as an index of demographic differences in average construct scores. As Sackett and Shen explain, "This is the majority mean minus the minority mean divided by the pooled within-group standard deviation.[9] This index expresses the group difference in standard deviation units, with zero indicating no difference, a positive value indicating a higher mean for the majority group, and a negative value indicating a higher mean for the minority group" [57].[10] In a 2008 review, Ployhart and Holtz provide an overview of *d* values for psychometric constructs common in employment [16], reproduced in Table 1.[11]

Table 1 Average standardized mean difference (*d*) for common selection instruments

Predictor	Black–White *d*	Hispanic–White *d*
General cognitive ability	0.99	0.83
Situational judgement (video)	0.31	0.41
Situational judgement (written)	0.40	0.37
Work sample	0.52	0.45
Job knowledge	0.48	0.47

To estimate the implications of a particular hiring test, the values summarized in Table 1 must be combined with information about the relevant selection context. In other words, how is the employer actually using the test to select candidates? The ultimate metric of interest is known as the *impact ratio* (IR), which is calculated as the selection rate for a given minority group divided by the selection rate of the majority group. Since 1978, EEOC guidance known as *the four–fifths rule* has suggested that an impact ratio below 0.8 may indicate disparate impact. In cases where average tests scores for the minority group and the majority group are

[8] As one psychometrician explains, "Publishers rarely tell a potential customer that their…tests are not suitable for the applicant sample expected…Already established group differences are usually buried deep within the many pages of the technical manual and may be explained briefly as being consistent with previous patterns of group differences in similar tests." See Jones, P.: Recruitment, Selection and Unconscious Bias. In Calvard, T., Cornish, T.: The Psychology of Ethnicity in Organisations. Bloomsbury Publishing, United Kingdom (2017).

[9] When separate standard deviations are not available for groups, the overall standard deviation is used.

[10] In layman's terms, *d* is the standard difference of two groups. For example, a *d* value of 1 indicates that group A outperforms group B by a full standard deviation. In the case of a standard IQ test, if group A has an average score of 100, group B would be expected to have an average score of 85.

[11] More recent research attempting to quantify the diversity–validity dilemma does not appear to be available in the I–O psychology literature. The same 2008 citation referenced here is also referenced in a 2021 article on AI-based personnel assessments written by prominent I/O psychologists Nancy Tippins, Frederick Oswald, and S. Morton McPhail. See Ref. [124].

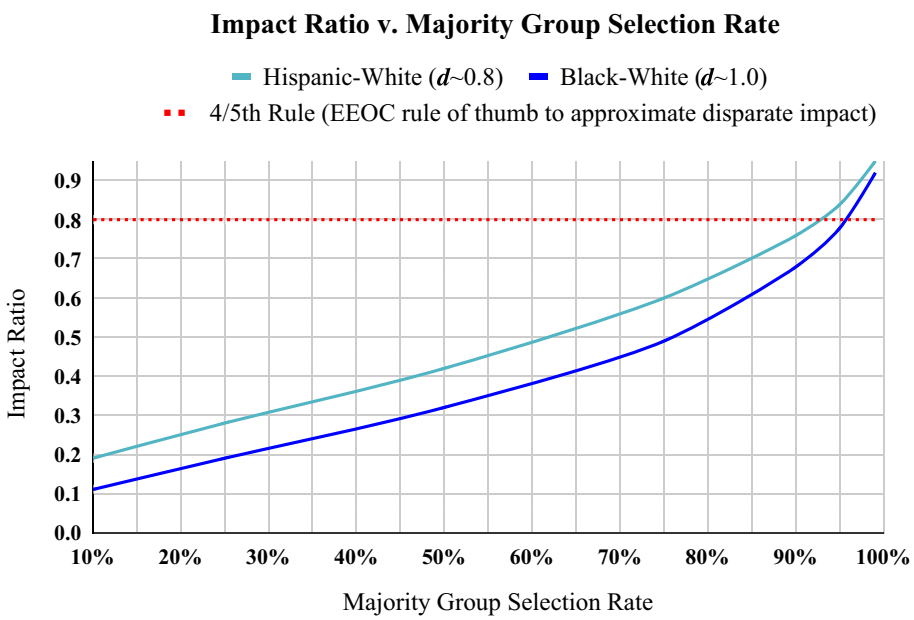

Fig. 1 Impact ratio observed for cognitive ability tests across varied levels of selectivity

very different (e.g., the *d* value is large), it is more likely that a test's IR will be relatively small. Group-level differences in scores on a hiring procedure are therefore a "precursor" of disparate impact [58], but it is worth noting that an exact impact ratio (IR) depends on how aggressively an employer uses a procedure to filter candidates. Generally, "the effects of group differences are greater as an organization becomes more selective (e.g., has a higher cutoff)" [59].

Consider the magnitude of disparate impact resulting when a cut-off score on a cognitive ability test is used to filter job candidates. While it is widely known that Black and Hispanic individuals tend to score lower on such assessments than White individuals, the practical implications of such gaps are seldom discussed. Using a methodology derived from I/O psychologists Sackett and Ellingson, Fig. 1 plots the impact ratio of cognitive ability tests as a function of the majority selection rate, or the proportion of White candidates who receive acceptable scores on the assessment [59]. Notably, neither the Black–White or Hispanic–White impact ratio falls above 0.8 unless the employer is already selecting over 90% of the White candidates in the pool. In other words, because the *d* value for cognitive tests is so large, it is virtually impossible for use of these procedures to not result in systematic exclusion of minorities.

Figure 2 alternatively focuses on the effects of tests with varying *d* values. The bar chart depicts the impact ratios observed when an assessment is used to screen in 50% of the White population. According to Sackett and Ellingson's methodology, the Black–White *d* value of 0.99 corresponds to only 16% of Black candidates being screened in, or an impact ratio of 0.32. For work sample tests, the *d* value of 0.52 means only 30% of Black candidates are screened in, or an impact ratio of 0.60. Strikingly, even when implemented in a relatively uncompetitive selection context, the vast majority of conventional hiring tests do not align with the EEOC's four–fifths rule.

While the above discussion demonstrates that the extent of racial impact in traditional assessments has held constant for decades, it is worth noting that I/O psychologists' attitudes toward this have changed over time. In the 1960s, assessments that consistently produced lower scores for Black Americans than their White peers were interrogated for possible "design defects" that were disadvantaging minority applicants [33]. But as testing experts tried and failed to build hiring procedures with less racial impact, questions began to emerge about whether avoidance of outcome disparities was actually necessary. By the 1970s, an increasing number of researchers were arguing that test bias should be framed solely in terms of under- or overprediction of performance for a given subgroup [42]. According to these commentators, analyses of fairness should instead look for evidence of *predictive bias*, also known as *differential validity*.[12]

In the trajectory of the hiring assessment industry, investigations of predictive bias have played an important role in sustaining I/O psychology's tolerance of disparate impact.[13]

[12] According to SIOP's professional guidelines, "Predictive bias is found when, for a given subgroup, systemic nonzero errors of prediction are made for members of the subgroup (Cleary, 1968; Humphreys, 1952)." See Ref. [71].

[13] As an example of how these investigations work, consider the SAT: if Black students consistently score lower than their White peers, but the groups go on to have similar college GPAs, the test is considered problematic; conversely, if Black students with low scores also go on to have low GPAs, the test is considered "fair." See Ref. [42].

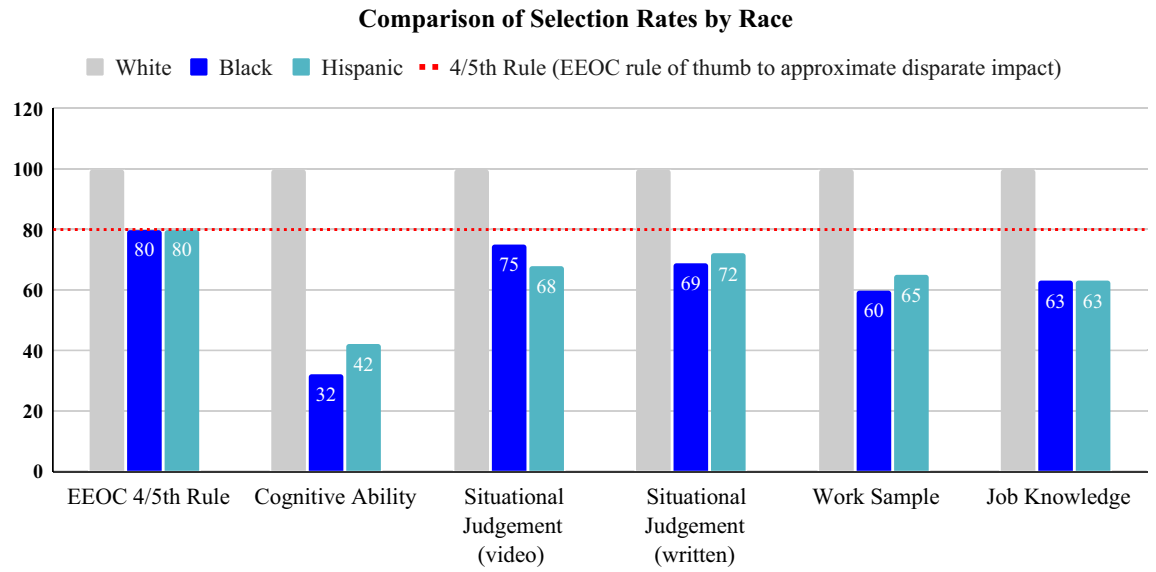

Fig. 2 Impact ratios observed for various tests (assume 50% white selected)

While this formulation of fairness as "lack of predictive bias" might seem reasonable at first glance, it requires a major assumption that is rarely true in selection contexts: the target variable cannot be biased against one of the subgroups. Guion called attention to this issue as early as 1965: "Most personnel research relies on ratings; the ratings of a potentially prejudiced supervisor can hardly be used in research on discrimination" [60]. In cases where a biased performance metric is used—such as if prejudiced white supervisors systematically assign lower ratings to black workers—"examinations of differential validity…may mask the presence of criterion bias and falsely indicate that no bias exists in either the test or criterion" [61]. Put differently, where I/O psychologists have used differential validity analyses to support the "fairness" of a test, historical employment data are treated as "ground truth," effectively characterizing the status quo as meritocratic.

3 Why a paradigm shift is needed in employment research and how machine learning can help

If the history of employment selection reveals anything, it is that traditional hiring procedures have been sustained by very premature conclusions about the nature of work, human potential, and job performance. In an effort to provide employers with actionable guidance on how to navigate anti-discrimination regulations, psychologists in the post-civil rights era were motivated to frame their research findings to date as universal truths. While it may have been important for testing experts to express resolute confidence in their validity studies in a world where all assessments had racial impact, the mindset has outlived its practical utility. Today, despite ongoing advancements in the evaluation of human aptitudes, some I/O psychologists remain steadfast in characterizing the field's knowledge of employment selection with a "mission accomplished mentality" [62]. In this section, we explain this mentality as a product of psychology's epistemological orientation, which has positioned the validity of established tests as truths that exist independently of available technology. In doing so, we set the stage to discuss machine learning as a tool for disrupting the dominance of inflexible theory-driven models for employment selection.

