Goldsmiths
UNIVERSITY OF LONDON

Fairness and bias in algorithmic recruitment tools: An interdisciplinary approach

Airlie Hilliard

Goldsmiths, University of London

Institute of Management Studies

September 2024

Thesis submitted in partial fulfilment of the requirements for the degree of Doctor of Philosophy (PhD).

Candidate Declaration of Authorship

I, Airlie Hilliard, confirm that this thesis and the work presented in it is entirely my own. Where I have consulted the work of others, this has been clearly acknowledged within the thesis.

Signature: Date: 20/09/24

Acknowledgements

I thank my supervisor Dr Franziska (Kiki) Leutner for introducing me to business psychology as an undergraduate, initially planting the seed to complete a PhD, and the applied opportunities she has given me since my placement year. I also thank Roger Thornham for the opportunities to work on real-life algorithm-driven psychometric assessments and making some of the data collection possible through these tools. I thank my second supervisor, Dr Nigel Guenole, for his guidance and the additional opportunities he has provided me with during my PhD.

I thank my friends and family for supporting me during this journey and keeping me sane and Ellie for continuing to invite me places even when I was glued to my laptop. I also thank my colleagues at Holistic AI for their support, encouragement, friendship, and for everything they have taught me.

Finally, a special thanks goes to Dr Emre Kazim and Dr Adriano Koshiyama, who made completing my PhD possible through sponsorship, mentored me, and provided me with countless learning opportunities.

Abstract

Industrial-organisational psychology and computer science are increasingly coming together to create innovative pre-employment selection tools that use non-traditional data and scoring methods to improve the test-taking experience and maximise test validity. However, the two fields have different approaches when it comes to scoring and measuring bias and fairness. Psychology uses simple algebra to score tests, whereas computer science uses predictive modelling and machine learning. Bias and fairness are distinct concepts in psychology but the same in machine learning. Accordingly, using two commercially created algorithmic selection tools, this thesis describes six empirical studies investigating the impact of using an image-based format and machine learning based scoring on test validity/accuracy, fairness, and bias.

In terms of scoring, this research found acceptable subgroup differences in personality and that a machine learning based approach can increase test validity in comparison to a manual-based approach. However, it also found that computer science approaches to mitigating bias can lack compatibility with psychological best practices and equal opportunity laws.

In terms of the effects of format, this thesis provides first data on the fairness of pre-employment tests for neurodivergent test-takers, where neurotypical test-takers have a more positive experience than neurodivergent in general. It found that image-based assessment formats present an opportunity to close the disparity in the experience of the two groups, although further research into the specific features that can support this is needed. Finally, it found that well-trained algorithms are generalisable to neurodivergent populations without causing biased outcomes.

Overall, this thesis provides the foundations for psychologists and computer scientists to work more collaboratively to maximise test validity, fairness, and accessibility of pre-employment tests while minimising biased outcomes.

Thesis summary

- **Chapter 1** –Thesis outline and introduction to algorithmic recruitment tools, bias in psychology and machine learning, fairness perceptions of algorithmic recruitment tools, and neurodiversity.
- **Chapter 2** – Interdisciplinary bias mitigation worked example using a computer science based approach and data from the validation of an image-based personality assessment designed for use in selection described in Chapter 3.
- **Chapter 3 (Studies One and Two**) – The creation and validation of an image-based personality measure scored using a machine learning based approach with adverse impact analysis.
- **Chapter 4 (Study Three)** – Builds on Chapter 3 to compare the impact of different machine learning based scoring approaches on the measure's validity and adverse impact.
- **Chapter 5 (Study Four)** – Interviews with neurodivergent adults on their experiences with recruitment tools and the potential for algorithmic formats to reduce barriers and make the process fairer and less biased.
- **Chapter 6 (Studies Five and Six**) – Comparison of the test-taking experience on a questionnaire-based and image-based assessment of personality between neurodivergent and neurotypical test-takers to explore the potential of image-based formats to close the gap and improve the experience for neurodivergent applicants.
- **Chapter 7** – General discussion that summarises main findings and their implications, research limitations, and future directions.
- **Chapter 8** – References.
- **Chapter 9** – Appendices.

Table of Contents

List of Tables

List of Figures

Chapter 1: Introduction

General Introduction

The overall purpose of this research is to investigate the feasibility of using machine learning scored image-based assessments in recruitment and their implications on fairness and bias, combining insights from industrial-organisational (I-O) psychology and computer science. While these tools – such as video interviews, CV sifters, and game- and image-based assessments – are validated by vendors, there is limited literature on the validity of these tools, how they are perceived by specific groups, and subgroup differences in performance. This is particularly problematic as the creation of fair and valid algorithmically scored recruitment tools requires a multidisciplinary approach, combining expertise from psychology and machine learning to ensure that data points and outputs are job-relevant. Moreover, despite image-based assessments gaining traction in practice (e.g., RedBull Wingfinder) and being offered by a growing number of vendors (e.g., HireVue, Traitify by Paradox), much of the existing research focuses on algorithmically-scored game-based assessments and video interviews. Accordingly, this research aims to fill that gap.

Given that over 40 million selection assessments are completed each year (Chamorro-Premuzic, 2017), it is imperative that recruitment tools are fair and unbiased in order to avoid disadvantaging qualified candidates and widen the talent pool. This is particularly important in the face of algorithmic recruitment tools since biases, even if derived from humans, can have a significant and widespread impact. This is because one of the major benefits of algorithmic tools is their ability to rapidly evaluate candidates (Lofink, 2021), meaning that a single tool may evaluate thousands of candidates each day. On the other hand, a human assessor can only review a fraction of the number of applications and can have a much shallower understanding of each candidate if they only focus on high-level details to maintain pace (e.g., Ladders Inc., 2018). Consequently, even if both the algorithm and human assessor shared the same biases, the impact of the algorithmic system would be further-reaching and

potentially more damaging. As such, it is important to understand how psychology and computer science can come together to ensure that algorithmic recruitment tools are valid, fair, and unbiased. To do so requires an understanding of how these concepts are operationalised and applied in practice in both computer science and I-O psychology in order to bring them together and ensure that candidates are not disadvantaged by these tools.

Within the field of computer science, a significant body of research has emerged addressing multiple aspects of algorithmic bias, including biased or unrepresentative training data and biased outcomes. However, this has resulted in over 20 different definitions of fairness being proposed, where the terms bias and fairness are used interchangeably (Verma & Rubin, 2018). These conceptualisations can range from simply not using protected attributes in models (fairness through unawareness; Kusner et al., 2017) to ensuring the true and false positive rates are equal across subgroups (equalised odds; Hardt et al., 2016) or that the rate of favourable outcomes is similar across groups (statistical parity; Dwork et al., 2012). While efforts have been made to consolidate these definitions into three types – independence, separation, and sufficiency (Barocas et al., 2023; Barocas & Hardt, 2017), where independence is closest to psychology conceptualisations of bias – these definitions focus on categorical models, rather than the continuous models that are more widely used in recruitment. Moreover, some machine learning approaches to mitigating bias are incompatible with recruitment tools since they could violate equal opportunity laws, particularly in the United States (US), if outcomes are changed based on subgroup membership (Civil Rights Act of 1991).

In contrast, in psychology, fairness and bias are distinct concepts (Society for Industrial and Organizational Psychology; SIOP, 2018). Here, bias is a subset of fairness that concerns the applicability of the regression line to multiple groups, known as predictive bias, and irrelevant sources of bias stemming from tools, known as measurement bias (SIOP,

2018). Fairness, on the other hand, is a social construct that is defined by equal group outcomes, equitable treatment of all test takers, comparable access to the construct that the assessment measures, and lack of bias (SIOP, 2018). In other words, fairness is more akin to the overall candidate test-taking experience. The majority of the research investigating fairness perceptions of algorithmic recruitment tools focuses on procedural fairness, which refers to perceptions of the procedure used to carry out the assessment, including assessment format and how applicants judge their ability to influence the outcome of an assessment (Gilliland, 1993).

Investigations into procedural fairness perceptions of algorithmic recruitment tools have led to mixed findings, where differences in perceptions of traditional and algorithmic formats may be driven by the differences in the synchronicity of the assessment formats, rather than the use of algorithms, particularly for video interviews (H. Y. Suen et al., 2019). In contrast, research into applicant reactions to algorithmic formats reveals more promising results. Indeed, machine learning based scoring is conducive to shorter measurements since a large number of datapoints can be extracted in a short amount of time to optimise accuracy (Atkins et al., 2014; Hilliard, Kazim, et al., 2022b; Leutner et al., 2023). Moreover, game-based assessments are more immersive, engaging, and satisfying than traditional measurements of equivalent traits and abilities (Georgiou & Nikolaou, 2020; Leutner et al., 2023). They can also reduce test-taking anxiety (Georgiou & Nikolaou, 2020; Mavridis & Tsiatsos, 2017), leading to a more pleasant candidate experience.

These benefits are not only applicable to game-based assessments; image-based assessments can also reduce test-taking time and offer additional, unique benefits due to their language-agnostic nature. This is because the lack of text could reduce cognitive demand and therefore improve accessibility for a variety of candidates, including those who are neurodivergent. Specifically, applicants with dyslexia and/or ADHD may benefit from the

reduced amount of text and greater reliance on visual processing (Bacon & Handley, 2010; De Beer et al., 2014a; Fassbender & Schweitzer, 2006). On the other hand, autistic applicants may benefit from their heightened ability to notice visual details (Skewes et al., 2015), but may struggle with interpreting the social cues represented in images (Ashwin et al., 2015). Either way, it is essential to understand how neurodivergent test-takers experience algorithmic recruitment tools, particularly given their growing use.

Moreover, given that neurodivergent populations are more prone to test-taking anxiety than neurotypical (Lewandowski et al., 2015; Nelson et al., 2014, 2015) and test-taking anxiety can impact performance (Hembree, 1988; McCarthy & Goffin, 2005), the game-like nature of image-based assessments might help to support performance by reducing anxiety. This could help to improve perceptions of the fairness of the selection procedure and also reduce subgroup differences. However, despite around 20% of the population being neurodivergent and neurodivergent employees being able to offer a number of strengths in the workplace (Doyle, 2020), there is very limited research into how neurodiverse applicants fare with recruitment tools in general, and no research into algorithmic tools or image-based formats in particular.

The remainder of this introductory section first provides an overview of the thesis before examining the relevant related work that laid the groundwork for this thesis. Specifically, it provides a summary of different types of algorithmic recruitment tools and examines how the two disciplines that collaborate to create such tools – namely I-O psychology and machine learning – define and mitigate bias. It then provides an overview of existing research into fairness perceptions of algorithmic recruitment tools before the focus is narrowed to the fairness implications of algorithmic selection assessments for neurodivergent job applicants.

Thesis outline

This thesis seeks to investigate the fairness and bias of algorithmic recruitment tools, predominantly through the lens of commercially developed image-based assessments due to their ease of access, combining insights from computer science and I-O psychology. There is currently a lack of research into the validity of and bias associated with algorithmic recruitment tools since this data is typically only contained in technical manuals and internal documentation. Moreover, given that these tools are generally trained using general populations, it is unclear whether the scoring algorithms are equally accurate for different subgroups and adverse impact analyses typically do not test for differences in neurotype or neurodivergence.

There is also a lack of research into image-based assessments in general, despite evidence of their validity (Leutner et al., 2017) and their potential to increase accessibility. Indeed, the experiences of neurodivergent applicants with recruitment tools in general are under-researched, meaning there is a lack of data to inform accommodations and adjustments that can make these assessments fairer. As such, this thesis seeks to combine insights from both fields to fill these significant research gaps by addressing four key research questions:

- RQ1: How are fairness and bias defined by psychology and machine learning?
    - RQ1.1: Can these definitions effectively be combined in practice?
- RQ2: Can psychology and computer science effectively be combined to measure personality through image-based assessments?
    - RQ2.1: Can this be done without causing subgroup differences?
    - RQ2.2: What is the best way to score these assessments to maximise performance and minimise adverse impact?
- RQ3: How do neurodivergent test-takers experience (algorithmic) recruitment tools?

    - RQ3.1: Do neurodivergent and neurotypical test-takers experience personality assessments differently?
    - RQ3.2: Are experiences with image-based formats enhanced compared to questionnaire-based formats?
- RQ4: Do the scoring algorithms developed using general populations have similar accuracy when applied to neurodivergent test-takers?
    - RQ4.1: Can this be done without resulting in subgroup differences?

Thesis structure

To answer these research questions, this thesis presents six empirical studies and a worked bias mitigation example as follows:

- **Chapter 2: Interdisciplinary bias mitigation: A worked example –** This chapter examines how the two fields can come together in order to effectively measure and mitigate algorithmic bias in recruitment tools, examining approaches from each field in turn. It is accompanied by a worked example of three bias mitigation techniques using data from the validation of an image-based personality assessment to examine the effectiveness of the mitigations and their appropriateness when applied to social contexts such as recruitment.

  Key findings: While one of the mitigation approaches successfully mitigated the bias, two out of the three mitigation approaches lacked compatibility with the assessment due to transforming the input data in a way that no longer made it meaningful and for a lack of compatibility with equal opportunity laws, respectively. This highlights the need to continue to investigate how the disciplines can come together to create unbiased, valid recruitment tools.

  RQ(s) investigated: RQ1 and RQ1.1.

- **Chapter 3: Studies One and Two -** This chapter aimed to provide initial evidence for the feasibility of image-based formats and machine learning based scoring and explore the presence of subgroup differences. To do so, in Study One, a bank of images mapped to the Big Five personality traits was created and refined using a data-driven approach. In Study Two, these image choices were then used to create machine learning based scoring algorithms to predict personality scores, which were then examined for subgroup differences.

  Key findings: The study provided evidence supporting the validity of algorithmically scored image-based assessments of personality, providing substantial evidence of the convergent and divergent validity of this approach and finding that the algorithms resulted in acceptable subgroup differences.

  RQ(s) investigated: RQ2 and 2.1.

- **Chapter 4: Study Three** – Study Three compared several different approaches to scoring the image-based assessment described in Chapter 3, investigating their effect on assessment validity and subgroup differences. Specifically, two machine learning based scoring approaches – Lasso and Ridge regression – ordinary least square regression, and a manual, summative approach were compared using different predictor combinations:
  - All: all images were entered into the models despite the trait they were intended to measure.
  - Mapped: images mapped to each trait in the validation described in Chapter 3 were used as predictors in each model.
  - Intended: images intended to measure each trait, regardless of whether they were mapped to this trait in the validation, were used as predictors in each model.

  Key findings: There was stronger validity evidence when the assessment was scored using machine learning compared to other approaches. Moreover, while both Lasso and

Ridge regression performed similarly, Lasso removes predictors from the model, performing well but using fewer predictors to do so compared to the other approaches. Furthermore, while the Lasso models had similar subgroup differences compared to the questionnaire-based measure, suggesting they could be genuine differences in personality, the Ridge and ordinary regression approaches diminished subgroup differences, which could indicate they are not detecting individual differences in personality as well and may have reduced utility.