3.1 When science is an impediment to innovation: the epistemology of testing

How did testing experts, who were supposedly encouraged to develop new assessments under Title VII, come to such definitive interpretations of existing research? In a word: epistemology, or how a field distinguishes justified belief from opinion. For I/O psychologists,[14] validity studies represent "accurate, precise, value-free and context-independent knowledge about the relationship between predictors and

[14] According to SIOP's professional guidelines on test validation, a key value of "systematic selection systems" developed by trained experts is "a reduction in the reliance on subjective decisions and their biases" thus providing "improved prediction of desired work outcomes and the avoidance of bias in employment decisions." See Ref. [71].

criteria" [63]. The perspective reflects the field's foundations in positivism, which assumes that social science truths exist and can be discovered using the same research methods as the natural sciences [64].[15] Positivism has been a useful framework for the social sciences that enforced falsifiable experimentation and the formation and reformulation of theories based on observable phenomena [65]. One consequence of this position, however, is that oft-cited older findings are frequently intepreted in the literature as robustly established truths [66]. In recent years, as foundational tenets of employment research have been revisited by scientists equipped with modern statistical methods and contemporary data sources, a replication crisis in the field has been revealed [67].

The fallout of the replication crisis is perhaps best demonstrated by the dominant perspective among employment researchers that cognitive ability tests predict job performance in virtually any role. Support for this position first emerged in the 1970s when researchers began experimenting with meta-analysis and "artifact" corrections to aggregate validity studies. Prior to the adoption of these methods, experiments on the efficacy of IQ tests as hiring tools had produced notoriously inconsistent results. For many employment researchers, this situation was untenable: "It was well recognized that until and unless some form of generalization of validity results was possible, personnel psychology would lack legitimacy…[and] would also fall short of the goal of all scientific endeavor—the discovery of general laws" [68]. With the advent of meta-analysis, I/O psychologists argued that it was possible "to demonstrate generalizable results… that [had been] obscured, distorted, or unclear" due to noise in the primary studies [69]. In the 1980s, when John Hunter and Frank Schmidt applied these methods to studies of employment procedures dating back to the early twentieth century, they concluded that cognitive ability (as measured by commercial IQ tests) is universally the most valid predictor of job performance [70]. Meta-analysis still proves to be a useful tool for aggregating trends over time; however, as with many statistical abstractions, the interpretation of these results is not without subjectivity.

The most noteworthy aspect of this conviction is not *which* types of abilities have been favored in the scholarship but *how* these conclusions are articulated. In order for a test to have "generalizable" validity, researchers must maintain that variable results in different situations "can be attributed to sampling error variance, direct or incidental range restriction, and other statistical artifacts" [71]. In other words, there must be something *inherent* about the relationship between the construct measured by a test and job performance.[16] As Landers and Behrend summarize, "I/O psychology [is] unusual among social sciences in [its] insistence that theory should always precede facts" [72]. The problem, however, is that when psychologists are committed to proving the validity of a *particular* theory, there is a risk of "clarifying a finding that was never really there in the first place" [73].

The science of human performance evaluation is of two minds: one which seeks validation of new measures with previously validated experiments, and another which grapples with a history of overt prejudice. The foundation of psychometrics by Francis Galton and other social scientists held that the existing social order represented a "natural" hierarchy on the basis of innate ability. Helms summarizes the implications of this thinking: "For Galton, race, intelligence, and environment (i.e., socioeconomic status) were tautological. White men of eminence were inherently more intelligent than everyone else, as demonstrated by their accomplishments, which occurred because they were more intelligent than others. Galton equated intelligence with the quality of these men's sensory or psychological attributes" [74]. Thus, traditional psychometric research began from the position that the human aptitudes worth measuring were those exhibited by "superior" individuals in the majority group.[17] I/O psychologists (and many other social and behavioral sciences) have begun to reconcile with their troubling past, leading to reconsideration of fundamental assumptions of measurement and human aptitudes.

Far from providing an "objective" and "value-free" picture of the correlates of candidate success then, the reality of the traditional literature on employment selection is more accurately described as repeated investigations into a

[15] The influence of positivism on psychology is particularly noteworthy, because the epistemology has historically encouraged researchers to uncritically view prior research as ground truth. According to academic Russell Keat, for the positivist, "what matters is not the 'context of discovery', but that of 'justification'…[meaning] little attention is paid to the manner in which theories are arrived at—only to their testing." Technical studies that rely on mathematical logic to confirm or disconfirm theories are therefore "to some extent, self-perpetuating." See Keat, R.: Positivism, Natural, and Anti-Naturalism in the Social Sciences. Journal for the Theory of Social Behavior (1971). https://doi.org/10.1111/j.1468-5914.1971.tb00163.x.

[16] Belief in the existence of a "true" predictor–criterion relationship is further demonstrated by the fact that I/O psychologists "place a high value on and great deal of effort toward [test] measurement precision" but pay "much less attention…to sampling." See Fisher, G., Sandell, K.: Sampling in Industrial–Organizational Psychology Research: Now What?. Industrial and Organizational Psychology (2015). https://doi.org/10.1017/iop.2015.31.

[17] On a related note, O'Boyle explains the influence of Charles Darwin on early differential psychologists: "If intellectual capacity is an inherited characteristic and those who are brighter rise to higher levels of eminence in society, then it stands to reason that eminence should run in families. To explore this question, Galton examined lists of people who have achieved some degree of recognition." See O'Boyle, Cherie.: History of Psychology: A Cultural Perspective. Taylor & Francis, United Kingdom (2020).

limited set of theories about human aptitudes. In general, these investigations are framed for the purpose of explaining the observed social order, meaning questions that may potentially detract from this goal are dismissed. Helms, for example, observes that "measurement experts have been quite resistant to…seek explanations for racial-group differences in test performance in the groups' mean scores" [75]. One article from Jeffrey Cucina and co-authors demonstrate the myopic commitment to established evidence: "Given the vast empirical support for existing theories of intelligence…we do not believe that future attempts to create new models of intelligence will be fruitful. Instead, new intelligence literature [should] focus on bolstering the existing body of research and addressing common misconceptions among laypersons" [76]. At the same time, as Rabelo and Cortina observe, "Social groups that are underrepresented and/or marginalized are often excluded from organizational research, so existing theories and frameworks may not even apply to them" [77]. Thus, the epistemology of assessment science has been to validate previous assessments, with their inherent biases toward existing social hierarchies.

3.2 A persistent trope: three factors perpetuating the fairness–validity tradeoff

As previously mentioned in this paper, I/O psychology's epistemological orientation is significant, because it affects the types of studies that are conducted in employment contexts, which in turn affects how organizations consider hiring procedures. The goal of mitigating disparate impact is readily dismissed as not possible when researchers assume that the most efficacious assessments have already been discovered. I/O psychologists express this sentiment to employers using the framing of the "diversity–validity dilemma," or the notion that selection methods with less disparate impact necessarily sacrifice the procedure's predictivity [16]. While many commentators have demonstrated an eagerness to espouse the "dilemma" as an uncontestable fact [78, 79], others have countered that the theory is largely a product of how personnel selection has studied over the last century. In particular, the pretense of a tradeoff has been supported by inflated validity benchmarks for traditional tests, reliance on narrow psychological theories, and use of rudimentary statistical models.

3.2.1 Issue 1: validity coefficients distorted by publication bias

The first issue sustaining the "diversity–validity dilemma" in the discourse on employment selection is the fact that commonly cited validity coefficients for traditional tests are severely inflated. By referencing effect sizes that are virtually unheard of in the psychology literature (e.g., IQ scores account for 30%[18] of job performance [70]), I/O psychologists have constrained demand for assessments with less adverse impact. For example, organizational consultant Nancy Tippins emphasizes that employers who attempt to reduce racial disparities in selection rates "will also suffer the costs associated with a lower-ability group of employees" [54]. However, in recent years, psychologists across all subdisciplines have come to acknowledge the high proportion of false-positive findings in the literature due to the "methodological flexibility" granted researchers in terms of data collection, analysis, and reporting [73]. In I/O psychology specifically, Kepes and McDaniel contend that the literature has disseminated "an uncomfortably high rate of false-positive results…and other misestimated effect sizes" over the years [67].[19]

Individual validity studies in I/O psychology are likely to overestimate effect sizes, but meta-analytical reviews have generally been used to support the existence of the diversity–validity tradeoff [16]. According to Kepes and co-authors, "publication bias poses multiple threats to the accuracy of meta-analytically-derived effect sizes and related statistics," but "research in the organizational sciences tends to pay little attention to this issue" [80]. As Siegel and Eder explain, "the detrimental effect of publication bias is exacerbated in meta-analyses" which "serve as statistical and also conceptual 'false anchors' by biasing subsequent power-calculations and serving as authoritative sources due to their higher citation rate" [81]. The distortion is even further compounded by unique validity generalization procedures used by I/O psychologists, which rely on statistical corrections based on loose assumptions about range restriction and criterion reliability [82]. As one example of

[18] As a point of comparison, it is worth noting that this estimate of the validity of IQ tests is as large as the effect size of sleeping pills as a treatment for insomnia. See Finn, S., Kay, G., Kubiszyn, T., Reed, G.: Psychological testing and psychological assessment: A review of evidence and issues. American Psychologist (2001). https://doi.org/10.1037/0003-066X.56.2.128.

[19] The authors also argue that the incentives for engaging in questionable methodology are particularly strong for I/O psychologists, because such work is often tied to commercial interests: "Given the scientist-practitioner emphasis…many I/O psychologists employed in consulting, industry, and government also…strive for publications with supported hypotheses in part to serve the reputational interests of their organizations." See Ref. [67]. Additionally, since the top journals in the field virtually never accept manuscripts with null results, researchers are "unlikely to start new studies if they cannot predict the hypothesized outcomes" See: Kepes, S.,; Orhan, M., Bal, P.M., Van Rossenberg, Y.: Bringing I-O Psychology to the Public: But What if We Have Nothing to Say? PsyArXiv (2022). https://doi.org/10.31234/osf.io/rnq2e.

the severity of this inflation, a 2021 article from Sackett and co-authors replicates the meta-analysis that has been used to benchmark the validity of cognitive ability tests for the last 40 years. Upon reconsidering the corrections proposed by Hunter and Schmidt in the 1980s, the authors find that the reported effect size of 0.51 is overestimated by as much as 70% [83].

3.2.2 Issue 2: inflexible thinking on human ability

A second issue that has contributed to the perpetuation of the diversity–validity tradeoff in personnel selection is the field's resistance to engage with interdisciplinary research on human ability and its measurement. As Goldstein and co-authors explain, I/O psychology has embraced the psychometric perspective on intelligence to the virtual exclusion of other models [53]. At a high level, psychometric theories of intelligence are informed by "studying individual differences in test performance on cognitive tests." Alternative models include cognitive theories (which "study the process involved in intelligent performance"), cognitive–contextual theories (which "emphasize processes that demonstrate intelligence in a particular context, such as a cultural environment"), and biological theories (which "emphasize the relationship between intelligence, and the brain and its functions") [84]. Importantly, the most common perspective on intelligence in the testing community was largely established in the early twentieth century, and has a fairly rigid perspective of intelligence as referenced to a specific social standard. In contrast, newer research from fields like neuropsychology and cognitive science treat intelligence as modular, flexible, situationally dependent, and multifaceted [62]. Although most testing professionals would agree that intelligence is best measured through multiple angles of assessment, the reality is that their definition of intelligence is implicitly linked to a general intelligence that one either has or does not. This contributes to particular cultures and people being more likely to be "intelligent".