RQ(s) investigated: RQ 2.2.

- **Chapter 5: Study Four** – Interviews were conducted with neurodivergent adults on their experiences with traditional and algorithmic recruitment tools. Interviewees were asked about barriers associated with pre-employment tests, including some that may be more unique to neurodivergent applicants compared to neurotypical, as well as adjustments to reduce the impact of these barriers. Questions were asked in relation to both traditional and algorithmic procedures to support Study Five.

  Key findings: Barriers when completing pre-employment tests included compatibility issues with the graphics and display, stigma associated with the disclosure or exposure of their condition, time pressures, and a lack of opportunity to show their unique skills. This, in turn, resulted in a stressful experience and a belief that recruitment tests are biased, were many barriers were specifically referenced as a result of being neurodivergent. Adjustments served as a way to reduce sources of the stressful experience and both sources of stress and way to reduce it were generally relevant for both algorithmic and traditional formats. However, algorithmic formats may be better able to give feedback compared to human recruiters, can be more customisable, and could be made to feel less like a test through gamification.

  RQ(s) investigated: RQ3.

- **Chapter 6: Studies Five and Six–** Study Five investigated the test-taking experience of neurodivergent test-takers compared to neurotypical on a questionnaire-based and image-based personality assessment, as well as whether an image-based format leads to a more positive test-taking experience for neurodivergent adults. Here, test-taking experience was investigated through ratings of motivation, concentration, comparative anxiety, fairness, ease, external attribution, and neurodivergent compatibility as well as open-ended responses. Study Six investigated subgroup differences and accuracy when scoring algorithms were applied to the participants who took the image-based assessment in Study Five.

  Key findings: Neurotypical test-takers had a more positive test-taking experience compared to neurodivergent, and the image-based format did not close this gap. Moreover, for neurodivergent test-takers, although the image-based format was perceived as fairer, it was rated more difficult, harder to concentrate on, and higher in external attribution. Test-takers with a diagnosis of autism or ADHD and autism rated the questionnaire-based measure as more compatible with their neurotype, while other diagnosis groups did not rate either format as more compatible. Finally, the algorithms developed using a general population had similar accuracy compared to the training and test data sets when applied to neurodivergent test-takers and novel subgroup differences based on condition were not identified, indicating the tool did not result in novel subgroup differences.

  RQ(s) investigated: RQ 3.1, 3.2, 4, 4.1

While the studies are related and all endeavour to contribute to the investigation of the validity, bias, and fairness of algorithmic recruitment tools through the lens of image-based assessment, they are distinct. As such, each chapter is presented in turn with its own discussion, recommendations for future research, and conclusion. However, the final section

of this thesis presents an overall discussion of the key findings collectively and their implications.

Ethical considerations

Ethical approval for the studies described was granted in April 2022 by the IMS Ethics Board, Goldsmiths, University of London. These studies posed no risk of physical harm to participants and minimal risk of psychological harm. Three quantitative datasets underpinned these studies: one for Study One, one for Studies Two and Three and the mitigation worked example, and one for Studies Five and Six, although datasets one and two were provided by a commercial partner. An additional supplementary dataset provided by an industry partner was used to support Study Six. All datasets were collected through Prolific Academic and all participant identities were anonymised through unique, Prolific-provided IDs. As such, participants were able to request to withdraw their data from the dataset at any point, including after the study, although this did not occur. Furthermore, the image-based personality assessments completed by participants were in the creation and validation stages, meaning that participants did not receive their personality scores or any feedback on their profiles, minimising potential harm. Study Four was underpinned by qualitative data in the form of interviews. For the interviews, it was not possible to keep participant identities anonymous during the interviews due to scheduling, but identities were anonymised with initials prior to analysis and not shared with anyone on the research team. Interviewees received the contact details for the primary researcher and supervisor and were also informed that they were able to withdraw at any point for any reason.

Overview of algorithmic recruitment tools

Algorithmic recruitment tools use algorithms to score pre-employment tests by predicting job-relevant constructs from a variety of data sources (Guenole et al., 2023; Hilliard, Guenole, et al., 2022). These data sources encompass both big data, such as social media footprints and chatbot conversations, and more traditional data, such as responses to

and behaviour while completing purposefully designed psychometric assessments (Albert, 2019; Guenole et al., 2023). The algorithms themselves can also vary in their complexity, from more simple linear models (e.g., Hilliard, Kazim, et al., 2022b) to complex models that combine data from multiple sources (e.g., Landers et al., 2022). While it is true that assessments of any format could be scored algorithmically, more traditional questionnaire-based measures are typically scored using a scoring key, while algorithmically scored assessments are typically more of a novel format. This includes video interviews, game-based and gamified assessments, and image-based assessments (Raghavan et al., 2020), all of which move away from traditional approaches to measuring job-relevant constructs with the aim of enhancing the candidate experience. The following subsections explore each of these three types of algorithmic assessment tools, focusing on game-based and image-based assessments to converge with the focus of this thesis.

Asynchronous video interviews

Asynchronous video interviews are increasingly being offered by vendors and require candidates to record responses to predetermined questions. Candidates are typically given around 30 seconds to prepare their response and up to two minutes to record their response, although the exact timings can vary by provider (Dunlop et al., 2022). In some cases, candidates are also given the opportunity to re-record their responses, although the majority of candidates given this option choose not to do so (Dunlop et al., 2022). Once submitted, responses are then evaluated by either human raters or algorithms.

When algorithms are used, they are typically trained to predict job-relevant constructs, such as personality, using features extracted from candidate responses, where these features may be verbal, paraverbal, or non-verbal (Hickman, Saef, et al., 2021). Here, verbal features refer to what candidates say, including the specific words used or categories and length of words used, and length of words used (Hickman, Bosch, et al., 2021). On the other hand, paraverbal features are vocal features that represent the way in which candidates

communicate, including pitch, speaking volume, duration of speaking, and duration of pauses (Hickman, Bosch, et al., 2021). Finally, non-verbal features include body language and facial expressions displayed during interviews, where facial expressions are typically interpreted using facial action units (Hickman, Bosch, et al., 2021). Facial action units were first developed by Ekman and Friesen in 1978 based on an anatomical mapping of facial movements that can be used to infer facial expressions. These action units can be extracted from video interviews and used as features in the predictive model. However, the use of facial analysis in video interviews is controversial due to concerns about how this may adversely affect applicants with neurological differences, as well as how the accuracy of inferences may vary depending on the candidate's race (Electronic Privacy Information Center, 2019). As such, some vendors have chosen to remove the non-verbal features from their algorithms (Kahn, 2021; Zuloaga, 2021).

Moreover, concerns have been raised about how video interviews could be susceptible to cheating using generative artificial intelligence (AI), where candidates may use tools such as ChatGPT to curate answers to questions and read them verbatim or enhance them with additional personal or contextual information (Canagasuriam & Lukacik, 2024). Indeed, preliminary findings indicate that using ChatGPT to generate responses to interview questions leads to more positive performance ratings without impacting the delivery of responses, although the use of the tool can impact honesty ratings (Canagasuriam & Lukacik, 2024). However, although methods are being developed to detect issues such as script sharing in interviews (Cornell, 2023), and answers generated by ChatGPT can contain similar language that could assist with the detection of AI-driven cheating (Canagasuriam & Lukacik, 2024), there is currently a lack of research into how algorithms may be adapted to detect and mitigate the effect of cheating in video interviews.

Game-based assessments and gamification

Unlike video interviews, game-based assessments are less prone to cheating and faking as they cannot be prepared for in the same way as video interviews and can be harder to manipulate. Game-based assessments engage test-takers in a core gameplay loop with the intention of bringing about a gameful experience while inferring specific abilities or constructs (Landers & Sanchez, 2022). A core gameplay loop is a set of repeated actions that a player undertakes during the completion of the game to meet its objectives (Guardiola, 2016; Landers & Sanchez, 2022), while a gameful experience is a voluntary motivation to pursue the goals of a game, where these goals must be perceived to be achievable (Landers et al., 2019; Landers & Sanchez, 2022). As with video interviews, game-based assessments can be scored manually with a scoring key or algorithmically, making use of additional data sources outside of pure responses in order to calculate scores. For example, Auer et al., (2022) developed a game-based assessment of cognitive ability and conscientiousness scored using machine learning that uses trace behavioural data from gameplay in combination with performance on the game to compute scores. Specifically, the algorithm combines data such as the number of correct and incorrect responses and rounds completed with additional data such as mouse movement and clicks for more sophisticated scoring. Moreover, F. Y. Wu et al. (2022) designed two game-based assessments to measure conscientiousness, where one required test-takers to earn as much revenue as possible by clicking on revenue-producing buildings (Click Town) and the other was in the form of word searches (Word Find). However, scored by algorithms that used GBA data such as time, game progression/levels completed, and cues to desired behaviour, the GBAs unintentionally measured cognitive ability instead of the intended facets of achievement-striving, self-discipline, and cautiousness (F. Y. Wu et al., 2022).

Although the term game-based assessment is often used as an umbrella term to describe any assessment that has elements of game, game-based assessments are distinct from

other formats such as gamified and gamefully designed assessments (Landers & Sanchez, 2022), where game elements include features such as animation, sound effects, instant feedback, levels of difficulty, and progress bars (Landers, Armstrong, et al., 2022). Similar to game-based assessments, gamefully designed assessments involve the creation of a new assessment, where game mechanics or concepts such as setting goals, creating immersion, and providing feedback through scoring are used to guide decision-making when designing an assessment (Landers & Sanchez, 2022). In contrast to game-based and gamefully-designed assessments, gamified assessments add game concepts and mechanics to existing assessments (Landers & Sanchez, 2022). There are multiple ways to do this, such as through game framing, where a traditional assessment is framed as a game, or through storification, where an assessment is gamified by converting it to a story (Landers & Sanchez, 2022). For example, McCord et al. (2019) created a storified fantasy game where test-takers play as a character that wakes up underground and must get to the surface through a series of tunnels. As the character moves through the tunnels, they encounter creatures and scenarios and are presented with a choice of actions to take that are mapped to personality traits, thus being used to infer personality. On the other hand, Collmus and Landers (2019) investigated the impact of framing cognitive ability assessments as puzzles and logic games compared to intelligence tests, finding game-framing to reduce completion time estimations.

Image-based assessments

Like game-based, gamified, and gamefully-designed assessments, image-based assessments are also more resistant to faking and cheating, particularly AI-driven cheating, due to the lack of a verbal component. Image-based selection assessments use image choices to measure job-relevant constructs such as personality and creativity. Early non-verbal personality assessments used images to depict certain actions and asked respondents to rate how likely they were to engage in the shown behaviour, therefore replacing the statement with an image but retaining the Likert-scale response approach (Paunonen et al., 1990, 2001).

However, more modern approaches have moved away from the use of tedious Likert scales, replacing response options rather than the question stem/statement with images. For example, Krainikovsky et al. (2019) tagged images with information relating to the objects, behaviour, emotions and scenery in the image and used choices to predict personality using a machine learning based approach. Although this resulted in low convergent validity with the NEO PI (Costa & McCrae, 2008), ranging from $r = .06$ for neuroticism to $r = .28$ for agreeableness, this measure was not created for use in selection.

In contrast, Leutner et al., (2017) developed an image-based assessment of creativity that was intended for use in selection, where test-takers were presented with text-based questions and asked to indicate which image is most like them from the options available. Scored using machine learning algorithms that predict scores on questionnaire-based measures of those traits, the assessment had good concurrent validity the target scale curiosity: $r = .35$, cognitive flexibility: $r = .50$, and openness to experience $r = .50$). Image-based formats can also be game-based or gamefully designed if they incorporate elements of game, such as sound effects or progress bars, to enhance the candidate experience.

In summary

Psychological theory is increasingly being combined with machine learning techniques to create innovative tools that can be used throughout the talent management lifecycle. Algorithmic assessment formats in particular combine psychological theory on the knowledge, skills, attributes, and other characteristics required to be successful in a role and the measurement of these competencies with machine learning based scoring algorithms that can make use of non-traditional data. While these assessments can take many formats, some of the most widely used algorithmic selection procedures are algorithmically scored video interviews, game-based assessments, and image-based assessments. Given the access available to image-based assessments of personality, these are used as the basis for the research in this thesis.

Bias in psychology and computer science

As the interdisciplinary approach to hiring is becoming increasingly widespread, it is being met with concerns about how the priorities of each discipline can align and the potential harms associated with the use of algorithms in recruitment. One of the most significant concerns is the potential of the algorithms to result in biased outcomes (Hunkenschroer & Luetge, 2022; Tippins et al., 2021) since they can perpetuate and amplify existing biases (Lloyd, 2018) and even small amounts can compound to have large effects (Hardy et al., 2021). However, concerns about bias are not merely a pessimistic prediction; there have already been multiple instances of algorithmic systems designed for use in recruitment resulting in biased outcomes. Perhaps the most well-known example of this is Amazon's decommissioned recruitment tool, which was found to be biased against female applicants and was, therefore, not deployed. The algorithm, which evaluated candidates' resumes, was trained using the resumes of those who had previously applied for technical roles at Amazon (Dastin, 2018), the majority of whom were male, reflecting the gender imbalance in the tech industry (PricewaterhouseCoopers, 2017). This led to the algorithm penalising any resumes that contained the word "women's", therefore being biased against female applicants who wrote about their membership to a women's team (Dastin, 2018).

LinkedIn and Facebook have also come under fire for bias in the algorithms used to display job adverts to users of the sites. The algorithm LinkedIn used to match job candidates with suitable opportunities was found to be biased against female users, with male users being referred for open roles more often than females. This was due to the fact that the algorithm ranked applicants on how likely they were to apply for the position that they were shown, and males are typically more determined when it comes to seeking out new opportunities and are consequently more likely to click on the ad (Beatrice, 2021; Wall & Schellmann, 2021). While this has now been resolved through the introduction of an

additional algorithm that ensures that there is a more balanced gender distribution in the targeting of job ads (Wall & Schellmann, 2021), the same cannot be said for Facebook. An investigation into the algorithms used by the site found that females are less likely to be shown an advert for a position in a male-dominated company compared to an advert for an almost identical position in a company with a more balanced gender distribution of employees. Even when qualifications are controlled for, this problem persists and is present across multiple roles and industries (Imana et al., 2021). Further, it is more expensive to display job adverts to females on Facebook, especially those aged between 25 and 44, since they are seen as a prized demographic (Lambrecht & Tucker, 2019). As such, there is a clear need to ensure that bias in algorithmic recruitment tools is prevented, identified, and mitigated. To do so requires an understanding of how bias is defined and mitigated in I-O psychology and machine learning, as well as the equality of opportunity legal landscape.

Equal opportunity and the law

Equal opportunity laws around the world protect individuals from discrimination in employment decisions based on certain protected attributes. For example, Article 21 of the European Union (EU) Charter of Fundamental Rights (European Union, 2000) prohibits discrimination on the grounds of race, age, sex, ethnicity, religion or belief, disability, ethnic or social origin, genetic features, language, political (or other) opinion, minority status, property, birth, and sexual orientation. The Charter takes a contextual, case-by-case approach to determine whether discrimination is occurring (Wachter et al., 2021). In the United Kingdom (UK), the Equality Act of 2010 protects individuals from discrimination in public services including employment. As with the EU Charter, the Equality Act takes a non-prescriptive approach to discrimination and bias, favouring a case-by-case evaluation and not specifying any particular evidence needed to support claims of non-discrimination in recruitment decisions (Equality Act, 2010). As such, I-O psychologists in the EU and UK often look to more prescriptive legal landscapes to determine how to measure bias in

recruitment (Hilliard et al., in press). Indeed, the US has some of the most prescriptive equal opportunity laws, where the Equal Employment Opportunity Commission's (EEOC) Uniform Guidelines on Employee Selection Procedures (EEOC, 1978) have governed the laws surrounding the use of pre-employment tests for the past 40 years. The Guidelines, which enforce the Civil Rights Act of 1964, have two major themes: outcome parity and validation studies.

Outcome parity

According to the Guidelines, individuals belonging to different subgroups should have equitable outcomes or outcome parity. This is operationalised in terms of equal hiring rates for groups based on sex and race/ethnicity as measured using the four-fifths rule, which says that adverse impact (differential hiring rates) occurs when the selection rate of one subgroup is less than four-fifths (.80) of the selection rate of the subgroup with the highest rate (EEOC, 1978). While this is generally taken as the rule of thumb, the Guidelines note that a ratio less than .80 might not indicate adverse impact, particularly if the analysis was carried out using a small sample or if the sample is not representative of the typical applicant pool (EEOC, 1978). When this rule is violated, the Guidelines stipulate that if there is another equally valid selection procedure available that is associated with less adverse impact, then the alternative should be used.

While the Uniform Guidelines endorse the use of the four-fifths rule to determine adverse impact, they do note that other metrics may be used, providing there is justification for this. For example, the two standard deviations rule, also known as the z-test, has been used in some court cases to measure adverse impact (PSI, 2018) and is also endorsed by the Federal Contract Compliance Manual (Office of Federal Contract Compliance Programs, 2020), which provides guidance on equal opportunity for workers. The two standard deviations rule compares the expected and observed hiring rates for different groups, where a value greater than two indicates adverse impact, and is suitable for use with smaller sample

sizes (Collins & Morris, 2008; Murphy & Jacobs, 2012). However, the two metrics can result in discrepant findings (S. B. Morris & Lobsenz, 2000), meaning I-O psychologists can often look to additional metrics to assess adverse impact. For example, the effect size of the difference in mean scores of subgroups can be examined using Cohen's *d*. This statistic, which is not affected by sample size, indicates that there is a small effect size when *d* = +/-0.20, medium when *d* = +/-0.50, and a large effect size when *d* = +/-0.80 (Cohen, 1992).

Although the Guidelines do not require adverse impact testing based on age since this is addressed by the Age Discrimination Act, some do extend their adverse impact analysis to include age (e.g., Fisher et al., 2017; Hilliard, Kazim, et al., 2022a; HireVue, 2020; Klein et al., 2015). More comprehensive adverse impact testing like this may become the norm as technology is increasingly used in recruitment since there are differences in the approach to and perceptions of technology based on age. For example, age is negatively related to technology self-efficacy (Ellison et al., 2020) and to perceptions of ease of use and usefulness of technology (Hauk et al., 2018). Older candidates also have more negative perceptions of the fairness and usability of video interviews compared to younger candidates (Basch & Melchers, 2019), but approach video interviews more formally, considering their attire, background, and body language more than younger applicants (McColl & Michelotti, 2019). Since background and accessories can influence algorithmic judgements of video interviews (Fergus, 2021), this could potentially result in adverse impact against younger applicants. However, Melchers & Basch (2021) found that younger applicants perform better on a game-based assessment of complex problem-solving than older applicants, although the effect sizes for the differences in performance were small and others report that age is not related to performance on game-based assessments of working memory, complex planning, numerical comprehension, or logical reasoning (Ellison et al., 2020). Given the mixed findings surrounding disparities in the performance of older and younger users, this indicates that age

differences could be an additional factor to consider during adverse impact analysis of algorithmic recruitment tools, something that is not suggested in the Guidelines.

Validation studies

While the Guidelines provide little guidance on mitigating bias, except for considering how cutoff scores influence adverse impact, they require evidence of the validity of a measure that is associated with adverse impact to justify its continued use (EEOC, 1978). Where evidence of the validity of a selection procedure is required, the Guidelines necessitate documentation to support three types of validity (EEOC, 1978):

- **Criterion-related validity** – evidence that the assessment is predictive of or correlated with important aspects of job performance.
- **Content validity** – evidence that the content of the assessment represents important aspects of job performance.
- **Construct validity** – evidence that the assessment measures characteristics that have been identified as being important for success in the role using a series of studies, which may include measures of content or criterion validity.

The Guidelines place restrictions on the types of validation studies that are suitable for measures of traits or constructs such as personality and intelligence, with content validity said to be unsuitable justification for the use of these measures. In addition, the Guidelines state that there is a lack of literature available in relation to construct validity, which could restrict validation studies of personality and intelligence to criterion-related validity. However, since the Guidelines were published over 40 years ago, there have been advancements in content validation studies, with multiple approaches suggested (Sackett, 1987).

A key feature of all three types of validity is job analysis, where the relevant behaviours, duties, and outcomes of a job are identified from job information. While the

Guidelines state that the way in which a job analysis is conducted is down to the discretion of the assessor, job analysis aims to ensure that the constructs being measured by the assessment are relevant to the attributes necessary to perform well in the job. Further, particularly for construct validity studies, the Uniform Guidelines require that the construct being measured be based on psychological theory. While in I-O psychology it is generally regarded as good practice to ensure that there is a theoretical basis for predictors (SIOP, 2018; Tippins et al., 2021), the same approach is not always shared when using algorithmically driven assessments; while each item in a questionnaire is carefully considered to ensure it has content validity, the large number of predictors used in machine learning algorithms makes it almost impossible to justify each predictor and how they interact in the model. Indeed, the value of machine learning is that it can process large datasets and identify unintuitive connections between data (Mitchell et al., 2021; Wachter et al., 2021), meaning that providing the assessment predicts relevant organisational behaviours, some may consider the theoretical basis of predictor interactions as nice to know (Tippins et al., 2021).

Algorithms using unstructured data may present additional challenges in terms of justifying predictors and how they are used by models. Indeed, if the datapoints included in the algorithm are not sufficiently investigated, this can affect the validity of the measure. For example, accessories such as glasses and headscarves, and objects visible in the background of a video interview such as a bookshelf or art can affect algorithmic judgements of personality from video interviews (Fergus, 2021). This highlights the need for caution when creating algorithms using a large number of datapoints, particularly when the algorithms are based on unstructured data.

In addition to using job analysis when designing a selection procedure, the Guidelines also provide recommendations for strengthening the conclusions of validation studies, such as cross-validation (EEOC, 1978). The use of cross-validation is also endorsed by I-O

psychology, where psychologists are encouraged to apply the scoring model to multiple samples from the same population (SIOP, 2018). Similarly, cross-validation is also used in machine learning to find the best model. Here, the dataset is divided into training data, which is used to train the model, and test data, which acts as an unseen sample. This is repeated using different combinations of training and test data to find the model that fits the data best (Schaffer, 1993). This approach is possible in machine learning since predictive measures are often based on Big Data, having sample sizes in their tens of thousands or even millions (Bachrach et al., 2012; Kosinski et al., 2013; A. H. Schwartz et al., 2013), allowing the dataset to be divided into sections of substantial size. In psychology, however, sample sizes are usually much smaller, typically in the hundreds. Therefore, to cross-validate, multiple samples are needed since dividing the data would result in very small sub-samples. When psychology and machine learning come together, sample sizes are usually in the hundreds or thousands (Hilliard, Kazim, et al., 2022a; Leutner et al., 2017; Quiroga et al., 2015, 2016), being much smaller than those that are typical of machine learning as participants need to be recruited to provide the data to train the algorithms. However, modern prediction methods that use a machine-learning based approach perform reasonably well when sample sizes are small (< 300), although larger samples perform better in terms of representing true cross-validation with independent samples (Putka et al., 2018). This is useful for I-O psychologists working with algorithmic recruitment tools as they can use the machine learning approach to cross-validation even when sample sizes are only modest, which is beneficial in situations where it is difficult to collect additional data to cross-validate models.

Bias in I-O psychology

I-O psychology builds upon the legal landscape to create best practices for measuring adverse impact and providing validity evidence, particularly in light of technological advancements. Indeed, publications are released from organisations such as the Society for Industrial and Organizational Psychology and the American Psychological Association to

outline best practices for I-O psychologists in line with current scientific knowledge and theory. Of particular relevance to I-O psychologists are the Principles for the Validation and Use of Personnel Selection Procedures (SIOP, 2018) and The Standards for Educational and Psychological Testing (American Educational Research Association, American Psychological Association & National Council on Measurement in Education, 2014), hereafter referred to as the Principles and Standards, respectively. Bodies outside of the US also release guidance on the use of selection assessments for I-O psychologists, with the British Psychological Society (2006) releasing guidelines specifically for the use of online tests, although they were published before the use of machine learning in recruitment was a common occurrence. The Principles (SIOP, 2018) are the most up-to-date guidelines available to I-O psychologists and touch on recent trends in recruitment such as gamification and the use of machine learning.

One of the key differences between the Principles (SIOP, 2018) and the Uniform Guidelines (EEOC, 1978) is the approach to group differences in scores. On one hand, the Guidelines assume that all group differences are indicative of adverse impact and stipulate that the use of selection procedures associated with adverse impact should be justified through evidence of the validity of the measure. On the other hand, I-O psychologists investigate the cause of group differences to a greater extent than is required by the Guidelines. In fact, instead of using the term adverse impact, the Principles refer to subgroup differences, and, instead of assuming that group differences indicate negative consequences, they acknowledge that group differences do not necessarily indicate bias (SIOP, 2018). Rather, group differences in scores can reflect true group differences in the work-relevant outcomes or behaviours assessed by the measure. If subgroup differences in scores are reflective of true group differences in ability, the Principles suggest that this can strengthen the validity of the measure (SIOP, 2018). However, when group differences in scores do not reflect differences in ability, this can weaken the validity of the measure. There are, therefore

two sources of variance in the scores of individuals: variance due to true individual differences in the construct and variance due to error (Vandenberg & Lance, 2000).

Another difference between the Principles and Uniform Guidelines is the approach to supporting the validity of a measure, where the Principles take a more holistic approach than the Guidelines. In particular, the Principles state that almost all relevant information about a selection procedure can contribute to determining its validity (SIOP, 2018), moving away from the trifurcated concept of validity adopted by the Guidelines. Moreover, the Principles explicitly define both fairness and bias but do not define adverse impact. Here, fairness is a social concept that can be defined as equal group outcomes, equitable treatment of all test takers, comparable access to the construct that the assessment measures, and lack of bias (SIOP, 2018). On the other hand, bias is defined as systematic errors in test scores that differentially affect different subgroups and can take two forms (SIOP, 2018):

- **Measurement bias** – occurs when irrelevant variance results in systematic group differences in scores and is a concern for both predictor and outcome variables. In other words, measurement bias occurs when group differences in scores do not reflect differences in ability. For example, when two subgroups are, on average, given different scores despite comparable job performance.
- **Predictive bias** – occurs when the same regression line cannot be applied to all subgroups. As a result of predictive bias, which is also known as differential prediction, individuals who have the same ability but belong to different subgroups are given different scores by the measure.

Measurement bias can occur at the item level, referring to differences in the probability of selecting a particular (or the correct) answer for different subgroups with the same underlying ability, or test level, referring to differences in the total scores for individuals belonging to different subgroups who have the same underlying ability. In the

item response theory literature, item and test bias are referred to as differential item functioning and differential test functioning, respectively (S. Stark et al., 2004). Although less widely researched than predictive bias, there are a number of ways to investigate measurement bias, including through examining the effect size of mean score differences for subgroups and through examining adverse impact at different cutoff scores (S. Stark et al., 2004).

In contrast, predictive bias focuses on the relationship between the scores derived from the measure and external criteria (S. Stark et al., 2004) and can be divided into non-compensatory and compensatory bias. The former occurs when an assessment yields different mean scores for different subgroups, despite them having the same level of competence on the construct being measured. The latter occurs when although the mean score for different subgroups is the same, there is a disparity in the variance in scores of each subgroup (Tay et al., 2022), and can be examined using moderated multiple regression using the predictor score, subgroup membership and the interaction between them to examine the fit of a common regression line for multiple subgroups. Here, differences in the slope or intercepts signal predictive bias (Berry & Zhao, 2015), although there has been some debate about whether intercept differences are as important as slope differences (Landers, Armstrong, et al., 2022).

Like the Guidelines, the Principles suggest that if multiple tests are compiled and used as the basis for selection decisions, then tests for bias should examine the total process instead of individual tests (SIOP, 2018). However, while the Guidelines emphasise that scores should be used to determine adverse impact, the Principles also recommend examining bias at the equational level, as well as score level. They also note that further research is needed to investigate the cause of different forms of predictive bias, especially with constructs other

than cognitive ability, and how to overcome issues associated with low statistical power when testing for predictive bias (SIOP, 2018).

Bias in Computer Science

In contrast to I-O psychology, computer science typically uses the terms bias and fairness interchangeably (Goel et al., 2018), and there is a lack of a single definition of bias/fairness because the concept has changed over time. Early definitions focused on differential prediction for different subgroups who were consistently scored too high or too low compared to other subgroups, while more recent definitions focus on differential item functioning and error rates (B. Hutchinson & Mitchell, 2019). Compiling the most widely used definitions of fairness, Verma and Rubin, (2018) identified 20 different definitions, some of which were known by multiple names. For example:

- **Fairness through unawareness** – a model is fair if it does not use protected characteristics to make predictions (Gajane & Pechenizkiy, 2017). This definition is incompatible with I-O psychology since proxy variables can be an issue; given that patterns of non-verbal communication vary between males and females in video interviews (Frauendorfer & Mast, 2014), this could be a proxy for determining gender, for example. Even if proxy features were removed from the model, it could still be argued that fairness through unawareness is not being met as some may assert that the model now knows the protected attributes as they were encoded in the features that were removed from the model.
- **Fairness through awareness** – knowing the protected attributes of individuals in the training data can make the model fairer if they are used to ensure that similar individuals receive similar predictions regardless of their subgroup membership (Dwork et al., 2012). If the model does not predict similar individuals similar scores, this indicates that it may be biased or unfair (Dwork et al., 2012).