Newer theories of human ability can mitigate racial impact in measurement. For example, intelligence tests developed for vocational purposes within the public sector have historically emphasized the role of *crystallized intelligence* (i.e., *acculturated learning*) as a driver of overall performance [85]. In contrast, task-based assessments may focus on information-processing abilities, like fluid intelligence (i.e., *reasoning ability*), working memory, and attention control, which are less likely to vary with factors like educational attainment [86]. Traditional tests also tend to measure a fairly narrow set of aptitudes and therefore imply "deficit thinking" about individuals who do not demonstrate those aptitudes [87, 88]. Conversely, broader models of human capability, like Howard Gardner's theory of multiple intelligences, emphasize the fact that individuals have different intellectual strengths and weaknesses [89]. Overall, as West-Faulcon summarizes, "differences in racial group average scores are smaller on tests based on more complete theories of intelligence (multi-dimensional conceptions of intelligence) than on [traditional] tests" [90].[20]

While multi-dimensional models of human abilities can theoretically facilitate the development of less-biased hiring procedures, it is worth noting that certain methods have practically constrained this possibility. Historically, whenever testing researchers have investigated the validity of "specific abilities" (in contrast to a "general ability" factor, or *g*), they have done so using "incremental validity analysis." Kell and Lang explain: "Scores for an external criterion (e.g., job performance) are regressed first on scores on *g,* with scores for specific abilities entered in the second step of a hierarchical regression. If the specific ability scores account for little to no incremental variance in the criterion beyond [*g*], the specific aptitude theory is treated as being disconfirmed" [91]. In other words, researchers often base analyses of specific abilities on the assumption that a person's general intelligence *causes* any variance in other aptitude tests, making it virtually impossible to isolate the predictive utility of unique constructs.[21] Further exacerbating this problem is the fact that testing research has historically been constrained to small sample sizes, limiting the number of independent variables that can be included in an

[20] In the last decade, some assessment researchers have worked to quantify the reduction in subgroup differences seen with newer ability tests. For example, one study by I/O psychologist Elliott Larson found that Black–White subgroup differences on an "information processing" measure of intelligence yielded a *d* value of 0.41, which is significantly lower than the *d* value of 1.00 typically reported for traditional intelligence tests. See Larson, C.: A Meta-Analysis of Information Processing Measures of Intelligence, Performance, and Group Score Differences. City University of New York ProQuest Dissertations Publishing (2019). https://academicworks.cuny.edu/gc_etds/3040. Another study by I/O psychologist Jennifer Ferreter found that, for various neuropsychological intelligence batteries, the observed Black–White *d* values all fell below 0.35. See Ferreter, J.: Subgroup Differences and Predictive Ability of Psychometric and Neuropsychological Intelligence Measures. City University of New York ProQuest Dissertations Publishing (2010). https://www.proquest.com/docview/763492070/abstract/D211848897314BF0PQ/1?accountid=12768.

[21] As an alternative to "incremental validity analyses," "relative importance analyses" provide an opportunity to "make more precise and informed decisions concerning the usefulness of predictor variables." However, these methods are rare in employment research, presumably because they are assumed to conflict with theory-driven conceptions of intelligence in the psychometric literature. See LeBreton, J. M., Hargis, M. B., Griepentrog, B., Oswald, F. L., Ployhart, R. E. A multidimensional approach for evaluating variables in organizational research and practice. Personnel Psychology (2007). https://doi.org/10.1111/j.1744-6570.2007.00080.x; Kell, H.J., Lang, J.: Specific Abilities in the Workplace: More Important Than g? Journal of Intelligence (2017). https://dx.doi.org/10.3390%2Fjintelligence5020013.

experiment at one time [92]. As a result, it has seldom been feasible to consider how different narrow aptitudes contribute to job performance.

3.2.3 Issue 3: models that do not support multi-objective optimization

A final characteristic of historical methods that have sustained the diversity–validity dilemma is the tendency to build assessments with the sole goal of maximizing validity coefficients. In technical terms, employment researchers generally approach hiring as a *single-objective optimization* problem (focusing on validity) rather than a *multi-objective optimization* problem (focusing on validity *and* fairness simultaneously). As De Corte and co-authors explain, I/O psychologists are often asked by employers if the available predictors could be combined in a manner "that comes close to the optimal solution in terms of level of criterion performance achieved but does so with less adverse impact." However, according to the authors, "practitioners do not know how to respond to such a request other than by trial and error with various predictor weights" [93]. Stated differently, rather than framing avoidance of disparate impact as an explicit goal of a hiring procedure at the outset, assessment developers have largely viewed group-level disparities "as an afterthought or an unfortunate consequence of the organization's attempt to attain a single goal of maximizing job performance" [94]. In considering this significant methodological oversight, Hattrup and Roberts conclude that "when it comes to [adverse impact] versus validity, it is less a dilemma and more a question that has not been answered or perhaps a question that has not even been asked" [95].[22]

3.2.4 Summary of issues

According to the disparate impact theory of discrimination, the types of hiring procedures that employers relied upon in the mid-twentieth century served as an impediment to equality of opportunity. Policymakers in the 1960s had believed that employers who sought guidance from I/O psychologists would inevitably implement better and fairer tests, but this perspective unfortunately overestimated how much would first need to change about the study of employment selection. As one recent commentary from ones and several other I/O psychologists summarizes: "There is an absence of innovation and new ideas in the field… 'modern' measures…have been used in employee selection for over 90 years" [96]. Unfortunately, the issues that initially led to the development of tests with significant disparate impact—limited psychological theories, unsophisticated analytical tools, and skewed expectations about effect sizes—have remained intact as impediments to progress. Psychologist Jennifer Wiley poses a question that aptly summarizes the problem: "Experts generally solve problems in their field more effectively than novices because their well-structured, easily activated knowledge allows for efficient search of a solution space. But what happens when a problem requires a broad search for a solution?" In such cases, domain knowledge can "confine experts to an area of the search space in which the solution does not reside" [97].

3.3 A path forward: three benefits of machine learning for overcoming scientific convention

While the risk that hiring assessments can perpetuate discrimination has been clear for decades, public attention to the issue has hardly been consistent since the passage of the Civil Rights Act. However, "the explosion in the use of software in important sociotechnical systems has renewed focus on the study of the way technical constructs reflect policies, norms, and human values" [98]. With the advent of big data and machine learning, many commentators have reengaged with societal consequences of employment procedures, though through a framing that largely how candidates have been evaluated for jobs since the early twentieth century. Much of the disconnect seems to stem from the misconception that employers have generally relied exclusively on human decision-making to screen, meaning that algorithms might be introduced as a substitute for the capricious and time-consuming judgements of flawed recruiters. Under this assumption, the benefits of machine learning to the employment process are therefore described as "efficiently winnowing down the increasingly large volume of applications that employers now regularly receive" [99] and "providing more objective outcomes than humans" [100]. Meanwhile, the risks of such technology are described as "amplifying biases of the past" [98], "facilitating and obfuscating employment discrimination" [101], and "inflicting unintentional harms on individual human rights" [102].

[22] Hattrup and Roberts also point to the lack of multi-level modeling in I/O psychology as a major barrier to meaningful investigations of the diversity–validity dilemma: "Adverse impact and validity are outcomes that exist at different levels of analysis…In particular, *validity*, in the context of research on adverse impact, refers to the prediction of individual job performance criteria….In contrast, adverse impact, diversity, and compliance with the 'four-fifths rule,' are outcomes that exist at an aggregate level of analysis." See Ref. [95]. While multilevel modeling was first developed by statisticians in the 1980s, and theoretically offers a strategy for simultaneously considering the micro- and macro-level dimensions of personnel selection, "traditional methods of analysis in I/O psychology do not account for the hierarchical nature of such observations." See Gilbert, J., Shultz, K.: Multilevel modeling in industrial and personnel psychology. Current Psychology (1998). https://doi.org/10.1007/s12144-998-1012-9.

In sum, the current discourse on innovations in employment selection is flawed for two reasons: (1) it ignores the fact that large corporate employers seldom rely solely on human decision-making and often use traditional hiring tests,[23] and (2) it fails to consider the benefits of recent technological advancements in light of the domain-specific challenges of employment selection that have sustained the "diversity–validity dilemma." Regarding the former point, some recent progress has been made. For example, one report from Upturn observes that "many employers are using traditional selection procedures at scale—including troubling personality tests—even as they adopt new hiring technologies" [103]. In light of broad societal engagement with the persistence of systemic racism in the U.S. since the death of George Floyd in 2020, the American Psychological Association has also acknowledged that "psychologists created and promoted the widespread application of…instruments that have been used to disadvantage many communities of color" [104]. Regarding the latter point, however, progress is lacking. We respond to this gap in the literature by explaining the utility of machine learning and big data in overcoming the three above-mentioned barriers to innovation in employment selection.

3.3.1 Advantage 1: providing realistic estimates of assessment validity

The first reason machine learning is useful for overcoming the diversity–validity dilemma is the provision of more realistic estimates of the predictive validity of a selection procedure in a particular context, using larger sources of data and modern validation techniques, such as out-of-sample validation. This advantage is crucial for ensuring that inflated validity estimates (cited from decontextualized meta-analyses) are not used to justify the use of a hiring procedure with significant racial impact.[24] In addition to cross-validation, inflated effect sizes are also now tempered with the availability of larger sample sizes, which have historically been rare in standardized testing research. Generally, "the reason effect sizes in many domains have shrunk is that they were never truly big to begin with, and it's only now that researchers are routinely collecting enormous datasets that we are finally in a position to appreciate that fact" [73]. Additionally, researchers have argued that inflated effect sizes for traditional assessment predictors are likely exacerbated by the use of ordinary least-squares (OLS) regression models in psychology. Speer and co-authors explain that, while the method "is particularly susceptible to capitalizations on chance…most modern prediction methods have advantages over OLS that help guard against overfitting" [105].

3.3.2 Advantage 2: identifying novel, context-specific predictors of job performance

A second benefit of machine learning in the employment selection context is making it easier for assessment researchers to consider a much broader scope of relevant constructs. In recent years, as I/O psychologists have increasingly come to terms with the limitations of traditional aptitude testing in facilitating progress, various commentators have argued that the field needs to revisit its earlier experimentation with inductive research strategies [106].[25] In terms of the advantages over purely theory-driven methods, inductive methods can help researchers "identify nonobvious, subtle relationships between items and the criterion that other scoring techniques might miss" [107] and "combine many facets of personality for the sake of understanding the comprehensive profile of a person" [108]. As in other contexts,

[23] Exact estimates of the proportion of employers who use pre-employment assessments vary. According to one 2021 report from the Talent Board, 65% of surveyed employers use pre-employment assessments and selection tests. The survey was conducted on a sample of North American employers, with 53% of responding companies generating over $1B in annual revenue and 74% having over 2,500 total employees. See Talent Board: 2021 North American Candidate Experience Research Report. https://www.thetalentboard.org/benchmark-research/cande-research-reports/.

[24] Out of sample testing is the process of training a model on a subset of data and testing it on a withheld sample. Hold-out validation involves permanently withholding data, while cross-validation works by repeatedly iterating through multiple training and testing samples of one dataset. This practice greatly improves estimates of model performance and generalization. Yarokoni and Westfall explain that "although the explicit use of cross-validation to quantify generalization performance is largely absent from contemporary psychological science, the practice has deep roots in the field. See Ref. [73]. For example, according to one 1931 commentary from an educational psychologist, "when a coefficient of multiple correlation (R) is derived from a given set of data, its value is likely to be deceptively large…This is particularly significant, because ordinarily the practical employment of a regression equation involves its use with data other than those from which it was derived." See Larson, S.C.: The shrinkage of the coefficient of multiple correlation. Journal of Educational Psychology (1931). https://doi.org/10.1037/h0072400.