- **Group fairness** (also known as statistical parity, equal acceptance rate, and benchmarking) – individuals from different subgroups should have the same probability of being assigned to the positive condition (Verma & Rubin, 2018). To contextualise this to recruitment, Black, Asian, Hispanic and White applicants should all have the same probability of being recommended for an interview by the model.
- **Conditional statistical parity** – extends group fairness, positing that when other legitimate influences are controlled for, the subgroups should have an equal likelihood of being assigned to the positive condition (Verma & Rubin, 2018). This means that when relevant factors such as education and experience are controlled for, candidates from all subgroups should have an equal likelihood of being recommended for an interview. While not completely consistent with the notions of (lack of) bias of I-O psychology, this is more compatible with the field because it acknowledges that there could be legitimate reasons for variations in performance between subgroups and, therefore, how individuals are classified.
- **Predictive parity** (also known as outcome test) – both the protected and non-protected attribute have an equal positive predictive value, calculated by dividing the number of true positives by the total number of positive cases.
- **Equalised odds** – the true and false positive rates of different subgroups should be equal (Verma & Rubin, 2018)
- **Conditional use accuracy equality** –satisfied when the probability of true positives and true negatives is equal for different subgroups. Implied by this definition is that there should be equal accuracy for different subgroups, something that is explicitly endorsed by the overall accuracy equality definition of fairness (Verma & Rubin, 2018).

In summary

The use of algorithms introduces several novel sources of bias that must be considered when designing, developing, and deploying algorithmic recruitment tools. While bias is a well-defined concept in I-O psychology, there is much less cohesion in computer science. Indeed, computer science does not have a single definition of bias, and definitions that do exist can be mutually exclusive, meaning computer scientists may be forced to choose the definition they are optimising their model with respect to. This can be problematic as not all definitions of bias are designed with social applications in mind, so could violate equal opportunity laws. On the other hand, I-O psychology has several well-established metrics for measuring bias and adverse impact that are typically aligned with equal opportunity laws. However, traditional approaches to bias mitigation in I-O psychology may not consider the novel sources of bias introduced by the use of algorithms. As such, it is essential that the compatibility of the two fields with respect to bias measurement and identification is investigated to avoid differential hiring rates based on irrelevant characteristics.

Fairness perceptions of algorithmic recruitment tools

Moving away from bias to fairness, there is an emerging body of research investigating the fairness perceptions of algorithmic tools. The most widely used model of fairness was developed by Gilliland (1993) and was derived from organisational justice theory. Gilliland's model conceptualises fairness perceptions of selection tools into two broad categories: distributive justice and procedural justice. Here, distributive justice concerns the equity, equality, and the fulfilment of needs while completing pre-employment tests. Specifically, a test is viewed as equal if unrelated characteristics such as gender or ethnicity do not influence outcomes and can be perceived as equitable if the outcomes of the assessment are in line with an applicant's expectations based on their previous success and qualifications (Gilliland, 1993). The fulfilment of needs builds on these elements, where candidates judge whether their needs are met to facilitate equal and equitable outcomes. This

includes providing accommodations to those who need them, for example, to ensure that outcomes are not unfairly disadvantaged by their disability (Gilliland, 1993).

On the other hand, procedural justice is concerned with the procedure used to conduct a pre-employment test. This type of fairness can be influenced by the test type, human resources policies, and human resources personnel and is driven by factors including job relatedness, opportunity to perform, feedback, and communication received, all of which affect the perceived ability to influence a decision (Gilliland, 1993). Much of the research regarding the perceived fairness of algorithmic recruitment tools investigates procedural justice since it is a concern about the use of algorithms shared by both candidates and human resources practitioners alike (Fritts & Cabrera, 2021; Mirowska & Mesnet, 2021).

Procedural justice perceptions of algorithmic recruitment tools

The perceived procedural justice of algorithmic recruitment tools is typically positive, although this can vary by assessment tool. For example, Georgiou and Nikolaou (2020) report that algorithmically scored game-based situational judgement tests are viewed as fairer than traditional situational judgement tests. In contrast, Suen et al., (2019) report that although there is a preference for synchronous video interviews compared to asynchronous, fairness perceptions do not vary when asynchronous video interviews are judged by humans compared to an algorithm, suggesting that it is the synchronicity that drives reactions rather than the use of algorithms. Concurring with this finding, a large-scale study examining the reactions of almost 645,000 real-life applicants from 46 countries to synchronous and asynchronous video interviews found more positive perceptions of the synchronous interview in terms of satisfaction and effectiveness (Griswold et al., 2022). More recent research reveals further contrasts, with algorithmic tools being seen as less fair when used in later stages of the recruitment funnel, such as for video interviews, and equally as fair as human ratings when used in earlier stages, such as for application screening (Köchling et al., 2022). Such perceptions can be examined at both a high level and a more granular level including in

relation to specific variables such as social presence, interpersonal treatment, perceived behavioural control, and consistency (Langer et al., 2019). Investigating procedural justice at this level can provide greater insight into the factors driving fairness perceptions, and how applicants interact with the tools. The following sub-sections examine the behavioural control, social presence, and creepiness of algorithmic recruitment tools to reflect the elements of procedural justice most frequently investigated in the literature.

Algorithmic recruitment and behavioural control

Perceived behavioural control refers to the extent to which candidates believe that they can control or influence an outcome with their behaviour (Langer et al., 2019). In other words, by engaging in certain behaviours, candidates believe they can influence hiring decisions. In contrast to the mixed findings of overall procedural justice, when narrowing the focus to perceived behavioural control, fairness perceptions of algorithmic recruitment tools are more consistent; several studies report that across different algorithmic recruitment tools, there is less perceived behavioural control. Indeed, despite being seen as more objective, algorithmic asynchronous video interviews are seen as having less opportunity to perform in comparison to human-rated asynchronous video interviews (Kaibel et al., 2019). Likewise, algorithmic resume screening tools are perceived to be less able to judge human character compared to human screeners (Lee, 2018), despite humans spending less than 10 seconds reading each resume (Ladders Inc., 2018) and being found to be biased against non-white sounding applicants (Bertrand & Mullainathan, 2004). Specifically, applicants perceive it to be unfair that algorithms cannot make exceptions whereas human raters can. This indicates that applicants perceive that there is less opportunity for them to perform with algorithmic judgements because of their objectivity, meaning that they are less able to manipulate the algorithm than human raters. There have also been claims that the use of algorithms can be reductionist and reduce the autonomy applicants have over their self-representation due to the rigid conception by algorithms of how attributes should or can be displayed (Aizenberg et al.,

2023). Therefore, efforts to make recruitment funnels more objective and standardised, and therefore fairer in terms of a lack of bias, have resulted in them being perceived as more unfair and sometimes unethical.

Algorithmic assessments and creepiness

Related to behavioural control is creepiness - a sense of ambiguous discomfort and not knowing exactly how to interact with the tool (Langer et al., 2019), where a lack of transparency and a limited ability to control or influence the technology can result in a tool being judged as creepy (Köchling et al., 2022; Langer et al., 2018, 2019; Tene & Polonetsky, 2013). This is particularly true for video interviews, where participants in two studies of automated video interviews judged them to be significantly creepier than video conference interviews (Langer et al., 2019) and asynchronous interviews evaluated by humans (Oostrom et al., 2024). Similarly, when compared to perceptions of human-evaluated telephone interviews, AI-supported video interview evaluations are rated as significantly creepier (Köchling et al., 2022). Moreover, the perception of AI as creepy mediates the relationship between the evaluation of the use of AI and organisational attractiveness, and concealing the use of AI results in negative reactions if a candidate discovers that the technology was used at a later date (Köchling et al., 2022).

A possible explanation for applicants perceiving less opportunity to perform and thus creepiness is a lack of knowledge about what the algorithm uses to make judgements. This is particularly due to the fact that algorithms are often black-box or glass-box systems (Cheng & Hackett, 2021), meaning the internals of the model are uninterpretable or unknown (Guidotti et al., 2018). However, using videos of virtual characters responding to and adapting to candidate behaviours, Langer et al. (2018) found that providing some explanation about how the character used inferences about non-verbal behaviour – such as facial expressions, gestures, and speech – to appropriately respond to candidates worsened perceptions of fairness and organisational attractiveness. Moreover, providing more

information to candidates about how algorithms make decisions can invoke concerns about privacy due to the data being used to compute judgements (Langer et al., 2021). This suggests that it is not the lack of knowledge about algorithmic decision-making that drives lower perceptions of opportunity to perform with algorithmic tools, but the fact that candidates are more confident in influencing human decision-making since the factors that humans consider when making judgements are more intuitive and superficial compared to datapoints used by algorithms. Therefore, the more subjective nature of human judgements is preferred because candidates believe that they are more able to influence the decision of the rater through impression management, which is not uncommon for candidates to use during their application (Weiss & Feldman, 2006).

Algorithmic recruitment and social presence

Perceptions of social presence are concerned with the extent to which applicants perceive there to be an interpersonal connection facilitated by empathy and warmth during an interaction (Langer et al., 2019). Research into algorithmically analysed video interviews and tests of performance found them to be rated as less personable, or lower in social presence, than manual ratings in a fictitious hiring scenario, despite participants not interacting with a human in either condition (Kaibel et al., 2019). This suggests that algorithmic judgements are perceived as being unfair as they are less able to reflect human values and replicate interpersonal exchanges as an algorithm cannot empathise with candidates like humans can. Indeed, despite acknowledging that algorithmic recruitment tools are more objective than human ratings and that human ratings can be biased, participants still reported perceiving algorithmic recruitment tools as less fair due to the lack of human connection and interaction (Mirowska & Mesnet, 2021). This could explain why algorithmic tools used earlier in the funnel, where there is typically less human interaction, are seen as equally fair to human ratings, while algorithmic tools used later in the funnel, such as during the interview stage,

are viewed as less fair than human ratings (Köchling et al., 2022) since there are differences in the level of human connection expected.

Concerns about lack of opportunity to perform are also echoed by human resource practitioners, some of whom believe that algorithms have artificial values instead of human values since the relationship between humans is very different to the relationship between humans and computers as the latter can become gamified and lack sincerity (Fritts & Cabrera, 2021). Removal of or limiting the human element in the recruitment process also has implications for the opportunities for recruiters to form connections with candidates (Li et al., 2021), particularly if much of the process is automated and recruiters are only involved in making the final decision based on algorithmic recommendations.

However, interestingly, a so-called algorithmic outrage deficit has been proposed, where gender discrimination in hiring supposedly results in less moral outrage when discrimination occurs due to algorithmic versus human biases (Bigman et al., 2022). Although algorithms might be trained based on human judgements, this deficit is theorised to be due to a shift in attribution, where algorithms are not held accountable for discriminatory decisions since they are not prejudicially motivated, whereas human decision-making can be driven by stereotypes and prejudices (Bigman et al., 2022). Despite this, it is vital that there is at least one human who is responsible for the outputs of algorithms to ensure that there are mechanisms for accountability. A summary of the key findings of these studies can be seen in Table 1.

In summary

Algorithmic assessments represent a significant opportunity to improve candidate experience and streamline processes for employers. Candidate perceptions of selection procedures can influence their likelihood to accept a role, and thus the talent an employer has access to. Investigations into how algorithmic assessments are perceived have led to largely mixed findings, although assessments used earlier in the selection process are typically

viewed more favourably than those used later in the process. Moreover, the objectivity of algorithms can negatively impact perceptions of these tools as exceptions cannot be made for individual candidates, and there is less of an opportunity to form connections. However, these investigations have typically examined specific types of fairness perceptions isolation, rarely providing a holistic view of experiences with algorithmic assessments.

Table 1

Summary of the key findings of the described studies investigating procedural justice.

Study	Assessment tool	Participants and Procedure	Key findings
Georgiou & Nikolaou, 2020	Game-based SJT	73 employees of an IT company + 88 control that completed a gamified or standard SJT; 131 students/alumni of a South European university that completed a gamified SJT	• Higher levels of process satisfaction and organisational attractiveness for the game-based SJT compared to the traditional form • Higher levels of perceived fairness through process satisfaction
Suen et al., 2019	Video interviews	180 members of a non-profit HR organisation in China that completed synchronous or (AI) asynchronous video interviews	• No difference in the fairness perceptions of synchronous and asynchronous video interviews • No difference in the fairness perceptions of human versus algorithmic rater • Preference for synchronous interviews
Köchling et al., 2022	AI support in screening and interviews	160 German employees presented with hypothetical hiring scenarios	• Decreased perceived opportunity to perform when AI support used for telephone or video interviews • No effect of AI support in earlier screening stages on perceived opportunity to perform
Langer et al., 2019	Video interviews	123 German participants presented with videos of others completing automated or conference interviews	• Automated interviews in the selection context are associated with less perceived behavioural control and lower levels of acceptance through lower social presence compared to automation in low-stakes contexts or synchronous video interviews
Mirowska & Mesnet, 2021	CV screening tool and video interview	Interviews with 33 French professionals about a presented scenario	• Interviews found that participants accepted algorithms to be more objective than humans but preferred human judgements, despite them being prone to bias
Kaibel et al., 2019	Application screening tool; online assessment and video interview	165 German employees presented with a fictitious scenario; 255 American MTurk workers who completed an online test and digital interview	• Across all of the tools, perceived opportunity to perform and social presence was lower for algorithmic judgements compared to human judgements
Lee, 2018	Screening tool	228 American MTurk workers presented with scenarios about hiring as well as work scheduling, work assignment, and work evaluation	• Human hiring decisions are judged as fairer than algorithmic as algorithms lack human intuition and cannot make exceptions

Neurodiversity

Neurodivergence is an umbrella term that describes differences in thinking and cognition, although there is a lack of an agreed-upon definition (Doyle, 2020). As a non-medical phenomenon, the conceptualisation of neurodivergence has been influenced by biodiversity or biological differences, social disability movements, and insights into cognitive functioning from psychology (Doyle, 2020; J. A. Hughes, 2021). In general, neurodiversity can be thought of as a continuum of differences, and the presentation and symptomology of conditions can vary between individuals (British Psychological Society, 2021). However, whereas neurotypical individuals typically show a flat profile on measures of specific cognitive abilities (e.g., verbal skills, working memory, visual skills, and processing speeds) where scores are similar across all abilities, neurodivergent individuals typically show a spiked pattern, scoring highly on some abilities and low on others (Weinberg & Doyle, 2017).

Neurodivergence encompasses many conditions, including attention deficit hyperactivity disorder (ADHD), dyslexia, and autism spectrum disorder (autism). These conditions are all thought to be developmental, meaning that they are present from birth and symptoms develop throughout childhood and adolescence (Weinberg & Doyle, 2017). Whereas ADHD and autism have more noticeable behavioural symptoms and are typically diagnosed through health services, dyslexia is associated with educational and practical challenges and is therefore more commonly diagnosed by psychologists and occupational therapists (Weinberg & Doyle, 2017). It is estimated that 10% of the population of the UK has dyslexia (Central Digital & Data Office, 2017), 1% have autism, and 10% have ADHD (NHS Digital, 2014). These figures are in line with global prevalence, with 15-20% having dyslexia (International Dyslexia Association, 2016), just under 1% having autism (Baxter et al., 2015), and around 5% having ADHD (Polanczyk et al., 2007). Accordingly, around 1 in 5 people are neurodivergent (Doyle, 2020). It is estimated that having ADHD increases the

chances of having dyslexia fourfold (Wagner et al., 2019), and the prevalence of ADHD in those with autism is around 40% (Rong et al., 2021). As such, some practitioners refer to individuals with these conditions as having a particular neurotype, such as a dyslexic neurotype, instead of someone who has a particular condition (e.g., someone with dyslexia) to account for those who have comorbid conditions or multiple neurotypes (McDowall et al., 2023).