[25] Relevant methods from the I/O psychology literature include "empirical keying" and "profile matching." Regarding the former, Mumford and Owens summarize: "Essentially, empirical keys select and weight items on the basis of their ability to discriminate the members of the criterion group from some reference group. For instance, in personnel selection a key might be developed to discriminate high performers…from the members of…the current pool of job applicants." See Ref. [111]. "Profile matching," on the other hand, "assesses the fit between an ideal or standard profile and test-taker's score" without relying on a linear model. See Glaze, R.: The Efficacy of Profile Matching as a Means of Controlling for the Effects of Response Distortion on Personality Measures. Doctoral dissertation, Texas A&M University (2012). https://hdl.handle.net/1969.1/148261.

inductive research applied to employment selection involves "bottom-up, data-driven, and/or exploratory" analyses [109]. Because machine learning and big data effectively facilitate the automation of such methods, the technology allows for consideration of more-nuanced information sources, including those derived from modern psychological instruments.

It is worth emphasizing that, historically, inductive research in the employment context has been used to interpret inventories whose items did not have inherently "correct" answers, such as biographical application forms, personality inventories, and situational judgement tests [110, 111]. This aspect of inductive research is particularly relevant in considering the interpretation of the type of data that is often collected from modern aptitude assessments. While the traditional psychometric approach examines the "top-down" relationship of intelligence to complex cognitive tasks, the "bottom-up" nature of cognitive psychology focuses on identifying lower-order information processes, like memory, attention, and perception [112]. Psychologists have long theorized that detection of between-subject variation in these processes could inform research on specific aptitudes, or "cognitive styles" [113]. Such research began in the 1940s and 1950s, with psychologists like George Klein observing that individuals tended to vary in their perceptions of changes to visual stimuli; while some people are "sharpeners" who notice contrasts, others are "levelers" who focus on similarities [114]. However, unlike a unidirectional ability, the utility of being a leveler or a sharpener is context-dependent; "the adaptiveness…depends on the nature of the situation and the cognitive requirements of the task at hand" [115]. Given a particular job then, inductive research strategies are useful for identifying the aspects of cognitive style associated with job performance. For example, an AI classification model might use an incumbent sample's cognitive styles as a training data set, distinguished from a general applicant pool as a reference set [9].[26,27]

[26] Similar empirical investigations have been proposed with respect to high-dimensional personality inventories, whereby lower order items can be explored as discrete independent variables. See Mottus, R., Bates, T., Condono, D., Mroczek, D., Revelle, W.: Leveraging a more nuanced view of personality: Narrow characteristics predict and explain variance in life outcomes. PsyArXiv (2017). http://dx.doi.org/10.31234/osf.io/4q9gv.

[27] It is worth emphasizing that the utility of machine learning in exploring novel predictors of employment performance is contingent on the availability of meaningful performance criteria. Objective, well-defined measures of job performance are essential for determining which individuals should be included in the training set for a given model.

3.3.3 Advantage 3: optimizing models based on specified fairness and validity goals

A final way in which recent advances in data science can help testing experts overcome the diversity–validity dilemma is through the use of more sophisticated modeling techniques that allow for iterations on a model to be tested for their utility in achieving multiple objectives. Because the range of solutions that can be proposed by machine learning techniques is inherently limited to the data provided, the options "are plausible or credible, but are nonetheless not certain" [116]. The implied uncertainty is an important feature of any research based on available empirical evidence, because in contrast to theory-driven models, the "correct" answer is somewhat open to interpretation. With data-driven models, then "the goal is not omniscient certainty but contextual certainty," such that "one can properly say, 'on the basis of the available evidence, i.e., within the context of the factors so far discovered, the following is the proper conclusion to draw'" [117]. Models can account for any number of observed factors in deciding on the "proper" conclusion, including "side effects" or "potential consequences" of organizational interventions, such as disparate impact yielded by a hiring procedure [118, 119].

Data science techniques that specifically facilitate evaluation of multiple goals in the employment context include the use of (1) constrained optimization during the model-building process and (2) pre-testing that model for disparate impact prior to deployment on candidates. Regarding the former, Roth and Kearns summarize: "Machine learning already has a 'goal,' which is to maximize predictive accuracy…Instead of asking or a model that only minimizes error…we ask for the model that minimizes error *subject to the constraint* that it not violate [a] particular notion of fairness 'too much'" [120]. Fairness-constrained training processes, also sometimes known as "fairness-aware" or "discrimination-aware," have been applied to various contexts where minorities have experienced systemic discrimination, with the goal being to find the version of a model that disrupts the paradigm of a fairness–validity tradeoff [121, 122]. Regarding the latter, with the availability of Big Data, once a model has been trained assessment developers, can conduct controlled tests to compare the scores of protected classes and look for evidence of disparate impact. If disparities are identified, pre-testing signals developers to interrogate underlying assumptions and make adjustments *before* the model is ever used to score real candidates [123].

3.3.4 Summary of advantages

In sum, when the benefits of machine learning are aligned to the specific problems that have plagued employment science for decades, there is a strong case to be optimistic about the

potential for progress. The ability of machine learning and Big Data to facilitate data-driven model development, and to iterate and refine those models with respect to fairness and validity goals, could theoretically drive innovative in an unprecedented manner. As stated in Trindel et al.:

> "There is an air of prescience in Title VII's simultaneous inclusion of business objectives and fairness considerations. For many decades, employers and the courts have struggled to navigate these dual considerations, because it was not entirely clear how they could be evaluated in tandem. Today's technology actually makes voluntary compliance with the 1964 statute significantly easier; with advancements in data science, employers are empowered with the ability to simultaneously consider the efficacy and fairness of many variations of a hiring procedure, effectively adhering to all three prongs of adverse impact theory at once." [123]

While the technical possibility may exist for such progress, the outstanding question is whether or not the theory bears out in reality. As Tippins and co-authors state, "[An] advantage often expressed by vendors of technology enhanced assessments is a lack of or reduction in adverse impact and/or increases in criterion-related validity, as compared with traditional testing. However, the empirical evidence behind such claims is often unavailable, making relevant comparisons impossible" [124]. Legal scholar Pauline Kim similarly notes that "implementing the best available technical tools can never guarantee that algorithms are unbiased" so "examining the actual impact of algorithms on protected classes" is critical [125]. For these authors and others, the basic concern is whether so-called fairness-aware algorithms actually have the desired effect of mitigating disparate impact.

4 A real-world evaluation of machine learning's theoretical advantages

Thus far in this paper, we have pointed out that neither employers nor researchers have been sufficiently incentivized to implement better, fairer hiring procedures, as the architects of the Civil Rights Act had hoped. While these barriers to innovation may not be featured in popular discourse, the fact that hiring technology is flawed is also hardly a secret. Other researchers have therefore investigated AI through the lenses described above: use of machine learning out-of-sample testing as a local validation procedure [126], data-driven identification of job-relevant predictor–criterion relationships [127, 128], and simultaneous optimization of selection models for fairness and validity [129, 130]. However, the literature currently lacks a demonstration of how machine learning-based selection models perform when used to screen candidates. In this section, we source data directly from pymetrics, an algorithmic screening platform currently used by dozens of Fortune 1000 companies. Our central question is: What are the comparative advantages of candidate selection by machine learning? Namely, can it actually select candidates without disparate impact?

Previous work has described pymetrics' use of behavioral assays from the cognitive science literature to evaluate job candidates' "soft skills" [9]. Similar to the logic of "profile matching" in the I/O psychology literature, the platform contrasts soft-skill data from top-performing incumbents to that of a general reference set, ultimately building hiring models that rely on narrower cognitive, social, and personality measures than those measured by traditional assessments. pymetrics is also an example of a platform that builds models to simultaneously mitigate disparate impact and maximize classification accuracy on out-of-sample testing. This platform address the potential for a fairness/validity tradeoff through the constrained optimization approach described in Sect. 3.c.iii above. This method maximizes model performance, within the constrained that model fairness exceeds a minimum threshold—in this context, the EEOC 4/5ths success threshold between all groups. Notably, the platform only deploys models marked as fair and performant, which is itself a refutation of the fairness/validity dilemma. While the platform is not unique in claiming to use the so-called fairness-aware machine learning methods, the system's automated procedures for avoidance of disparate impact were the subject of a third-party audit in 2020 [131].

Over an 18-month period in 2019 and 2020, 26 North American employers commissioned the platform to build job models for an average of 2 roles per organization. Prior to the model-building process, a sample of at least 50 successful incumbents *in each role* completed the 25-min behavioral assessment to provide training data, or a "success profile" for the role.[28] In total, 60 models were built, each of which served as a custom standardized assessment.These models were built for employers in diverse industries (40% finance, 15% consumer goods, 12% consulting, 8% real estate, 8% HR, 7% logistics, and 10% other) to predict success in a variety of role types (23% office and administrative support,

[28] The definition of "successful incumbents" is unique to a particular employer and role. Depending on an organization's hiring priorities, pymetrics works closely with the employer to determine relevant performance criteria. Objective and quantifiable metrics, such as revenue data for a sales role, are preferred over subjective measures like supervisor ratings. pymetrics also recommends employers choose as diverse a set of top performers as possible to provide training data. For additional information about the platform's efforts to identify appropriate training samples with as little bias as possible, see Ref. [9].

Table 2 Summary of concurrent validity of machine learning models ($k = 60$)

Metric	Mean	SD
$r_{biserial}$	0.393	0.092
Accuracy	0.739	0.062
Recall	0.792	0.061
ROC AUC	0.749	0.052

28% business and financial operations, 15% computer and mathematical, 13% management, 13% sales and related, and 7% other). The concurrent criterion-related validity of each model was estimated prior to deployment, using *k*-fold cross-validation (Table 2). Cross-validation involved testing how successfully a model trained on 80% of the incumbent sample could identify the remaining 20%, repeated several times using different segments of the data.

Candidates who were screened by the platform took the same behavioral assessment as incumbents, and their results are translated into a fit score based on alignment to the incumbent profile. Candidates who received a percentile-ranked fit score above the 50th percentile were considered "recommended" by the model, and those who fell below the threshold were "not recommended." Upon deployment, these models were each used to screen applicant pools that included an average of 7000 job candidates. The platform, therefore, provided a total of approximately 400,000 predictions about future job success.

We considered four group-level fairness measures: (1) the impact ratio comparing Black and White candidates, (2) the impact ratio comparing Hispanic and White candidates, (3) the impact ratio comparing Female and Male candidates, and (4) the impact ratio comparing candidates who request disability-related accommodations in the hiring process and those who do not. We focused on these classes, because critics of AI hiring procedures have frequently cited bias against racial minorities, women, and individuals with disabilities as primary concerns [132–134]. Additionally, these demographic groups have all notably faced disproportionate levels of unemployment for decades [135–138]. Other classes certainly face discrimination in the labor market, but we were limited to studying those that are commonly reported in the hiring process and that consistently make up at least 2% of applicant pools, per the legal definition of disparate impact.