However, there is growing debate around the whether the prevalence of autism is underestimated. Originally developed by Allison et al. (2012), the AQ10 is a short, 10-item measure designed to aid referral decisions for a full autism diagnostic that asks individuals to report the extent to which they agree with each of the 10 statements. Scores are calculated by binarising responses to the scale, giving a maximum score of 10. In the initial creation and validation of the measure, a cut-off score of six represented that an individual should be referred for a full diagnostic assessment (Allison et al., 2012). However, the National Institute for Health and Care Excellence, the public body of the Department of Health and Social Care in England, which publishes guidance on treatment and diagnosis, incorrectly recommended to use a cut-off score of seven for the assessment, meaning that some individuals with a score of six that should have been referred for a full assessment over the past 10 years have not been (Waldren et al., 2021, 2022). As such, the prevalence of autism, in England in particular, is likely to be higher than current statistics indicate, with estimates suggesting that autism prevalence could be almost twice as high as official figures report (O'Nions et al., 2023). Either way, neurodivergent individuals make up a considerable proportion of the population.

Neurodivergence as a Protected Characteristic

Many individuals with these conditions are still able to participate in things like employment with the help of accommodations, with 22% of those with autism and around 27% of those with specific or severe learning difficulties in employment (Office for National Statistics, 2021b). However, since being neurodivergent can come with many barriers to

participation in activities such as work, the term neurominority is increasingly being adopted to describe those who are neurodivergent in acknowledgement of their disadvantage with a range of life outcomes (Doyle & McDowall, 2022). Indeed, classed as disabilities, ADHD, dyslexia, and autism are protected attributes under equal opportunity laws around the world, meaning that individuals with these conditions are protected from employment discrimination based on their disability status. These protections are enforced by the Equality Act of 2010 in the UK and the Americans with Disabilities Act (ADA; 1990) and the ADA Amendments Act of 2008 in the US, for example. Under the ADA, employers are prohibited from denying applicants employment based on their disability status if they can perform the essential functions of the position, with or without accommodations. Reasonable accommodations can include ensuring that facilities are accessible to those with disabilities, job restructuring, modified schedules, or the modification or acquisition of equipment (ADA; 1990).

While the ADA (1990) signposts some of the forms that so-called reasonable accommodations can take, the accommodations required by each person are individual. Nevertheless, common physical adjustments can generally be grouped into those that address auditory stimulation, visual stimulation, social stimulation, and the provision of resources (Weber et al., 2022). There is, however, a lack of robust evidence on the accommodations that routinely alleviate barriers in the workplace for neurodivergent employees (Weber et al., 2022). As such, it is important that an individualised approach informed by a workplace needs assessment is taken, where a qualified assessor identifies the weaknesses of the individual based on their examinations and assessments and recommends specific accommodations that would support the individual (Moody, 2015). Accommodations are typically not costly to implement, are often supported by co-workers (Schur et al., 2014), improve perceptions of inclusiveness in the workplace, and foster a more positive organisational climate and culture (Hartnett et al., 2011). Further, accommodations granted to

neurodivergent, or disabled, employees can often be beneficial to other members of the team if implemented widely (Bonaccio et al., 2020; De Beer et al., 2014; Hartnett et al., 2011; Leather & Kirwan, 2012).

Employers are required to meet requests for such accommodations to the best of their ability unless there is evidence that doing so would result in undue hardship or a significant financial impact, and failure to meet requests for accommodations can result in legal action (e.g., Abreu, 2018; Hensel, 2017; McEvoy, 1993). However, despite qualified individuals being entitled to such accommodations by law, there is often a reluctance for neurodivergent employees to disclose their condition to their employer (Bonaccio et al., 2020; Lindsay et al., 2021; Locke et al., 2017), typically out of fear of judgement, discrimination, or stigma (Lindsay et al., 2021). Accordingly, they can resort to implementing their own workarounds to support themselves without the need for employer intervention (Locke et al., 2017).

Moreover, despite the fact that disability is a protected characteristic, the Uniform Guidelines (EEOC, 1978) do not require adverse impact analysis based on disability as, like age, it is protected by other, specific legislation. This lack of adverse impact assessment is a particular concern for invisible disabilities, such as being neurodivergent, where it can be harder to observe whether a procedure has a discriminatory outcome. As such, the burden can be on applicants to prove that they have been discriminated against due to their disability, and some may not be aware that the procedure is discriminatory against them to initiate proceedings to begin with. Therefore, the lack of testing for bias against neurodivergent and disabled applicants could create a barrier to accessing employment.

Neurodiversity in the workplace

Given the considerable rates of neurodivergence in the general population, many employers will have a notable number of neurodivergent employees – whether they know it or not. However, much of the research into neurodiversity is centred around education and how learning can be supported at school, with much less focus on neurodivergence in

adulthood (Leather & Kirwan, 2012). Notwithstanding this, recent years have seen increased efforts towards understanding neurodiversity in the workplace, driven at least in part by greater awareness of the symptoms of these conditions, resulting in adults to recognising the symptoms in themselves and seeking adulthood diagnoses (London & Landes, 2021; Russell et al., 2022; Sayal et al., 2018; Solmi et al., 2022; Zhu et al., 2018).

As a result of this increased understanding, conversations are also emerging around the skills that neurodivergent employees can bring to the workplace, barriers to success, and effective accommodations. This is aligned with the notions of positive psychology, which studies the circumstances in which people or groups flourish (Gable & Haidt, 2005) and can be applied to the organisational context by focusing on how strengths can be unlocked and applied to promote positive workplace behaviours (Lorenz et al., 2016). Accordingly, there are increasing efforts focusing on how accommodations can be effectively introduced to support the performance of neurodivergent and other disabled employees, typically emerging in the form of diversity, equity, and inclusion initiatives. Indeed, large, global companies such as Google and Goldman Sachs have launched such initiatives, and a number of charities and organisations exist to support neurodivergent individuals in obtaining and maintaining employment, including the Access to Work scheme in the UK. In the following subsections, the impact of dyslexia, ADHD, and autism at work are discussed, as well as evidence for successful accommodations for each condition.

Dyslexia in the workplace

Dyslexia is a hidden disability that does not result in physically observable symptoms but is characterised by functional difficulties in school and the workplace (Doyle & McDowall, 2022). Although it is not generally diagnosed in a clinical setting, there is evidence for dyslexia having biological origins, including neuro-anatomical differences, genetics, and differences in brain activation (Erbeli et al., 2022; Maisog et al., 2008; Mascheretti et al., 2017; Ramus, 2014; Richlan et al., 2009; Shaywitz et al., 2006; Tamboer et

al., 2016). Dyslexia is characterised by impaired phonological skills, short-term memory, sequencing ability, and visuospatial skills, meaning that those with dyslexia can not only struggle with reading and spelling, but can also forget appointments, struggle with pronouncing long or complex words, take longer to process information presented to them, and have messy handwriting due to poor hand-eye coordination (Moody, 2010). These challenges can lead to anxiety and lower self-esteem surrounding specific tasks since their condition can present barriers to achieving something they are otherwise capable of intellectually (Jordan et al., 2014; Nalavany et al., 2018; Novita, 2016).

According to the model of adult dyslexic success derived from interviews with successful adults with learning disabilities including dyslexia, there are a number of factors that can contribute to workplace success in dyslexic employees (Gerber et al., 1992). This includes benefitting from social support, having a job that draws on individual strengths, and persistence and resilience in the face of adversity (Gerber et al., 1992). Indeed, many adults with dyslexia can go on to have successful careers (Leather et al., 2011) or go on to start their own businesses (Logan, 2009; Logan et al., 2008). The workplace achievements and performance of dyslexics are often supported by the unique strengths that they can bring to the workplace, including creativity, out-of-the-box thinking, and problem-solving ability due to their ability to mentally rearrange information and processes (Beetham et al., 2017; De Beer et al., 2014; Kannangara et al., 2018; Sauter & McPeek, 1993). This is particularly true when dyslexic employees are given accommodations in the workplace to unlock these strengths and remove unnecessary barriers.

Although the needs of each dyslexic employee differ and tailored support should be informed by a needs assessment, broadly, effective accommodations include coaching to support the identification of effective coping strategies (Beetham et al., 2017) or improve working memory and self-efficacy (Doyle & McDowall, 2019), additional time during

selection assessments and training, and technological aids such as text to speech software (Leather & Kirwan, 2012). Dyslexic employees could also be supported by using dyslexia-friendly formats when providing written information, including using a sans-serif font such as Arial (Evett & Brown, 2005; Rello & Baeza-Yates, 2013), and ensuring that the contrast between the text and background is not too high by using a cream or yellow background when using black text, where possible (Rello & Baeza-Yates, 2017; Rello & Bigham, 2017). Such accommodations are also likely to benefit other employees (Leather & Kirwan, 2012), particularly if the readability of information is improved (Evett & Brown, 2005).

However, despite the effectiveness of accommodations when provided, many are reluctant to disclose their dyslexia to their employer, often due to concerns about bias, stigma, or being seen as incompetent (Alexander-Passe, 2015; Beetham et al., 2017; Gerber & Price, 2008; Marshall et al., 2020; McLoughlin, 2015; D. K. Morris & Turnbull, 2007). Dyslexic employees, therefore, can instead resort to creating their own workarounds, including using dictation or read aloud software, adjusting font size, and using online spellcheck tools (Locke et al., 2017).

ADHD in the workplace

ADHD is an applied condition that is characterised by differences in behaviour or communication (Doyle & McDowall, 2022). Like dyslexia, ADHD is a hidden disability that presents in the form of three types of behaviour: inattention, disorganization, and hyperactivity-impulsivity. Generally, inattention and disorganisation present as difficulty staying on task and losing belongings, while hyperactivity-impulsivity presents as fidgeting and interrupting other people at a level that is disproportionate to age and developmental status (American Psychiatric Association, 2013). However, there are also sex differences in the symptoms of ADHD; males are more likely to display externalising symptoms and thus engage in disruptive behaviour, while females are more likely to show internalizing symptoms, having higher levels of anxiety and depression (Gershon, 2002; Levy et al., 2005;

Loyer Carbonneau et al., 2021) and being more likely to display symptoms of inattention. This could explain the higher rates of ADHD diagnosis in males than females (American Psychiatric Association, 2013) due to the diagnostic criteria being more aligned with male than female phenotypes.

Within the workplace, issues with executive functioning that come with ADHD can manifest in the form of being late, missing deadlines, and misplacing materials (Mao et al., 2011; Nadeau, 2005), as well as issues with attendance, general organisation, and interrupting others (Sarkis, 2014). Employees with ADHD can also struggle with teamwork, getting on with supervisors, and meeting their own expectations or perceived potential (Fuermaier et al., 2021). However, like those with dyslexia, adults with ADHD can go on to have successful careers and display a number of strengths in the workplace, such as increased creativity and problem-solving, resilience, and hyper-focus on tasks that are found interesting (Hoogman et al., 2020; Sarkis, 2014; Sedgwick et al., 2019; Steele et al., 2021; Weinberg & Doyle, 2017), which can help to mitigate issues with procrastination.

Since three broad behaviours categorise ADHD, accommodations can be targeted at each. For inattention, accommodations include being provided with a quiet room to concentrate, flexible working, and noise-cancelling headphones. On the other hand, hyperactivity can be aided through interventions such as ensuring that there are breaks in long meetings, and organisation issues can be aided by making use of calendar reminders, providing both written and verbal instructions, and providing well-structured notes (Adamou et al., 2013; Drehmer & LaVan, 1999; Mao et al., 2011; Nadeau, 2013), although the accommodations that each employee will benefit from are individual (Patton, 2009). Coaching has also been explored as a more personalised approach to supporting individuals with ADHD in the development of organisational and time management skills (D. R. Parker & Boutelle, 2009; Prevatt, 2016; Prevatt & Yelland, 2015). However, this is something that is

typically implemented in college, with a lack of research on the benefits of coaching for employees with ADHD (Sarkis, 2014).

Approached from a more medical perspective than dyslexia, there is evidence for biological origins of the condition, including genetics and differences in brain activation and structure (Bayerl et al., 2010; Dibbets et al., 2010; Elia & Devoto, 2007; Nakao et al., 2011; Sharp et al., 2009; Sigi Hale et al., 2007; Thapar, 2018). The dopaminergic reward system is also thought to play a significant role in ADHD symptomology (R. Stark et al., 2011), where individuals with ADHD do not receive adequate stimulation. As such, there is evidence that stimulants such as caffeine can help to sustain attention and focus by changing neural activity (Kahathuduwa et al., 2020). The symptoms of ADHD can also be managed through medication, with many effective treatments using stimulants to alter brain activity and activation, although this medication can result in side effects such as headaches (Cascade et al., 2010; Pan et al., 2022). However, like with dyslexia, there can be a reluctance to disclose ADHD to employers due to concerns about stigma and discrimination (Arnold et al., 2010; Masuch et al., 2019; McIntosh et al., 2023; Thomas et al., 2022), which can reduce access to useful accommodations.

Autism in the workplace

As with ADHD, autism is an applied neurodevelopmental condition (Doyle & McDowall, 2022) but is characterised by differences in social communication and interaction that persist across multiple contexts, which can lead to impaired social relationships. Autistic individuals also display repetitive behaviours, interests, or activities (American Psychiatric Association, 2013). Previously, autistic disorder, Asperger's disorder, and pervasive developmental disorder were all considered separate disorders but have since been consolidated in the Diagnostic and Statistics Manual Version Five (American Psychiatric Association, 2013) into a continuum representing the severity of symptoms in relation to social communication and restrictive repetitive behaviours. As with dyslexia and ADHD, it

has been proposed that autism has a biological basis, and there is considerable evidence that it is a highly heritable condition (Gaugler et al., 2014; Geschwind, 2011; Muhle et al., 2004; Thapar & Rutter, 2021), although there is not yet a consensus on the neural mechanisms or differences in brain activation that result in autism symptomology (Byrge et al., 2015; Harris et al., 2006; F. Zhang & Roeyers, 2019).

Despite the deficits that characterise autism, many autistic adults are still capable of successfully obtaining and maintaining employment, and autistic people are employed across various sectors (Baldwin et al., 2014; Lorenz et al., 2016). Roles associated with more repetition and less spontaneity generally suit autistic employees' ways of thinking better (Baldwin et al., 2014; Lindsay et al., 2021), but autistic employees also pay attention to detail and can concentrate on tasks for a long period of time (Lindsay et al., 2021). Positions in the technology industry can particularly draw on these strengths (K. R. Johnson et al., 2020). Autistic people are also often honest and efficient (Baldwin et al., 2014), methodical, punctual, and consistent (Hagner & Cooney, 2005). Additionally, they are creative and able to offer an alternative viewpoint due to their different way of thinking (Cope & Remington, 2022).