To determine the disparate impact of each model, we used demographic data that were voluntarily provided by candidates in an exit survey. The average response rate for the exit survey was 85% (SD = 13%). For the disability-related impact analysis, before beginning the assessment, candidates were offered accommodations for three disabilities: colorblindness, dyslexia, and ADHD. This allows standard testing accommodations to be applied automatically, as specified by best practices in the educational psychology literature.[29] For each model, an average of 6% of candidates (SD = 2%) elected to take an accommodated version of the assessment. According to EEOC regulations, adverse impact testing is inappropriate for groups that are less than 2% of the candidate pool. Using this rule, our analyses include 52 models for the Black–White comparison ($n = 392{,}448$); 56 models for the Hispanic–White comparison ($n = 404{,}952$); 60 models for the Female–Male comparison ($n = 412{,}219$), 17 models for colorblindness accommodations ($n = 59{,}604$); 18 models for dyslexia accommodations ($n = 189{,}557$); and 44 models for ADHD accommodations ($n = 347{,}959$).

Across the models with sufficient minority sizes (e.g., at least 2%), we found that the average minority-weighted IR for Black versus White candidates was 0.93 (SD = 0.10). For Hispanic versus White candidates, it was 0.97 (SD = 0.04). For Female versus Male candidates, it was 0.98 (SD = 0.05). For accommodations: colorblindness was 1.06 (SD = 0.14), dyslexia was 1.01 (SD = 0.10), and ADHD was 1.00 (SD = 0.10). In addition to these top-level results, we report impact ratios for subgroups of models by ONET Job Family (Tables 3 and 4). Overall, the analysis demonstrates that when a model is built to mitigate disparate impact, it does.

5 Discussion

Automated methods of building employment selection procedures have certainly proliferated in recent years, but much of the discourse about the advantages and risks has felt oddly detached from the nature of existing hiring methods. While it is certainly true that machine learning can make evaluating job candidates efficient and consistent, as many commentators have suggested, employers have been relying on standardized selection procedures precisely for this purpose for nearly a century. Additionally, while it is true that machine learning can introduce harms in the form of systematizing bias and obscuring discrimination, these effects are already pervasive due to widespread use of traditional assessments in many industries.

A more productive evaluation of the implications of machine learning for employment selection begins by identifying the problems that have sustained the suboptimal nature

[29] While accommodations may be desirable for additional mental disabilities, research on adjustments to testing procedures has generally been conducted within educational and developmental psychology. For this reason, a robust literature exists on how to accommodate these three disabilities as they are common, likely to be diagnosed in childhood, and likely to affect school performance. Further research is needed to determine how to best accommodate other mental disabilities in testing contexts. In cases where candidates feel they need an accommodation that the platform cannot provide, employers are legally required to provide alternative evaluation procedures.

Table 3 Disparate impact ratio results by ONET job family (race and gender)

Job family	Black				Hispanic				Female			
	Models	*n*	Mean	SD	Models	*n*	Mean	SD	Models	*n*	Mean	SD
Business and financial ops	13	89,918	0.88	0.05	16	103,539	0.96	0.04	17	103,891	1	0.06
Computer and mathematical	7	47,908	0.86	0.06	7	47,908	0.99	0.03	9	53,706	0.93	0.04
Management	8	22,168	0.91	0.12	8	22,168	0.93	0.06	8	22,168	0.98	0.08
Office and admin support	14	148,186	0.85	0.05	13	147,069	0.95	0.03	14	148,186	0.97	0.05
Sales and related	8	18,393	0.93	0.14	8	18,393	1	0.08	8	18,393	1	0.03
Other	4	65,875	1.04	0.03	4	65,875	1.02	0.02	4	65,875	1.02	0.01
All	52	392,448	**0.93**	0.1	56	404,952	**0.97**	0.04	60	412,219	**0.98**	0.05

Mean and standard deviation are weighted based on the number of relevant minority candidates scored. Job families with 2 models or less are excluded

Table 4 Disparate impact ratio results by ONET job family (disability accommodations)

Job family	Colorblindness				Dyslexia				ADD/ADHD			
	Models	*n*	Mean	SD	Models	*n*	Mean	SD	Models	*n*	Mean	SD
Business and financial ops	1	1359	N/A		3	21,928	1.08	0.07	11	98,405	0.96	0.12
Computer and mathematical	4	8837	0.97	0.15	1	1044	N/A		7	54,713	0.97	0.08
Management	4	17,874	1.02	0.12	1	1558	N/A		5	8974	1.25	0.29
Office and admin support	2	17,854	N/A		12	162,796	1	0.1	13	165,951	1.02	0.06
Sales and related	5	12,790	1.07	0.18	1	2231	N/A		7	19,083	0.95	0.13
Other	1	890	N/A		0	0	N/A	1		833	N/A	
All	17	59,604	**1.06**	0.14	18	189,557	**1.01**	0.1	44	347,959	**1.00**	0.10

Mean and standard deviation are weighted based on the number of relevant minority candidates scored. Job families with 2 models or less are excluded

of hiring procedures for the last 50 years. Regulation is often conceived of as the primary impediment to innovation, but the implicit goal of Title VIII of the Civil Rights Act was for employers to use science and evidence to produce better, fairer methods of screening job candidates, regardless of their demographic identity. However, meaningful progress toward this goal was never made, as demonstrated by the persistent trope in the I/O psychology literature that an "inevitable" tradeoff exists between fairness and validity in personnel selection. While regulation explicitly rejected the notion of such a tradeoff, the epistemology of testing experts effectively barred innovation in the field. Employers were therefore limited by the state of assessment technology in terms of their options for complying with anti-discrimination law.

In recent years, as the broader field of psychology has grappled with its replication crisis and society has reengaged with the systemic nature of discrimination, the flawed assumptions of employment researchers have come to light. The tendency of psychologists to view foundational research as established fact has been replaced with a recognition that theories of human hierarchy developed in the early twentieth century come with significant cultural baggage. Various commentators have recognized that the so-called fairness–validity dilemma in I/O psychology has largely been a product of how the field has framed its investigations. Specifically, by prioritizing established theories of human ability, developing assessments solely to maximize predictive validity, and overstating the benefits of legacy tools, employment researchers made it exceedingly difficult for hiring procedures to emerge outside the confines of the fairness–validity tradeoff.

Unfortunately, employment law is a context that has come to be defined by risk aversion and deference to precedent. Even though the shortcomings of traditional employment tests have been apparent since the latter half of the twentieth century, little has been done to deviate from the methods that employers first grew accustomed to shortly after the Civil Rights Act was passed. The practical constraints of the employment environment—small samples, litigation concerns, and subjective metrics, to name only a few—have dovetailed with the deductive epistemology of I/O psychologists to disincentivize exploratory research. From this perspective, it seems very unlikely that machine learning could further exacerbate the extent of disparate impact in the

modern hiring process, since it is highly entrenched under the status quo.

On the contrary, if the fundamental problems of employment research are related to reliance on simplistic theories and cumbersome manual research, an extensive literature suggests that machine learning can certainly help address these issues. Importantly, the point of the technology is not to replace subject matter experts or to argue for undiscerning use of irrelevant candidate information. Rather, the goal is to align the benefits of the technology to the specific challenges that have plagued employers and their advisors for decades. Where a large subset of psychometricians have spent the last century collecting evidence to identify universal selection methods, machine learning, and big data can help unpack more subtle and context-specific models. Where testing experts have sought to maximize the predictive validity coefficients yielded in scholarly articles, machine learning can optimize an assessment to align with multiple organizational practical objectives. Where employers have struggled with the robustness of manual validity studies and the technical challenges of proactively analyzing disparate impact, machine learning can facilitate efficient and iterative backtesting on large sample sizes.

The theoretical possibility of machine learning to improve on the hiring process has been discussed by various commentators. In some cases, authors have also conducted empirical investigations into the narrow benefits of the technology, often using synthetic data or very old samples. However, such piecemeal research is only so effective, particularly given the extent of public scrutiny regarding automation in high-stakes decision-making contexts. To provide a more grounded perspective on the real-world implications of machine learning, it is important to source data from a commercial AI platform that actually implements several of the prominently discussed advantages of the technology. pymetrics is useful for this exercise, because the platform utilizes the so-called fairness-aware training to build custom soft skill assessments for employers, relying on assays sourced from the cognitive and behavioral science literature to measure various job-relevant aptitudes. Overall, results indicate that the theorized benefits of machine learning for hiring bear out in practice: models that are simultaneously optimized for fairness and validity in the training process also result in much fairer outcomes when used to screen real candidates.

In considering the potential offered by machine learning to address the issue of disparate impact, it is worth anticipating a few likely reactions. At a high level, criticisms fall into two categories: concerns that technology will shield employers from disparate impact liability and complaints about the supposed "quality" of algorithmic decisions. Scholars interested in the former category emphasize the potential for AI systems to conceal disparities in outcomes from candidates [139], avoid scrutiny by claiming intellectual property rights [101], and provide statistical support for spurious predictors [124]. While such risks may be technically possible with machine learning technology, to focus on them suggests that progress toward the eradication of discrimination rests on disparate impact litigation. The reality is that, even absent automated screening tools, it is not especially difficult for employers to provide the necessary evidence to support the validity of a hiring practice that disadvantages a protected class. To put it differently: given that many employers perfected strategies for using biased hiring procedures lawfully decades ago, it seems unlikely that any additional shield offered by algorithms will meaningfully change their calculus.

The second category of criticisms about AI for hiring is made up of claims that fairness-aware algorithms are ineffective at identifying successful candidates. According to these researchers, hiring models that proactively mitigate group-level disparities necessarily trade accuracy for fairness [140], ignore characteristics that genuinely predict job performance [141], fail to account for inherent differences between groups [142] and hinder optimal prediction [143]. This literature further emphasizes that efforts to reduce the disparate impact can be highly unfair to qualified members of the majority group and warns that underqualified minority candidates will be harmed when they fail on the job. Such critiques are ostensibly directed at algorithmic hiring technology, but they are more precisely attacks on selection procedures that do not assume historical trends reflect objectivity. Consider Hardt and co-authors' argument that group-level statistical parity is "seriously flawed" as a fairness metric, because it "permits that we accept the qualified applicants in one demographic, but random individuals in another" [143]. Barocas and co-authors similarly describe algorithmic fairness constraints are "crude", because they "don't incorporate a notion of deservingness" for members of disadvantaged groups [141]. In both instances, the authors imply that "qualified" and "deservingness" are unambiguous concepts. Much like the deductive epistemology of I/O psychology, Birhane observes that such assumptions are common in computational research because of the field's roots in the Western "rationalist" worldview [144].

While machine learning skeptics who focus on the "effectiveness" of selection tools may be well-intentioned, their position largely mischaracterizes the nature of hiring, ignores the realities of historical discrimination, and misunderstands the purpose of anti-discrimination law. As a science, employment selection is characterized by irreducible unpredictability due to factors like unreliable performance criteria, small sample sizes, varied organizational environments, and ambiguous predictor constructs. For all of these reasons and others, testing professionals have repeatedly emphasized that the notion that a single hiring procedure

could perfectly summarize a candidate's "deservingness" is demonstrably false. Highhouse offers a useful perspective: "People seem to believe that, as long as the applicant is the right person for the job and the applicant is accurately assessed, success is certain…This represents a refusal…to recognize that many determinants of performance are not knowable at the time of hire…There is no such thing as perfect prediction in this domain" [145]. Demands that hiring procedures must be "accurate" largely ignore the very reason facially neutral hiring procedures are subject to anti-discrimination law: there are countless value judgements embedded in the definitions of how "qualified" someone is for a job. As Cooper and Abrams eloquently summarize, "Giving validity to an accuracy metric that has a dependency on past unfairness inherently advantages privileged groups; it is aligned with maintaining the status quo, as there is no way to splice out the past unfairness on which it is conditioned" [146].

Critics of fairness-aware AI who are worried about validity also fundamentally mischaracterize the purpose of disparate impact analysis in the context of employers' efforts to voluntarily comply with the Civil Rights Act. For example, Solon Barocas and co-authors suggest that, rather than attempting to mitigate disparities when racial impact is found, "the employer should strive to understand the reasons for this difference in success" and perhaps offer compensatory on-the-job training and feedback for rejected applicants [141]. Nothing in anti-discrimination law supports this position; the point of measuring subgroup disparities is not to gather evidence of "inherent" differences in abilities, but to encourage thoughtful consideration of a more equitable model. Rejection of all alternative models contributes to what D'Ignazio and Klein refer to as "deficit narratives," which "reduce a group or culture to its 'problems,' rather than portraying it with the strengths, creativity, and agency that people from those cultures possess." Put differently, wholesale rejection of efforts to mitigate bias can only be supported by ignoring the fact that diverse candidates have been excluded from the workforce *for reasons that have nothing to do with their job-relevant abilities.*

This research is not without its limitations. The perspective that machine learning can facilitate the development of hiring procedures with less disparate impact is certainly not meant to be interpreted in universal terms. Our investigation is confined to a single commercial platform that explicitly builds models specifically in accordance with EEOC guidelines.[30] Future research is obviously needed to understand the broader utility of machine learning in mitigating disparate impact in platforms that might assess different candidate characteristics using different instruments. It is also worth noting that our estimate of criterion-related validity relies on out-of-sample testing for classification accuracy. While reliance on concurrent validation is very common in employment research, future research should explore the predictive validity of machine learning models deployed in real-world contexts. Finally, while we use the existing literature to contrast the extent of disparate impact in traditional assessments vis-a-vis the AI-based approach, the sample of candidates was obviously not held constant across all procedures. Future research could investigate both the impact ratios and the concurrent validity observed when the same group of people is evaluated using different technology, though such research would require significant collaboration from an employer.

Many articles have provided recommendations for improved fairness in algorithmic decision-making, and many more are soon to follow. Although the purpose of the present article is to highlight the real successes that can be gained through fairness-aware machine learning, we can nonetheless summarize the innovations in this field as recommendations for future research: (1) identify causal and incidental sources of bias in data, (2) minimize use of source data that has historically led to bias, (3) measure algorithmic outcomes by protected group status in whichever application the algorithm is deployed, (4) minimize differences between groups during model training, within specific contexts such as geographies or campaigns, (5) test algorithms for fairness outcomes before deployment, and (6) monitor algorithmic fairness throughout the model lifecycle. While we are wary to codify these as more than suggestions, as legal, ethical, and regulatory requirements almost assuredly vary, the simple practice of monitoring outcomes and questioning historically biased data sources should hopefully prove broadly applicable.

6 Conclusion

Commentaries on AI for hiring that focus on the potential for technology to make fast and objective decisions are solving the wrong problem. The field of employment selection is not one that will ever be characterized by straightforward answers about the optimal way to align people to jobs they will go on to succeed in. This ambiguity of this context is not new; for much of the last century, researchers have attempted to grapple with it by insisting on simple models that could at least provide employers with *some* useful information across a very broad range of contexts. Machine learning and big data simply offer a different set of tools for navigating an ambiguous problem space by acknowledging the nuance,

[30] The data used to train these models is limited to psychometrically valid constructs borrowed from well-established literature, and predictors included are always validated by subject matter experts who have separately conducted a job analysis to understand the specific context in which the model will be used.

imprecision, and multidimensionality. One of the major benefits of the flexibility and exploration facilitated by this technology is the opportunity to proactively consider disparate impact in the development of hiring procedures. Given how persistent this problem has been throughout the history of employment testing, that benefit should be strongly weighed in any cost–benefit analysis of AI for hiring.

Author contributions All authors contributed to the study conception and design. Material preparation, data collection, and analysis were performed by LB, JD, and SK. The first draft of the manuscript was written by SK and all authors commented on previous versions of the manuscript. All authors read and approved the final manuscript.

Funding No funding was received for conducting this study.

Availability of data and materials The data that support the findings of this study are available from the corresponding author upon reasonable request.

Declarations

Conflict of interest The first and last author of this manuscript are current full-time employees of pymetrics, while the second and third are former full-time employees of pymetrics. The data reported in this manuscript were therefore supplied directly by pymetrics.

References

1. Fry, R., Kennedy, B., Funk, C.: STEM jobs see uneven progress in increasing gender, racial, and ethnic diversity. Pew Research Center. https://www.pewresearch.org/science/2021/04/01/stem-jobs-see-uneven-progress-in-increasing-gender-racial-and-ethnic-diversity/ (2021)
2. Stevens, P.: Companies are making bold promises about greater diversity, but there's a long way to go. CNBC. https://www.cnbc.com/2020/06/11/companies-are-making-bold-promises-about-greater-diversity-theres-a-long-way-to-go.html (2020).
3. Harrison, S.: Five years of tech diversity reports—and little progress. Wired. https://www.wired.com/story/five-years-tech-diversity-reports-little-progress/ (2019).
4. Title VII of the Civil Rights Act of 1964, 42 U.S.C. §2000e
5. Jones, K., Arena, D., Nittrouer, C., Alonso, N., Lindsey, A.: Subtle discrimination in the workplace: a vicious cycle. Ind. Organ. Psychol. (2017). https://doi.org/10.1017/iop.2016.91
6. Greenberger, S.: A productivity approach to disparate impact and the civil rights act of 1991. Or. L. Rev. (1993). https://via.library.depaul.edu/lawfacpubs/881
7. Selmi, M.: Was the disparate impact theory a mistake?, UCLA L. Rev. (2006). https://scholarship.law.gwu.edu/faculty_publications
8. Wallace, P.A.: Testing of minority group applicants for employment. U.S. Equal Employment Opportunity Commission: Office of Research and Reports, Washington, D.C. (1966)
9. Polli, F.E., Kassir, S., Dolphin, J., Baker, L., Gabrieli, J.: Research brief 18: cognitive science as a new people science for the future of work. MIT Task Force on the Work of the Future (2021). https://workofthefuture.mit.edu/research-post/cognitive-science-as-a-new-people-science-for-the-future-of-work/
10. Edelman, L.B.: Legal ambiguity and symbolic structures: organizational mediation of civil rights law. Am. J. Sociol. (1992). https://doi.org/10.1086/229939
11. Li, D., Raymond, L., Bergman, P.: Hiring as exploration. Natl. Bur. Econ. Res. (2020). https://doi.org/10.2139/ssrn.3630630
12. Kleinberg, J., Ludwig, J., Mullainathan, S., Sunstein, C.R.: Algorithms as discrimination detectors. Proc. Natl. Acad. Sci. USA (2020). https://doi.org/10.1073/pnas.1912790117
13. Gonzalez, M., Capman, J., Oswald, F., Theys, E., Tomczak, D.: "Where's the I-O?" Artificial intelligence and machine learning in talent management systems. Personnel Assess. Dec. (2019). https://doi.org/10.25035/pad.2019.03.005
14. Sharma, S., Zhang, Y., Rios Aliaga, J., Bouneffouf, D., Muthusamy, V., Varshney, K.: Data augmentation for discrimination prevention and bias disambiguation. In: AIES '20: Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society (2020). https://doi.org/10.1145/3375627.3375865
15. Pena, A., Serna, I., Morales, A., Fierrez, J.: Bias in multimodal AI: testbed for fair automatic recruitment. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (2020). https://doi.org/10.1109/CVPRW50498.2020.00022
16. Ployhart, R.E., Holtz, B.C.: The diversity–validity dilemma: Strategies for reducing racioethnic and sex subgroup differences and adverse impact in selection. Pers. Psychol. (2008). https://doi.org/10.1111/j.1744-6570.2008.00109.x
17. Stryker, R., Docka-Filipek, D., Wald, P.: Employment discrimination law and industrial psychology: social science as social authority and the co-production of law and science. Law Soc. Inq. (2012). https://doi.org/10.1111/j.1747-4469.2011.01277.x
18. Link, H.C.: Employment psychology: the application of scientific methods to the selection, training and grading of employees. Macmillan, United States (1919)
19. Yerkes, R. M.: Army Mental Tests. H. Holt, United States (1920)
20. Miller, N.E.: Psychological Research on Pilot Training. U.S. Government Printing Office, Washington, D.C. (1947)
21. Koppes Bryan, L.L., Vinchur, A.J.: A history of industrial and organizational psychology. In: Kozlowski, W.J. (ed.) The Oxford Handbook of Organizational Psychology, vol. 1, pp. 22–78. OUP USA, Spain (2012)
22. Rubin. R.B.: The Uniform Guidelines on Employee Selection Procedures: Compromises and Controversies. Catholic University Law Review (1979). https://scholarship.law.edu/lawreview/vol28/iss3/7
23. Goslin, D.A.: The Search for Ability: Standardized Testing in Social Perspective. Russell Sage Foundation, New York (1963)
24. Haney, C.: Employment tests and employment discrimination: a dissenting psychological opinion. Indus. Rel. LJ (1982). https://doi.org/10.15779/Z38QP7R