However, in the workplace, autism can manifest in the form of difficulty communicating with others and performing the required tasks, particularly if instructions are unclear or when vague language is used (Lorenz et al., 2016). Autists may also struggle with roles where there is a lack of routine (Lindsay et al., 2021). In the job application phase, over-specific job descriptions that are not a true reflection of the role can also present a barrier for autistic job seekers (Nagib & Wilton, 2020; Vincent & Fabri, 2022) due to the so-called black-and-white thinking style that characterises autism, where thoughts are typically binary with a lack of grey area in between (E. Stark et al., 2021). Consequently, if autistic job seekers do not meet all of the specifications outlined in the job description, this can lead them

to believe they are not qualified for the job, preventing them from applying to a position they may well be qualified for (Vincent & Fabri, 2022).

Nevertheless, there are a number of accommodations that can be implemented to support autistic people in the workplace, including a consistent schedule or routine, flexible working arrangements, direct communication and instructions, and providing opportunities for check-ins and to ask for more information (Baldwin et al., 2014; Hagner & Cooney, 2005; Lorenz et al., 2016; Waisman-Nitzan et al., 2021). However, as with dyslexia and ADHD, autistic employees can be reluctant to disclose their condition to employers due to concerns about stigma and discrimination (Huang et al., 2022; Lindsay et al., 2021; Romualdez, Walker, et al., 2021), or might selectively disclose to only certain people in the workplace (Romualdez, Heasman, et al., 2021), although those that do disclose often report that it had a positive impact (Romualdez, Heasman, et al., 2021), particularly in relation to workplace supports and accommodations (Romualdez, Walker, et al., 2021).

Using technology to support neurodivergent job applicants

While research into the support mechanisms that can be put in place for neurodivergent individuals is typically focused on education, emerging research has begun to examine how technology can be used to support neurodivergent job seekers. Although many of these interventions are autism-focused, a number of studies have examined the efficacy of training inventions to support performance during video interviews (Burke et al., 2018; Rosales & Whitlow, 2019; Smith et al., 2021). Others have used video-based interventions (Fontechia et al., 2019) and virtual reality (Bozgeyikli et al., 2017) to support vocational training. Given the potential of technology to support neurodivergent job seekers and the fact that the deployment of alternative assessment formats was accelerated by the pandemic (Strazzulla, 2020), this demonstrates that novel assessment formats could help to make pre-employment tests more accessible. Image-based assessments, for example, largely remove the language element, which could be beneficial for individuals with dyslexia and ADHD who

can favour visual ways of thinking (De Beer et al., 2014; Fassbender & Schweitzer, 2006). The more game-like nature of image-based assessments may also help to alleviate or reduce test-taking anxiety (Georgiou & Nikolaou, 2020; Mavridis & Tsiatsos, 2017), which can be especially beneficial for neurodivergent applicants who are more prone to test-taking anxiety than neurotypical populations (Lewandowski et al., 2015; Nelson et al., 2014, 2015).

In summary

Workplaces are made up of neurodiverse individuals that represent a rich pool of unique strengths. While neurodiversity is often studied in the context of education, progress is increasingly being made to research neurodiverse individuals in the workplace to understand their strengths and how they might be supported through adjustments and interventions. However, there is a lack of research into how neurodivergent individuals may be supported while completing pre-employment tests; the majority of research focuses on the training that can be delivered to prepare for them, rather than how pre-employment tests could be made more accessible. Notwithstanding this, algorithmically scored image-based assessments represent a potential avenue to improve the accessibility of pre-employment tests, particularly for applicants who are neurodivergent.

Chapter 2. Interdisciplinary bias mitigation: A worked example

I-O Psychology and Machine Learning on Test Bias: A Reconciliation and Work Agenda

Airlie Hilliard,[1,2] Nigel Guenole,[1] Franziska Leutner,[1] Cristian Munoz,[2] Emre Kaizm,[2,3] Adriano Koshiyama[2,3]
[1]Institute of Management Studies, Goldsmiths, University of London, New Cross, London SE14 6NW, UK
[2] Holistic AI, 18 Soho Square, London W1D 3QH, UK
[3]Department of Computer Science, University College London, Gower St, London WC1E 6EA, UK

Abstract

Industrial-organisational psychology and computer science are increasingly being applied alongside each other to create innovative assessments for screening job candidates. While these tools can increase the efficiency of the hiring process and enhance the candidate experience, combining artificial intelligence with psychometric assessments presents novel risks, including additional sources of bias. This is a concern for both psychologists and computer scientists. Although they differ in their approach to measuring and mitigating bias, they have the same end goals in mind and could benefit from each other's approaches. *This situation demands a reconciliation of perspectives so that practitioners across the fields can communicate with each other and with the organisations that make use of their respective skills.* In this article, we review the different approaches of psychology and computer science to identifying and addressing bias, focusing on the independence, separation, and sufficiency definitions used in computer science. We provide a worked bias mitigation example to explore the effectiveness of three bias mitigation approaches from computer science using data from an algorithmically scored image-based assessment of personality designed for use in selection, finding that the in-processing approach is most compatible with I-O psychology for this assessment. We end with a reconciliation agenda that recommends that greater alignment of the two fields could result from updated legislation specific to algorithmic recruitment tools, additional guidance from professional bodies, and additional training for practitioners.

Introduction

Industrial-organisational (I-O) psychology and computer science are two fields making important contributions to addressing bias in algorithmic recruitment tools. However, bias can have different meanings to psychologists and computer scientists who each take different courses of action when observing what they consider bias to be. This situation demands a reconciliation of perspectives in order that practitioners across the fields can communicate with each other and with the organisations that make use of their respective skills. The two views of bias, of course, are not mutually exclusive; there are computer scientists who are also psychologists, and psychologists who are also computer scientists. But for this article, where differences exist, the common perspectives of each will be set against one another to highlight similarities and resolve differences in interpretations of bias in algorithmic recruitment tools. It is also consistent with a real difference in the philosophy and approach between industrial psychology and computer science that is evident in the literature from each field.

This article advances the proposition that psychologists and computer scientists often have the same end goals in mind, and would profit from studying each other's approaches. A reconciliation of what each field means by bias is a prerequisite for this to occur. Specifically, we provide an overview of how bias is defined in I-O psychology, focusing on bias in the context of algorithmic recruitment tools (Tay et al., 2022), and a number of common ways to mitigate bias, covering techniques applicable to the majority of selection assessments as well as those specific to algorithmic tools. We then provide an overview of the several definitions of bias/fairness that are used in computer science (Verma & Rubin, 2018) before narrowing the focus to three more consolidated definitions of bias: independence, separation, and sufficiency (Barocas et al., 2023; Barocas & Hardt, 2017). Noting that independence is most aligned with psychology's notions of bias, particularly the four-fifths rule for measuring

adverse impact, we present a bias mitigation worked example that examines the effectiveness of three computer science mitigations on reducing subgroup differences in scores on an image-based assessment of personality (Hilliard, Kazim, et al., 2022a). Given that, in the worked example, two out of the three examined mitigation approaches lacked compatibility with I-O psychology, we end with a reconciliation agenda for how the two fields could learn from each other and more effectively collaborate in order to effectively mitigate bias in algorithmic recruitment tools. Key recommendations include greater education on both sides, updated laws that reflect technological advances and developments in best practices, and the education of policymakers to ensure that legislation targeting algorithmic tools is effective and actionable.

Algorithmic recruitment and ethical AI

In recent years, recruitment has become increasingly interdisciplinary, combining I-O psychology with computer science to deliver artificial intelligence (AI) and machine learning driven recruitment tools, which are being implemented in industry in a number of different ways. For example, to identify and aggregate qualified candidates (HireEZ, Arya, AmazingHiring), engage with applicants through chatbots (Paradox's Olivia, ThisWay, Brazen), analyse video interviews (HireVue, MyInterview, Gecko), and assess candidates through game- and image-based assessments (HireVue, Traitify, pymetrics). The uptake of these tools was accelerated by the pandemic, which saw demand for asynchronous video interviews soar (HireVue, 2021).

However, these tools are not without their risks, particularly since algorithms can perpetuate and amplify existing biases (Lloyd, 2018) and even seemingly small amounts of bias can compound to have large effects (Hardy et al., 2021). Perhaps the most well-known example of this is Amazon's aborted recruitment tool, which was found to be biased against female applicants and was therefore retired prior to deployment. The algorithm was trained to

evaluate candidates' resumes on data from previous applicants to technical positions at the company (Dastin, 2018), the majority of whom were male, reflecting the gender imbalance in the tech industry (PricewaterhouseCoopers, 2017). Since the algorithm was not optimised to evaluate female candidates, it penalised any resumes that contained the word "women's", and was, therefore, biased against female applicants who wrote about their membership to a women's team (Dastin, 2018).

Whilst Amazon is perhaps the most well-known example of an algorithmic recruitment tool being biased, it is not the only offender; LinkedIn and Facebook have also come under fire for bias in the algorithms used to display job adverts to users of the sites. LinkedIn's job matching algorithm, which is used to show users of the site suitable positions based on their profile, was found to be biased against females as it was optimised to show ads to users who were more likely to view the role and apply. However, males are typically more determined when it comes to seeking out new opportunities and are consequently more likely to click on the ad and apply (Beatrice, 2021; Wall & Schellmann, 2021), meaning that male users were being referred for open roles more often than female users. This has now been resolved through the introduction of an additional algorithm that ensures that there is a more balanced gender distribution in the targeting of job ads (Wall & Schellmann, 2021). Similarly, an investigation into the algorithms used by Facebook found that, for an identical position, females are less likely to be shown an advert for a position in a male-dominated company compared to a company with a more balanced gender distribution of employees. Even when qualifications are controlled for, this problem persists and is present across multiple roles and industries (Imana et al., 2021). It is also more expensive to display job adverts to females on Facebook, especially those aged between 25 and 44, since they are seen as a prized demographic (Lambrecht & Tucker, 2019).

These examples are just some of the several instances of bias present in algorithmic recruitment tools that have been exposed in recent years, highlighting the need for robust approaches to measuring and mitigating bias, particularly since recruitment is a high-risk and ethically critical context (Kazim et al., 2021). As this interdisciplinary approach to hiring becomes more widespread and possibly the norm, greater guidance is needed on how best to align the priorities of each discipline and ensure that the algorithms are being applied in a way that minimises potential harm. Accordingly, publications discussing concerns associated with the use of algorithms in recruitment have begun to emerge, with one of the major concerns being the ethical application of these systems in the high-risk context of recruitment, particularly surrounding how to ensure that the use of algorithms does not result in biased outcomes (Hunkenschroer & Luetge, 2022; Tippins et al., 2021)

Ethics in artificial intelligence

AI ethics is concerned with the psychological, political and social impact of AI (Kazim & Koshiyama, 2020), and research in this field focuses on the principles, policies and regulations that can be implemented to minimise the harm resulting from the technology (Jobin et al., 2019; Siau & Wang, 2020). AI ethics frequently draws on concepts from philosophy, particularly utilitarianism (the consequences of principles or rules), human and civil rights, and virtue (resulting from an individual's good character; Kazim & Koshiyama, 2020). Playing a prominent role in AI ethics is fairness, particularly with respect to bias (Kazim & Koshiyama, 2020). Here, bias refers to both treatment and impact, where the treatment of individuals by algorithms and the impact of these algorithms on individuals should be free from bias (Lipton et al., 2018).

AI ethics has recently started to gain traction and, in response, principles and frameworks have been developed by individuals or groups attempting to hold themselves (and others that follow them) accountable (Floridi et al., 2018; Floridi & Cowls, 2019; Jobin et al., 2019; Siau & Wang, 2020). Beyond this, there have also been calls for increased

governance in this space to further increase accountability since many of these principles are not put into practice due to issues concerning clarity, a lack of consensus, and a lack of interpretability (Kazim & Koshiyama, 2020). Towards this, something that is gaining traction is algorithm audits (Koshiyama et al., 2024; Raji et al., 2020), which can be defined as the practice of assessing, mitigating and assuring an algorithm in terms of legality, ethics, and safety (Kazim et al., 2021). Audits can be carried out by internal or external parties (Raji et al., 2020) and are increasingly being required by emerging AI laws.

Ethics in I-O psychology

Concerns about the ethical implications of AI in recruitment are also echoed by I-O psychologists. For example, concerns have been raised about the extent to which candidates can give informed consent about the collection and use of their data when using algorithmic recruitment tools since candidates might not know or be able to control what data is being collected about them (Tippins et al., 2021). To this end, laws have been passed in Illinois (Illinois General Assembly, 2020) and New York City (The New York City Council, 2021), requiring candidates to be informed about the information being collected about them and used during the evaluation of their performance. Additionally, Hunkenschroer & Luetge (2022) provide a review of ethical considerations relevant to algorithmic recruitment tools from a number of perspectives, including from a practitioner, legal, and technical perspective. As well as raising issues about consent, they also outline other concerns such as transparency about decisions and implications for holding developers, vendors and recruiters accountable. They also offer strategies to mitigate ethical concerns, taking into account governance, professional standards, and technical approaches.

Unlike the computer science field, where ethical frameworks are predominantly created by individuals or groups as opposed to being released by governing bodies, there is official guidance on ethical best practices that psychologists must follow. For example, the American Psychological Association's (2017) Ethics Code outlines best practices such as the

use of methods to reduce or eliminate bias resulting from assessments, and the interpretation of test scores in a way that does not result in discrimination. While this ethics code is not specific to algorithmic recruitment practices and instead outlines ethical principles for psychology as a whole, in their Framework, Tay et al., (2022) offer some guidance (although unofficial) on how machine learning can be applied in an ethical way to algorithmic psychological assessments, which they contextualise to recruitment. Specifically, they focus on Machine Learning Measurement Bias (MLMB), which they define as a machine learning model that performs differently for different subgroups, predicting individuals with the same underlying ability different scores due to their subgroup membership.

In addition, audit frameworks specifically for algorithmic recruitment tools have been developed, and they recommend that audits can be carried out in relation to verticals including bias, transparency (in the governance and decision-making procedures, and system explainability), safety or robustness (accuracy of the algorithm when applied in different contexts or to different datasets), and privacy (concerning the data the model was trained on or that is being used to compute an output; Kazim et al., 2021). Some progress has already been seen towards the auditing of algorithmic recruitment tools, with pymetrics (C. Wilson et al., 2021) and HireVue (O'Neil Risk Consulting and Algorithmic Auditing, 2020) both making audits of their tools public, and due to the mandating of audits for algorithmic recruitment tools (The New York City Council, 2021), we are likely to see more of this in the future.