25. Seashore, H.: Ethical problems of the industrial psychologist. Pers. Psychol. (1949). https://doi.org/10.1111/j.1744-6570.1949.tb01674.x
26. Dobbin, F.: Inventing Equal Opportunity. Princeton University Press, United Kingdom (2009)
27. Legislative History of Titles VII and XI of Civil Rights Act of 1964. U.S. Government Printing Office, Washington, D.C. (1968)
28. Wax, A.L.: Disparate Impact Realism. William & Mary Law Review (2011). https://scholarship.law.wm.edu/wmlr/vol53/iss2/9v
29. Epstein, R.A.: Forbidden Grounds: The Case Against Employment Discrimination Laws. Harvard University Press, United Kingdom (1992)
30. Papers of John F. Kennedy. Presidential Papers. President's Office Files. Legislative Files. Special Message on Civil Rights, 28 February 1963. https://www.jfklibrary.org/asset-viewer/archives/JFKPOF/052/JFKPOF-052-016
31. Firefighters v. City of Cleveland, 478 U.S. 501 (1986)
32. Albemarle Paper Co. v. Moody, 422 U.S. 405 (1975); United States v. N. L. Industries, Inc., 479 F.2d 354, 379 (CA8 1973)
33. Henson, C., Title VII Works—That's Why We Don't Like It. Miami Race & Soc. Just. L. Rev. (2012). https://scholarship.law.missouri.edu/facpubs/517/
34. Bartholet, E.: Application of title VII to jobs in high places. Harv. Law Rev. (1982). https://doi.org/10.2307/1340570
35. Ash, P.: The implications of the Civil Rights Act of 1964 for psychological assessment in industry. Am. Psychol. (1966). https://doi.org/10.1037/h0023906
36. Gordon, E., Rubain, T.: Bias and alternatives in psychological testing. J. Negro Educ. (1980). https://doi.org/10.2307/2295093
37. Berk, R.: Handbook of Methods for Detecting Test Bias. Johns Hopkins University Press, United Kingdom (1982)
38. Wagner, R.K.: Intelligence, training, and employment. Am. Psychol. (1997). https://doi.org/10.1037/0003-066X.52.10.1059
39. Hartigan, J., Wigdor, A.: Validity Generalization Minority Issues, and the General Aptitude Test Battery. National Academy Press, Washington, D.C. (1989)
40. Bridgeman, B., Buttram, J.: Race differences on nonverbal analogy test performance as a function of verbal strategy training. J. Educ. Psychol. (1975). https://doi.org/10.1037/h0077030
41. Hollenbeck, J.R., Whitener, E.M.: Criterion-related validation for small sample contexts: an integrated approach to synthetic validity. J. Appl. Psychol. (1988). https://doi.org/10.1037/0021-9010.73.3.536
42. Cleary, T.A.: Test bias: prediction of grades of Negro and White Students in integrated colleges. J. Educ. Meas. (1968). https://doi.org/10.1111/j.1745-3984.1968.tb00613.x
43. Thorndike, R.L.: Concepts of culture-fairness. J. Educ. Meas. (1971). https://www.jstor.org/stable/1433959
44. Darlington, R.B.: Another look at "cultural fairness". J. Educ. Meas. (1971). https://www.jstor.org/stable/1433960
45. Proceedings of the ETS Invitational Conference. Educational Testing Service, United States (1976)
46. Hutchinson, B., Mitchell, M.: 50 years of test (un)fairness: lessons for machine learning. In: Proceedings of the Conference on Fairness, Accountability, and Transparency (2019). https://doi.org/10.1145/3287560.3287600
47. López, I.: A nation of minorities: race, ethnicity, and reactionary colorblindness. Stanford Law Rev. (2007). http://www.jstor.org/stable/40040347
48. McGinley, A.C.: The Emerging Cronyism Defense and Affirmative Action: A Critical Perspective on the Distinction Between Colorblind and Race-Conscious Decision Making Under Title VII. Scholarly Works (1997). https://scholars.law.unlv.edu/facpub/163
49. Jensen, A.: Race and mental ability. In: Symposium of the Institute of Biology on "Racial Variations in Man" (1974). https://eric.ed.gov/?id=ED114432
50. Bryan, L.K., Vinchur, A.J.: Industrial-organizational psychology. In: Freedheim, D.K., Weiner, I.B. (eds.) Handbook of Psychology, vol. 1: History of Psychology. Wiley, Germany (2012)
51. Guion, R.M.: Autobiography. https://www.siop.org/Portals/84/docs/Presidents/Guion_Robert_M.pdf (1984)
52. Cole, N.S., Zieky, M.J.: The new faces of fairness. J. Educ. Meas. (2001). https://doi.org/10.1111/j.1745-3984.2001.tb01132.x
53. Goldstein, H.W., Scherbaum, C.A., Yusko, K.P.: Revisiting *g*: intelligence, Adverse Impact, and Personnel selection. In: outtz, jl (ed.) adverse impact: implications for organizational Staffing and High Stakes Selection. Taylor & Francis, United Kingdom (2010)
54. Tippins, N.T.: Adverse impact in employee selection procedures from the perspective of an organizational consultant. In: Outtz, J.L. (ed.) Adverse Impact: Implications for Organizational Staffing and High Stakes Selection. Taylor & Francis, United Kingdom (2010)
55. Arnold, D.W.: How many of your hiring tools exhibit disparate impact? https://wonderlic.com/wp-content/uploads/2017/05/DisparateImpact.pdf (2017)
56. Jenoff, P. The case for candor: application of the self-critical analysis privilege to corporate diversity initiatives. Brook. L. Rev. (2011). https://brooklynworks.brooklaw.edu/blr/vol76/iss2/4
57. Sackett, P.R., Shen, W.: Subgroup differences on cognitive tests in contexts other than personnel selection. In: Outtz, J.L. (ed.) Adverse Impact: Implications for Organizational Staffing and High Stakes Selection. Routledge/Taylor & Francis Group, United Kingdom (2010)
58. Van Iddekinge, C.H., Morgeson, F.P., Schleicher, D.J., Campion, M.A.: Can I retake it? Exploring subgroup differences and criterion-related validity in promotion retesting. J. Appl. Psychol. (2011). https://doi.org/10.1037/a0023562
59. Sackett, P.R., Ellingson, J.E.: The effects of forming multipredictor composites on group differences and adverse impact. Personnel Psychol (1997). https://doi.org/10.1111/j.1744-6570.1997.tb00711.x
60. Guion, R.: Employment tests and discriminatory hiring. In: Employment Service Review. U.S. Government Printing Office, Washington, D.C. (1966)
61. Stauffer, J., Gaither, L.: A reanalysis of the predictive validity of the general aptitude test battery. Bus. Stud. J. (2011). https://www.abacademies.org/articles/bsjvol3si22011.pdf
62. Pl, H., Scherbaum, C., Goldstein, H., Ryan, R., Yusko, K.: I-O psychology and intelligence: a starting point established. Ind. Organ. Psychol. (2012). https://doi.org/10.1111/j.1754-9434.2012.01430.x
63. Schmidt, C.: Validity as an action concept in IO psychology. SA J. Ind. Psychol. (2006). https://doi.org/10.4102/sajhrm.v19i0.1477
64. Pietersen, H.J.: An epistemological view of industrial/organizational psychology: some perspectives and implications for future knowledge development. South Afr. J. Psychol. (1989). https://doi.org/10.1177/008124638901900206
65. Johnson, P., Cassell, C.: Epistemology and work psychology: new agendas. J. Occup. Organ. Psychol. (2001). https://doi.org/10.1348/096317901167280
66. Prilleltensky, I.: The Morals and Politics of Psychology. State University of New York Press, United States (1994)
67. Kepes, S., McDaniel, M.: How trustworthy is the scientific literature in industrial and organizational psychology? Ind. Organ. Psychol. (2013). https://doi.org/10.1111/iops.12045

68. Landy, F.: Validity generalization: then and now. In: Murphy, K. (ed.) Validity Generalization: A Critical Review. Lawrence Erlbaum Associates, United States (2003)
69. DeGeest, D., Schmidt, F.: The impact of research synthesis methods on industrial–organizational psychology: the road from pessimism to optimism about cumulative knowledge. Res. Synth. Methods (2011). https://doi.org/10.1002/jrsm.22
70. Schmidt, F., Hunter, J.: The validity and utility of selection methods in personnel psychology: practical and theoretical implications of 85 years of research findings. Psychol. Bull. (1998). https://doi.org/10.1037/0033-2909.124.2.262
71. Principles for the Validation and Use of Personnel Selection Procedures. American Psychological Association (2018). https://www.apa.org/ed/accreditation/about/policies/personnel-selection-procedures.pdf
72. Landers, R., Behrend, T.: When are models of technology in psychology most useful? Ind. Organ. Psychol. (2017). https://doi.org/10.1017/iop.2017.74
73. Yarkoni, T., Westfall, J.: Choosing prediction over explanation in psychology: lessons from machine learning. Perspect. Psychol. Sci. (2017). https://doi.org/10.1177/1745691617693393
74. Helms, J.: A legacy of eugenics underlies racial-group comparisons in intelligence testing. Ind. Organ. Psychol. (2012). https://doi.org/10.1111/j.1754-9434.2012.01426.x
75. Helms, J.: Fairness is not validity or cultural bias in racial-group assessment: a quantitative perspective. Am. Psychol. (2006). https://doi.org/10.1037/0003-066X.61.8.845
76. Cucina, J., Gast, I.F., Su, C.: g 2.0: factor analysis, filed findings, facts, fashionable topics, and future steps. Ind. Organ. Psychol. (2012). https://doi.org/10.1111/j.1754-9434.2012.01424
77. Rabelo, V., Cortina, L.: Intersectionality: infusing I-O psychology with feminist thought. In: Curtin, N., Cortina, L., Roberts, T., Duncan, E. (eds.) Feminist Perspectives on Building a Better Psychological Science of Gender. Springer International Publishing, Germany (2016)
78. McDaniel, M., Kepes, S., Banks, G.: The uniform guidelines are a detriment to the field of personnel selection. Ind. Organ. Psychol. (2011). https://doi.org/10.1111/j.1754-9434.2011.01382.x
79. Schmidt, F.: The role of general cognitive ability and job performance: why there cannot be a debate. Hum. Perform. (2002). https://doi.org/10.1080/08959285.2002.9668091
80. Kepes, S., Banks, G.C., McDaniel, M., Whetzel, D.L.: Publication bias in the organizational sciences. Organ. Res. Methods (2012). https://doi.org/10.1177/1094428112452760
81. Siegel, M., Eder, J.: Times are changing, bias isn't: a meta-meta-analysis on publication bias detection practices, prevalence rates, and predictors in industrial/organizational psychology. J. Appl. Psychol. (2021). https://doi.org/10.1037/apl0000991
82. Richardson, K., Norgate, S.: Does IQ really predict job performance? Appl. Dev. Sci. (2015). https://doi.org/10.1080/10888691.2014.983635
83. Sackett, P.R., Zhang, C., Berry, C.M., Lievens, F.: Revisiting meta-analytic estimates of validity in personnel selection: addressing systematic overcorrection for restriction of range. J. Appl. Psychol. (2021). https://doi.org/10.1037/apl0000994
84. Gardner, M.K.: Theories of intelligence. In: Bray, M., Kehle, T. (eds.) The Oxford handbook of school psychology. Oxford University Press, United Kingdom (2012)
85. Roberts, R., Goff, G.N., Anjoul, F., Kyllonen, P.C., Pallier, G., Stankov, L.: The armed services vocational aptitude battery (ASVAB): little more than acculturated learning (Gc)!? Learn. Individ. Differ. (2000). https://doi.org/10.1016/S1041-6080(00)00035-2
86. Burgoyne, A., Mashburn, C., Engle, R.: Reducing adverse impact in high-stakes testing. Intelligence (2021). https://doi.org/10.1016/j.intell.2021.101561
87. Ford, D.: Intelligence, Testing, and Cultural Diversity: Concerns, Cautions, and Considerations. The National Research Center on the Gifted and Talented (2004). https://files.eric.ed.gov/fulltext/ED505479.pdf
88. Ford, D., Harris, J.J., Tyson, C., Scott, M.F.T.: Beyond deficit thinking: providing access for gifted African American students. Roper Rev. (2002). https://doi.org/10.1080/02783190209554129
89. Gardner, H.: A multiplicity of intelligences. Sci. Am. **9**, 19–23 (1998)
90. West-Faulcon, K.: More intelligent design: testing measures of merit. University of Pennsylvania J. Constit. Law (2011). https://scholarship.law.upenn.edu/jcl/vol13/iss5/2
91. Lang, J., Kell, H.J.: General mental ability and specific abilities: their relative importance for extrinsic career success. J. Appl. Psychol. (2020). https://doi.org/10.1037/apl0000472
92. Oswald, F., Behrend, T., Putka, D., Sinar, E.: Big data in industrial-organizational psychology and human resource management: forward progress for organizational research and practice. Annu. Rev. Organ. Psych. Organ. Behav. (2020). https://doi.org/10.1146/annurev-orgpsych-032117-104553
93. De Corte, W., Lievens, F., Sackett, P.: Combining predictors to achieve optimal trade-offs between selection quality and adverse impact. J. Appl. Psychol. (2007). https://doi.org/10.1037/0021-9010.92.5.1380
94. Murphy, K.: How a broader definition of the criterion domain changes our thinking about adverse impact. In: Outtz, J.L. (ed.) Adverse Impact: Implications for Organizational Staffing and High Stakes Selection. Taylor & Francis, United Kingdom (2010)
95. Hattrup, K., Roberts, B.: What are the criteria for adverse impact? In: Outtz, J.L. (ed.) Adverse Impact: Implications for Organizational Staffing and High Stakes Selection. Taylor & Francis, United Kingdom (2010)
96. Ones, D., Kaiser, R., Chamorro-Premuzic, T., Svensson, C.: Has Industrial-Organizational Psychology Lost Its Way? Society for Industrial and Organizational Psychology. https://www.siop.org/Research-Publications/Items-of-Interest/ArtMID/19366/ArticleID/1550/Has-Industrial-Organizational-Psychology-Lost-Its-Way (2017).
97. Wiley, J.: Expertise as mental set: the effects of domain knowledge in creative problem solving. Mem. Cognit. (1998). https://doi.org/10.3758/bf03211392
98. Mulligan, D., Krooll, J., Kohli, N., Wong, R.: This Thing Called Fairness: Disciplinary Confusion Realizing a Value in Technology. In: Proceedings of the ACM on Human-Computer Interaction (2019). https://doi.org/10.1145/3359221
99. Girouard, M.: Big data, bigger risk: recognizing and managing the perils of using algorithms in recruiting and hiring. RAIL: J. Robot. Artif. Intell. Law. (2019) https://nilanjohnson.com/wp-content/uploads/2019/05/Girouard.pdf
100. Langenkamp, M., Costa, Al, Cheung, C.: Hiring fairly in the age of algorithms. Arxiv https://doi.org/10.48550/arXiv.2004.07132 (2020)
101. Ajunwa, I.: The auditing imperative for automated hiring. Harvard J. Law Tech. (2021). https://doi.org/10.2139/ssrn.3437631
102. Yam, J., Skorburg, J.: From human resources to human rights: Impact assessments for hiring algorithms. Ethics Inf. Technol. (2021). https://doi.org/10.1007/s10676-021-09599-7
103. Rieke, A., Janardan, U., Hsu, M., Duarte, N.: Analyzing the Hiring Technologies of Large Hourly Employers. Upturn (2021). https://www.upturn.org/work/essential-work/
104. APA Council of Representatives. Apology to People of Color for APA's Role in Promoting, Perpetuating, and Failing to Challenge Racism, Racial Discrimination, and Human Hierarchy in U.S. https://www.apa.org/about/policy/racism-apology (2021)
105. Speer, A.B., Christiansen, N.D., Robie, C., Jacobs, R.: Measurement specificity with modern methods: using dimensions, facets,