While ethical frameworks and audits are a step in the right direction to ensuring that machine learning driven recruitment solutions are free from bias, real progress is difficult if there is a lack of consensus on what is meant by bias and how to mitigate it. Definitions of bias differ between computer science and I-O psychology, and even within each of these disciplines, perceptions of bias and approaches to mitigating it can vary. In the next sections,

we examine the approaches of I-O psychology and computer science to defining and mitigating bias, comparing the two disciplines and examining the compatibility of the two fields through a worked example before suggesting a reconciliation agenda.

Bias in (algorithmic) recruitment

I-O psychologists are bound by equal opportunity laws in the region they are designing and using selection tools in, although they may look to more prescriptive international laws such as those in the US when operating in regions such as the EU and UK that take a case-by-case approach (Hilliard et al., forthcoming). Indeed, in the US, psychologists are bound by several equality opportunity laws, including the Civil Rights Act of 1964, which is enforced by the Equal Employment Opportunity Commission's Uniform Guidelines on Employee Selection Procedures (Equal Employment Opportunity Commission; EEOC, 1978). The Uniform Guidelines require adverse impact testing to be carried out to examine differential hiring rates based on sex and race/ethnicity and endorse the four-fifths rule of thumb, which says that the hiring rate of one group should not be less than four-fifths of the rate of the group with the highest rate (EEOC, 1978). However, the Uniform Guidelines have had few updates since their publication in 1978 (Tippins et al., 2021), and concerns have been raised about the suitability of the four-fifths rule for measuring adverse impact since it is a rule of thumb (Cascio & Aguinis, 2001). Concerns have also been raised from a legal perspective about the difficulty in proving a test to be discriminatory due to the adverse impact of the overall selection process being prioritised over individual components, as well as the lack of guidance on measuring test fairness (Rubin, 1978).

Fortunately, psychologists also receive guidance on best practices from publications from professional bodies such as the Principles for the Validation and Use of Personnel Selection Procedures published by the Society for Industrial Organizational Psychology (SIOP; 2018). Here, fairness is considered a social concept that refers to equal group

outcomes, equitable treatment of all test takers, comparable access to the construct that the assessment measures, and lack of bias (SIOP, 2018). On the other hand, bias is defined as systematic errors in test scores that differentially affect different subgroups and can take two forms (SIOP, 2018):

- **Measurement bias** – occurs when irrelevant variance results in systematic group differences in scores and is a concern for both predictor and outcome variables. In other words, measurement bias occurs when group differences in scores do not reflect differences in ability.
- **Predictive bias** – occurs when the same regression line cannot be applied to all subgroups. Also known as differential prediction, this results in individuals who have the same ability but belong to different subgroups being given different scores by the measure. It is tested through moderated multiple regression using the predictor score, subgroup membership and the interaction between them, with differences in the slope or intercepts signalling predictive bias (Berry & Zhao, 2015).

However, this guidance often does not reflect the unique challenges posed by algorithmic tools; while the Principles (SIOP, 2018) mention tools such as game-based assessments, there is a lack of differentiation between algorithmic tools and traditional procedures. In light of this, SIOP has published guidelines on the validation and use of AI-based assessment in employee selection that are based on five key principles (SIOP, 2023):

1. AI-based assessments should accurately predict future job performance or other relevant outcomes.
2. AI-based assessments should produce consistent scores upon re-test that reflect job-related characteristics.
3. AI-based assessments should produce fair and unbiased scores.

4. AI-based assessments should be used appropriately with operational considerations in mind.
5. Decision-making related to AI-driven assessments should be adequately documented to facilitate verification and external auditing.

Moreover, despite the lack of updates to the Uniform Guidelines, the EEOC has recently released guidance on the use of AI in recruitment, particularly in regard to candidates with disabilities. This guidance outlines potential ways that the use of algorithmic recruitment tools could violate the Americans with Disabilities Act, which protects disabled individuals from unjustified discrimination. For example, failure to provide reasonable accommodations for those who need them would violate the Act, as would using algorithmic tools to screen for disabilities. Nevertheless, this guidance has not been codified in law and focuses on just one potential avenue for bias (EEOC, 2022). However, these publications do not provide guidance on how bias in algorithms should be measured and mitigated. Fortunately, Tay et al., (2022) have proposed their Framework for investigating and mitigating MLMB, when a machine learning model has differential functioning for different subgroups. The framework outlines two major sources of MLMB:

- **Biased data** – differences in subgroup performance on the ground truth, in behavioural expression, or feature computing.
- **Biased algorithm training** – where there is a differential relationship between features and the ground truth score for different subgroups in terms of differential weighting or transformation of predictors (Tay et al., 2022). Therefore, examining the relationship between the actual and predicted scores can help to determine whether subgroup differences are due to the model amplifying or diminishing genuine differences in ability.

In the absence of genuine subgroup differences, MLMB can manifest both in the form of differential relationships between the ground truth and predicted scores, and differences in the accuracy of the model for different subgroups (Tay et al., 2022). The manifestation of bias as a differential relationship is similar to the notion of predictive bias described by the Principles (SIOP, 2018); if the algorithm predicts different scores for individuals who have the same ground truth score but belong to different subgroups, this indicates that the predictors are used differentially for different subgroups, signalling that predictive bias is present. This bias can take the form of either compensatory (same mean scores, different distributions) or non-compensatory (different mean scores) bias, therefore reflecting the bias that can occur with traditional selection procedures. These differences can be examined for the overall model (distribution level) by comparing the mean and variance of scores predicted for different subgroups with the same mean and variance on the ground truth measure or for a particular range of scores (score level) by comparing the scores predicted for individuals from different subgroups with the same ground truth score and repeating this for multiple ground truth levels (Tay et al., 2022).

Differential accuracy, when the model accuracy varies across different subgroups (Tay et al., 2022), can result in predictive bias since the same model does not equally fit different subgroups (SIOP, 2018). It can be observed by testing whether using different models for different subgroups improves the fit. However, within the computer science field, a number of metrics exist to measure the predictive accuracy of models. For categorical classifications, metrics such as accuracy, precision, recall, F1 scores and area under the curve can be used (Powers, 2020), with some of these metrics capable of being converted to metrics such as Cohen's *d* values and point-biserial correlation (Salgado, 2018), which I-O psychologists are more familiar with. When evaluating the accuracy of continuous outcomes, metrics such as

R^2, mean squared error, and correlation coefficients are typically used (Rosenbusch et al., 2021).

I-O psychology approaches to addressing bias

According to the Uniform Guidelines, when adverse impact occurs, another measure with less adverse impact must be found or evidence of the validity of the procedure must be provided as a justification for its continued use (EEOC, 1978). On the other hand, the approach of I-O psychology is to create models that allow for true group differences while minimising unjustified differences. The approach used to address bias depends on its source. While the following is not an exhaustive list of strategies (see Ployhart & Holtz, 2008), we provide an overview of some different approaches used in I-O psychology to reduce adverse impact, which can result from biased selection procedures, that are applicable to both traditional approaches and algorithmic. Indeed, when the source of bias in recruitment algorithms is the ground truth data or the way scores are used (e.g., using high cutoff scores), some of the traditional approaches to mitigating bias can be applied, including removing items associated with bias, providing that this does not greatly alter the representation of the construct being measured (Byrne & van de Vijver, 2010). Conversely, when the source of bias is the algorithm itself, additional steps can be taken, as will be explored in the following sections.

Mitigating bias in assessment procedures

More simple approaches to mitigating bias in recruitment include using an assessment associated with less adverse impact, only allowing those meeting some minimum requirements to apply, or increasing test-taking time. The order of assessments when multiple hurdles are used may also be altered, or the cognitive demands of measures might be minimised by using visual-based formats (Tippins, 2009). If the source of bias is socio-cognitive biases that manifest during observer ratings, a strategy for mitigating this could be to use a standardised or structured approach (Levashina et al., 2014) and comprehensively

train raters (Aguinis et al., 2009). Other sources of bias require more careful consideration in order to mitigate it; when bias is present in the form of differential item functioning, ensuring that the measure has a good distribution of items that favour each subgroup can reduce differential test functioning since the items balance out and mitigate bias at the test level (S. Stark et al., 2004). If bias is not mitigated at the test level, items associated with non-equivalence can be removed from the model, providing that they do not significantly alter the psychometric integrity of the scale (Byrne & van de Vijver, 2010).

Mitigating bias in scores and predictors

Other approaches to mitigating bias place more emphasis on the way that scores are handled or interpreted. Although it may seem logical to select candidates with the highest scores through a top-down approach, this can result in adverse impact if some subgroups routinely score lower than others. Thus, while this approach can maximise the utility of the selection tool, it can also result in bias. To overcome this, a within-group percentile approach has been proposed (Sackett & Wilk, 1994), whereby raw scores are converted to percentiles and ranked, with the top performers from each subgroup being selected via stratified selection. While this approach can overcome adverse impact, this method should not be used since it not only reduces the utility of the selection measure (Cascio et al., 1995), but can also be seen as token diversity, which can be viewed unfavourably (Rixom et al., 2022), and can be considered as illegal in the US under the Civil Rights Act of 1991.

Instead, cutoff scores can be used, where all individuals scoring above the threshold are considered suitably qualified for the position. Cutoff scores can either be related to the overall selection process, where applicants must reach a certain total score across multiple measures (compensatory model), or to individual assessments, where a cutoff score is set for each measure (multiple hurdles; SIOP, 2018). Under the Uniform Guidelines, cutoff scores should reflect the minimum ability necessary for adequate job performance and should consider the potential for adverse impact. I-O psychology best practices also encourage the

probability of a qualified person answering an item correctly or being able to eliminate incorrect responses, as well as the importance of each item, to be taken into consideration (Cascio et al., 1988). Other factors that may also be taken into consideration when deciding on appropriate cutoff scores are the difficulty of the assessment, whether multiple thresholds are needed, and the type and quality of the data available for use in determining these scores (L. Mueller et al., 2007). Careful consideration of appropriate cutoff scores can help to reduce adverse impact, with particularly high scores potentially resulting in subgroup differences if few members of minority subgroups exceed the threshold (S. Stark et al., 2004).

Closely linked to cutoff scores is banding, where individuals with scores within a specified range are treated the same (SIOP, 2018). For example, individuals scoring above the primary cutoff score might progress to the next step in the hiring process, those scoring between the primary and secondary scores may be retained but not immediately progressed until those in the above band have been dismissed, and those scoring outside of these bands would not progress any further (Cascio et al., 1995). The Principles (SIOP, 2018) note that there are multiple approaches to determining bands, one of which is based on the assumption that scores within a band are considered not significantly different from each other. The selection of individuals from within the band can take the form of random or non-random selection, where the latter approach is based on stratification, meaning that the demographics of individuals selected from the band reflect the subgroup distribution in the wider population (Cascio et al., 1995).

Rather than fixed bands, sliding bands can also be used. Here, the width of the band is retained (e.g. 10 points) but as individuals from the top end of the band are chosen and removed from the potential pool, the band retains the same width but slides to include lower scores to better reflect the iterative process involved in selecting candidates (Cascio et al., 1995). For example, if the initial band was from 87 to 97 but the only individuals with scores

of 97 and 96 were chosen and removed from the pool, the new band would be from 85 to 95. Like with fixed banding, individuals can be selected from the bands randomly or non-randomly, with non-random sampling increasing the diversity of hires and resulting in less adverse impact, while having little effect on the mean test score of those who are selected (Cascio et al., 1995). To choose the width of these bands, one approach considers the standard error of measurement and the standard error of the difference between scores (Cascio et al., 1995), although it is controversial as some suggest that the assumption that all individuals within a band do not significantly differ in terms of true scores is incorrect (Campion et al., 2001). Nevertheless, banding in general, particularly when narrow bands are used (Campion et al., 2001), can be an effective way of retaining the utility of a selection tool while simultaneously reducing adverse impact (Campion et al., 2001; Cascio et al., 1995). Moreover, banding is endorsed by the Principles (SIOP, 2018), and has been subject to little legal debate (Campion et al., 2001).

Another common approach to reducing the risk of adverse impact is broadening the criterion space to include non-task-related measures of job performance (Tippins, 2009), such as an individual's contribution to the social and organisational factors that enable others to perform (Murphy, 2009). This is because group differences in scores are often considerably larger than group differences in actual job performance (Murphy, 2009), suggesting that job performance is not a unidimensional concept. Indeed, a multivariate approach to assessing job performance has been suggested, whereby the criteria used to measure job performance is bifurcated to address both task-related and contextual performance. The majority of the research into predicting job performance focuses on the task performance dimension, using assessments of cognitive ability to predict performance despite the fact that it is widely acknowledged that cognitive ability assessments are associated with adverse impact. On the other hand, contextual job performance is better predicted by non-cognitive traits such as

personality (Murphy, 2009), which are associated with a much lesser risk of adverse impact (Hogan et al., 1996). Therefore, broadening the criterion space to include contextual measures of job performance can lessen the overall adverse impact of a selection procedure since measures associated with higher and lower adverse impact are combined (Tippins, 2009). A benefit of this approach is argued to be that instead of treating adverse impact as an afterthought, where the aim is to design the most valid approach possible and then mitigate any subgroup differences, the effectiveness of the measure is defined by both the validity of the approach and the associated adverse impact, allowing bias to be considered at the time of design (Murphy, 2009).

The goal of optimising multiple outcomes (i.e., quality of hires and adverse impact) of a selection process during its creation, as suggested by the multivariate approach outlined by Murphy (2009), is shared by the Pareto-optimal selection approach. Pareto-optimal selection enables multiple outcomes, as well as constraints such as cost per applicant and the proportion of applicants to be selected, to be considered when developing a selection system (De Corte et al., 2011). With a Pareto-optimal trade-off, there may be multiple combinations of predictor weightings that result in the same level of mean performance but only the combination that is associated with the lowest adverse impact will be the Pareto-optimal one. When the Pareto-optimal weighting is deviated from to improve one outcome, this will be to the detriment of the other. For example, a reweighing of predictors to increase the quality of hires could result in greater adverse impact if the optimal weights are deviated from (De Corte et al., 2007). Using this approach, I-O psychologists can generate a variety of trade-off scenarios, allowing them to make an informed decision on how the predictors can be weighted to ensure high-quality hires and little adverse impact, with De Corte et al. (2011) providing guidance for using a Pareto-optimal approach in both practical and research settings.

Addressing data bias

Data bias occurs when there are differences in the ground truth scores used to train the model, behavioural expression, or feature computing (Tay et al., 2022), and is, therefore, something that can be mitigated for algorithmic tools specifically. Data bias occurs when the outcome measure (e.g. self-reported personality) used to train the algorithm gives individuals with the same underlying ability but different subgroup memberships different scores or has differences in the intervals of scores of different subgroups (Tay et al., 2022). One source of data bias is the underrepresentation of certain subgroups within the training data. This results in the algorithm only being optimised to evaluate some subgroups and not being generalisable to the groups that are not well-represented in the data (Buolamwini & Gebru, 2018). To mitigate this, a diverse range of respondents should be sampled to ensure the training data is as representative as possible. Computer science can offer a resolution in this case by using over- and under-sampling (Junsomboon & Phienthrakul, 2017) to ensure that there is adequate representation of minority groups in the sample and that it is not dominated by majority subgroups.