and items from personality assessments to predict performance. Appl. Psychol. (2021). https://doi.org/10.1037/apl0000618
106. Spector, P., Rogelberg, S., Ryan, A.M., Schmitt, N.: Moving the pendulum back to the middle: reflections on and introduction to the inductive research special issue of journal of business and psychology. J. Bus. Psychol. (2014). https://doi.org/10.1007/s10869-014-9372-7
107. Reiter-Palmon, R., Connelly, M.: Item selection counts: a comparison of empirical key and rational scale validities in theory-based and non-theory-based item pools. J. Appl. Psychol. (2000). https://doi.org/10.1037/0021-9010.85.1.143
108. Early, R.: The Use Of Personality Profiling As A Means To Assess Person-Organizational Fit To Inform Personnel Decisions. Wayne State University Dissertations (2016). https://digitalcommons.wayne.edu/oa_dissertations/1636
109. Woo, S.E., O'Boyle, E.H., Spector, P.: Best practices in developing, conducting, and evaluating inductive research. Hum. Reource Manag. Rev. (2017). https://doi.org/10.1016/j.hrmr.2016.08.004
110. Bergman, M., Henning, J., Drasgow, F., Juraska, S.: Scoring situational judgment tests: once you get the data, your troubles begins. Int. J. Sel. Assess. (2006). https://doi.org/10.1111/j.1468-2389.2006.00345.x
111. Mumford, M., Owens, W.: Methodology review: principles, procedures, and findings in the application of background data measures. Appl. Psychol. Meas. (1987). https://doi.org/10.1177/014662168701100101
112. Pretz, J., Sternberg, R.: Unifying the field: cognition and intelligence. In: Sternberg, R., Pretz, J. (eds.) Cognition and intelligence: Identifying the mechanisms of the mind. Cambridge University Press, United Kingdom (2005)
113. Kozhevnikov, M.: Cognitive styles in the context of modern psychology: toward an integrated framework of cognitive style. Psychol. Bull. (2007). https://doi.org/10.1037/0033-2909.133.3.464
114. Klein, G.: The personal world through perception. In: Blake, R., Ramsey, G. (eds.) Perception: An approach to personality. Ronald Press Company, New York (1951)
115. Messick, S.: Cognitive Styles in Educational Practice. Educational Testing Service. Princeton, New Jersey (1982)
116. Chater, N., Oaksford, M., Hahn, U., Heit, E.: Inductive logic and empirical psychology. In: Gabbay, D., Woods, J. (eds.) Inductive Logic. Elseiver Science, Netherlands (2011)
117. Locke, E.: The case for inductive theory building. J. Manag. (2007). https://doi.org/10.1177/0149206307307636
118. Watts, L., Gray, B., Medeiros, K.: Side effects associated with organizational interventions: a perspective. Ind. Organ. Psychol. (2020)
119. Messick, S.: Test validity: a matter of consequence. Soc. Indic. Res. (1998). https://doi.org/10.1023/A:1006964925094
120. Kearns, M., Roth, A.: The Ethical Algorithm: The Science of Socially Aware Algorithm Design. Oxford University Press, United Kingdom (2019)
121. Corbett-Davies, S., Pierson, E., Feller, A., Goel, S., Huq, A.: Algorithmic decision making and the cost of fairness. In: Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (2017). https://doi.org/10.1145/3097983.3098095
122. Kamiran, F., Calders, T.: Data preprocessing techniques for classification without discrimination. Knowl. Inf. Syst. (2012). https://doi.org/10.1007/s10115-011-0463-8
123. Trindel, K., Kassir, S., Bent, J.: Fairness in algorithmic employment selectin: how to comply with title VII. ABA J.Labor Empl. Law. **35**, 241–287 (2021)
124. Nancy, T., Oswald, F.L., McPhail, S.M.: Scientific, legal, and ethical concerns about ai-based personnel selection tools: a call to action. Personnel Assess. Dec. (2021). https://doi.org/10.25035/pad.2021.02.001
125. Kim, P.: Auditing Algorithms for Discrimination. University of Pennsylvania Law Review (2017). https://ssrn.com/abstract=3093982
126. Allen, K., Affourtit, M., Reddock, C.: The Machines Aren't Taking Over (yet): an empirical comparison of traditional, profiling, and machine learning approaches to criterion-related validity. Personnel Assess. Dec. (2020). https://doi.org/10.25035/pad.2020.03.002
127. Sajjadiani, S., Sojourner, A.J., Kammeyer-Mueller, J.D., Mykerezi, E.: Using machine learning to translate applicant work history into predictors of performance and turnover. J. Appl. Psychol. (2019). https://doi.org/10.1037/apl0000405
128. Stewart, R.D., Mottus, R., Seeboth, A., Soto, C.J., Johnson, W.: The finer details? The predictability of life outcomes from Big Five domains, facets, and nuances. J. Pers. (2021). https://doi.org/10.1111/jopy.12660
129. Zafar, M.B., Valera, I., Rodriguez, M.G., Gummadi, K.P.: Fairness constraints: mechanisms for fair classification. In: Proceedings of the 20th International Conference on Artificial Intelligence and Statistics. https://doi.org/10.48550/arXiv.1507.05259 (2015)
130. Geden, M., Andrews, J.: Fair and Interpretable Algorithmic Hiring using Evolutionary Many-Objective Optimization. Association for the Advancement of Artificial Intelligence (2021). https://www.aaai.org/AAAI21Papers/AISI-1438.GedenM.pdf
131. Wilson, C., Ghosh, A., Jiang, S., Mislove, A., Baker, L., Szary, J., Trindel, K., Polli, F.: Building and auditing fair algorithms: a case study in candidate screening. In: Proceedings of the ACM Conference on Fairness, Accountability, and Transparency (2021). https://doi.org/10.1145/3442188.3445928
132. Whittaker, M., Alper, M., Bennett, C., Hendren, S., Kaziunas, L., Mills, M., Ringel Morris, M., Rankin, J., Rogers, E., Salas, M., & Myers West, S. (2019). Disability, Bias, AI. AI Now Institute. https://ainowinstitute.org/disabilitybiasai-2019.pdf
133. Dastin, S.: Amazon scraps secret AI recruiting tool that showed bias against women. Reuters. https://www.reuters.com/article/us-amazon-com-jobs-automation-insight/amazon-scraps-secret-ai-recruiting-tool-that-showed-bias-against-women-idUSKCN1MK08G (2018)
134. Raghavan, M., Barocas, S. (2019) Challenges for mitigating bias in algorithmic hiring. Brookings. https://www.brookings.edu/research/challenges-for-mitigating-bias-in-algorithmic-hiring/.
135. Ajilore, O. On the Persistence of the Black–White Unemployment Gap. Center for American Progress. https://www.americanprogress.org/issues/economy/reports/2020/02/24/480743/persistence-black-white-unemployment-gap/ (2020)
136. Zamarripa, R. Closing Latino Labor Market Gap Requires Targeted Policies To End Discrimination. Center for American Progress. https://www.americanprogress.org/issues/economy/reports/2020/10/21/491619/closing-latino-labor-market-gap-requires-targeted-policies-end-discrimination/ (2020)
137. Boesch, D., Phadke, S. When Women Lose All the Jobs: Essential Actions for a Gender-Equitable Recovery. Center for American Progress. https://www.americanprogress.org/article/women-lose-jobs-essential-actions-gender-equitable-recovery/ (2021)
138. Ross, M., Bateman, N. Only four out of ten working-age adults with disabilities are employed. Brookings. https://www.brookings.edu/blog/the-avenue/2018/07/25/only-four-out-of-ten-working-age-adults-with-disabilities-are-employed/ (2018)
139. Ajunwa, I.: The "black box" at work. Big Data Soc. (2020). https://doi.org/10.1177/2053951720938093

140. Menon, A.K., Williamson, R.C.: The cost of fairness in binary classification. In: Proceedings of the ACM Conference on Fairness, Accountability and Transparency (2018). https://proceedings.mlr.press/v81/menon18a.html
141. Barocas, S., Hardt, M., Naryanan, A.: Fairness and Machine Learning: Limitations and Opportunities (2021). fairmlbook.org
142. Corbett-Davies, S., Goel, S.: The Measure and Mismeasure of Fairness: A Critical Review of Fair Machine Learning. Arxiv (2018). https://5harad.com/papers/fair-ml.pdf
143. Hardt, M., Price, E., Srebo, N.: Equality of Opportunity in Supervised Learning. Conference on Neural Information Processing Systems. https://arxiv.org/pdf/1610.02413.pdf (2016)
144. Birhane, A.: Algorithmic injustice: a relational ethics approach. Patterns (2021). https://doi.org/10.1016/j.patter.2021.100205
145. Highhouse, S.: Stubborn reliance on intuition and subjectivity in employee selection. Ind. Organ. Psychol. (2008). https://doi.org/10.1111/j.1754-9434.2008.00058.x
146. Cooper, A.F., Abrams, E.: Emergent unfairness in algorithmic fairness-accuracy trade-off research. In: Proceedings of the AAAI/ACM Conference on Ethics, AI, and Society (2021). https://doi.org/10.1145/3461702.3462519