As well as biases in the ground truth data, there may also be biases in the predictor data. While traditional approaches to measuring constructs through self-reports or observations produce structured data, the application of machine learning algorithms to measuring these constructs has led to the increased use of unstructured data that has no particular format, including text and video (Tanwar et al., 2015). However, when this non-traditional data is used in predictive models, it can result in data bias. For example, behavioural expression can vary by subgroup, with female candidates speaking faster and smiling, nodding and varying their volume more than male candidates in interviews (Frauendorfer & Mast, 2014); older candidates maintain better eye contact during video interviews than younger candidates (McColl & Michelotti, 2019); and White candidates maintain greater eye contact than Black candidates (Fugita et al., 1974). Thus, even if

subgroups have the same ground truth distribution (the mean and variance of their ground truth scores are equivalent), their behavioural expression may differ, meaning that for some subgroups, there may be a stronger relationship between behavioural expression and the construct (Tay et al., 2022), which can lead to differential functioning. To prevent this, caution should be taken when including behaviours that are known to vary between subgroups in the model, and their weighting should be considered.

Related to this is the issue of feature computing, where there can be non-equivalence in the features used to compute the scores of different subgroups. The extraction of these features usually relies on another algorithm, which itself may be biased (Tay et al., 2022). For example, the inference of an individual's traits from video interviews relies on voice recognition to transcribe their responses and sometimes facial recognition and analysis software to interpret facial expressions. This is problematic as there are disparities in the accuracy of voice recognition (Bajorek, 2019) and facial recognition software (Buolamwini & Gebru, 2018) for different groups. Indeed, the accuracy of Google's voice recognition software, which is supposedly among the most accurate (Palanica et al., 2019), varies with accent (Tatman, 2017; Tatman & Kasten, 2017) and is less accurate for female (Tatman, 2017; Tatman & Kasten, 2017) and non-White users (Tatman & Kasten, 2017). The accuracy of text classification tools also shows bias against Black individuals (Blodgett & O'Connor, 2017). Further, facial recognition software, which may be used to extract non-verbal communication from video interviews (Kassab & Kashevnik, 2024), is less accurate for female and darker-skinned individuals (Buolamwini & Gebru, 2018) and an incident with Google's image classification algorithm resulted in offensive classifications of photos of Black individuals (The Verge, 2018). Moreover, commercial gender recognition tools are the most accurate for lighter-skinned males and least accurate for darker-skinned females, with error rates of up to 34.7% and 7.1%, respectively (Buolamwini & Gebru, 2018).

While much of the foundational research highlighting the disparities in the performance of these tools for demographic groups is almost a decade old, more recent research indicates that the issue is still pervasive and has not been widely resolved with advancements that have been made with the technology. Indeed, transcription software has still been demonstrated to be less accurate for non-native English speakers when transcribing English language (Dubois et al., 2024) and for non-White speakers compared to White (Wassink et al., 2022; Zolnoori et al., 2024). Moreover, commercial facial recognition tools still demonstrate disparities in their accuracy for males and females, although it has been suggested that differences in hairstyles could be driving this disparity since the tools may not be sufficiently optimised to evaluate female hairstyles (Bhatta et al., 2023). Similarly, while disparities still exist in the accuracy of facial recognition tools for lighter-skinned versus darker-skinned individuals, recent research indicates that this could be due to exposure differences in images of darker-skinned and lighter-skinned individuals, with the algorithms better optimised to evaluate over-exposed images that are more common with lighter-skinned individuals (H. Wu et al., 2023).

Another issue is the digital divide, with non-White individuals being more likely than White to only have access to the internet through their mobile data while at home (Tsetsi & Rains, 2017), which can lack reliability compared to broadband (Baltrunas et al., 2014). Non-Whites are also more likely than Whites to only have access to the internet at home via a smartphone (Fairlie, 2017; Tsetsi & Rains, 2017), which can lack the capabilities of a laptop or computer (Fairlie, 2017). As a result, individuals belonging to minority subgroups are more likely to have missing or incorrectly classified data, thus affecting the feature computing of these subgroups. The Framework (Tay et al., 2022) suggests that to mitigate this, features associated with bias should be removed from the model. However, since the

majority of the commercial tools available are either designed to be accessed via a mobile device, this reduces the impact of disparities in access to technology.

When removing features from the model, caution should be taken to ensure that the debiased algorithm retains its accuracy (Verma et al., 2021) and psychometric integrity. The removal of features would likely have a greater impact on traditional measures that only use a small number of predictors as they can have a larger influence compared to machine learning models that utilise hundreds of predictors. Nevertheless, the performance of the algorithm should be examined when mitigating bias to ensure that both bias and accuracy can be balanced. An applied example of the removal of features that could potentially result in bias comes from the concerns raised about the non-verbal features used in HireVue's video interviews after a complaint was made raising concerns about how analysis of non-verbal features such as facial expressions and tone of voice might impact those with disabilities (Electronic Privacy Information Center, 2019). Since advances in natural language processing meant that the performance of the model could be maintained even in the absence of facial features (Zuloaga, 2021), the decision was made to exclude these features from the model altogether. However, this approach lacks sophistication and could be an area that computer science can make valuable contributions to, as will be discussed in the section below.

Addressing algorithm training bias

In the absence of biased training data, bias can still occur during the training phase if the model uses different features for different subgroups. For example, if the algorithms used to score video interviews use features such as pitch and pause duration for one subgroup and features such as word count and loudness for another subgroup, this would constitute differential feature use (Hickman, Bosch, et al., 2021). This is equivalent to different algorithms being used for different subgroups (Tay et al., 2022). This type of bias can be examined by looking at the feature list for different subgroups and can be mitigated by retraining the model to use the same features for all subgroups if disparities are found (Tay et

al., 2022). Algorithm training bias can also occur when although the same features are used for all subgroups, there is differential weighting of these features for each group. Alternatively, the transformations used on the features may be non-equivalent for different subgroups (Tay et al., 2022). This can occur when the transformation relies on the distribution of scores and this distribution is not equivalent for each subgroup. For example, normalisation uses the mean and standard deviation of scores to make the transformation, meaning that if the subgroups are normalised separately, this could result in different transformation functions, thus violating the Civil Rights Act of 1991. To avoid this, when transformations are used, they should be applied identically to all subgroups to avoid bias. Additionally, feature ablation can be used to remove predictors from the model to examine their influence on outcomes (Kerz et al., 2022; Kumar et al., 2018) and can be applied to identify the predictors that are associated with subgroup differences (Tay et al., 2022).

Bias in computer science

Within the computer science field, fairness and bias are not as clearly conceptualised as in the I-O psychology field. Indeed, Verma & Rubin, (2018) identified 20 different definitions of bias in the field, with some definitions being known by multiple names. The lack of a consolidated approach to how bias is defined in computer science can become problematic, particularly since different definitions can be incompatible with each other (Kleinberg et al., 2016). For example, fairness through awareness and fairness through unawareness are opposites, with the former using protected attributes to ensure fairness and the latter not including any information related to protected attributes to create fair models. Indeed, Verma and Rubin (2018) applied the 20 definitions to a classification model based on a German Credit Dataset, where individuals were predicted to have either good or bad credit scores. They found a mismatch between the fairness definitions, where the model was fair according to some definitions and unfair according to others. This is problematic as it could

potentially allow developers and others to manipulate bias or fairness assessments since they would be able to choose metrics that give them favourable results.

As a result of the impractical number of definitions of fairness in computer science, efforts have been made towards grouping them into a small number of principles. Some key contributors to this are Barocas and Hardt, who suggested that the definitions of fairness definitions proposed in the field so far can be represented by three principal criteria: independence, separation, and sufficiency (Barocas et al., 2023; Barocas & Hardt, 2017). These definitions are likely something that will be unfamiliar to I-O psychologists not trained in computer science and rely heavily on statistical notation, which we contextualise to recruitment to aid understanding. Specifically, we use the example given by (Barocas & and Hardt, (2017) where an algorithm is being used to recruit a software engineer and is used to decide whether or not to display the job ad to an individual based on their browsing history. Here, X refers to features of the individual (data extracted from browsing activity), A refers to the sensitive feature (e.g. gender) bias is being minimised with respect to, Y is the target variable for prediction (whether or not someone is a software engineer), C is the classification (whether the ad is shown), and R is a score given to an individual to indicate how likely they are to click on the ad (Barocas & Hardt, 2017).

- **Independence** exists when the classification or prediction C (whether the ad is shown), is independent of the protected variable A (gender). This relationship is denoted $\mathrm{C} \perp \mathrm{A}$, and can be represented by the equation $\mathbb{P} = \{C = c|A = a\} = \mathbb{P}\{C = c|\, A = b\}$. This can be read as 'the probability of a candidate being shown the ad is equal for males and females'.

  This notion of fairness is similar to statistical parity and, outside of computer science definitions, is similar to the four-fifths rule, which essentially measures whether hiring rates are independent of subgroup membership. Because of the

similarity to the four-fifths rule, this definition of fairness is the most useful for I-O psychologists working with machine learning. However, this definition of fairness is limited since subgroup membership may be correlated with the target variable Y, meaning that when independence is forced, C is no longer a perfect predictor of Y (Barocas & Hardt, 2017). In other words, if subgroup membership is correlated with whether an individual is predicted to be a software engineer, then being shown the ad is not a perfect predictor of being a software engineer since individuals who are not software engineers may be shown the ad due to their subgroup membership, despite not being a software engineer. Further, independence permits what is referred to as laziness in machine learning, where high predictive accuracy in a majority group can mask low predictive accuracy in the minority group due to sample size discrepancies across groups (Barocas & Hardt, 2017).

- **Separation** constraints require that the predicted value R (likelihood of clicking on the ad) and the sensitive attribute A (gender) are independent conditional on Y (being a software engineer), denoted $R \perp A|Y$. This can be represented by $\mathbb{P} = \{(R = r \mid Y = y)|A = a\} = \mathbb{P} = \{(R = r \mid Y = y)|A = b\}$ and interpreted as 'the probability of an individual being predicted to click the ad or not is equal for males and females, conditional on software engineer status'. In other words, for both males and females, the probability of being predicted to click the ad is equal, providing that they are a software engineer.

  In contrast to independence, separation allows the predictor R (whether an ad is clicked), to be equal to the target variable Y (software engineer status), denoted $R = Y$. This assumes that software engineers will click on the ad and is known as optimal combability. It also means that the sensitive attribute A (gender) and the

target variable Y (software engineer status) can be correlated (which is useful due to the gender imbalance in tech), as well as the predictor R and sensitive attribute A. In addition, separation penalises laziness since the same true and false positive rates are required across groups, meaning that the error rates must be the same in all groups. However, this approach assumes that the target variable Y is reliable, but this is not always the case (Barocas & Hardt, 2017).

- **Sufficiency** exists when we do not need to know the sensitive attribute A (gender) to predict the criterion Y (software engineer status) given R (whether the ad is clicked) and is denoted $Y \perp A|R$. In other words, sufficiency occurs when R is predictive of Y independent of A (Barocas & Hardt, 2017). Thus, regardless of whether an individual is male or female, clicking on the ad is related to software engineer status.

  Although not strictly the same as fairness through unawareness, sufficiency does encompass this definition of fairness, where sensitive attributes do not need to be known or included in the model to predict an outcome. This definition is the most incompatible with I-O psychology notions of fairness since psychologists can use knowledge of protected attributes to inform their conclusions about the bias present in a tool; having knowledge of protected attributes enables adverse impact testing and examination of the factors that could be influencing subgroup differences.

Computer science approaches to addressing bias

Given that there are several definitions of bias in computer science, there are many more proposed ways to mitigate it. While in I-O psychology most approaches to mitigating bias are concerned with how the scores are used (e.g., using cutoff scores and banding) and the model used to score candidates (e.g. considering the weighting of variables and intercept and slope

equivalence), computer science approaches focus more on the training data and how predictors are used, with model constraints and data transformations being proposed to mitigate bias. In general, computer science bias mitigation approaches fall into one of three categories (Raghavan & Barocas, 2019):

- **Pre-processing** – Data is processed before being fed into the unmodified algorithm by removing sensitive attributes from or adjusting the features so that they are not correlated with subgroup membership. However, this approach relies on knowing the protected attributes and could be considered disparate treatment if the scores for different groups are adjusted. It also poses the question of how to treat or remove proxies of protected characteristics.
- **In-processing** – The model is optimised during training by introducing constraints that reduce the influence of protected attributes on model outputs.
- **Post-processing** – The model outputs are adjusted after training so that scores are not related to subgroup membership. However, this can be considered disparate treatment in I-O psychology.

To limit the scope of this article, we focus on the three fairness criteria (independence, separation, and sufficiency) proposed by Barocas and Hardt (2017) and give examples of how these definitions can be imposed through pre-processing, training constraints and post-processing approaches. Since independence is most relevant to I-O psychology, we elaborate on this concept the most and provide a summary and key references for each approach in Table 2. We also provide a worked example of bias mitigation using a real-life dataset, applying the pre-processing approach Learning Fair Representations (Zemel et al., 2013), in-processing approach Prejudice Remover Regularizer (Kamishima et al., 2011), and post-processing approach Equalized Odds (Hardt et al., 2016), all of which are explored below.

Achieving independence – Compatible with recruitment

Recall that independence occurs when the classification C is independent of the protected attribute A. To remove protected attributes and the features that encode them from the training data, many of the approaches to achieving independence are implemented during the pre-processing phase. Zemel et al. (2013) propose a learning algorithm to achieve independence, which is aimed at facilitating fairness both at an individual level (where similar individuals receive similar scores) and at a group level (where the proportion of individuals receiving a positive classification is equal across different subgroups). Their algorithm is used to create a representation of the original data that maximises its similarity to the original form while minimising the inclusion of protected characteristics (subgroup membership) by creating clusters of datapoints that have equal proportions of data from each subgroup. The new data, which is said to be an intermediate representation, can then be entered into a machine learning algorithm as the basis for prediction. Building on Zemel et al.'s (2013) approach, Louizos et al. (2016) proposed a deep variational autoencoder to encode the data, separate the sensitive attributes from the other datapoints, and penalise the resulting data using a maximum mean discrepancy term to limit differences in distributions of the subgroups to achieve independence.

Others have proposed that a chain of conditional models can be used to transform the features in the training data. Lum and Johndrow (2016) describe two approaches; univariate adjustment, where one feature is changed at a time, and multivariate adjustment, where multiple features are changed concurrently. Through these conditional models, a new set of variables that are independent of the protected attributes is created. Count variables (e.g., number of promotions, years of experience) are adjusted using Poisson or binomial regression, linear variables using linear regression, and binary variables using logistic regression, resulting in similar accuracy but less bias than the unmitigated model. Others take a more specific approach and focus on language, proposing how bias in word embeddings

can be mitigated since many words are associated with gender stereotypes. Giving the example of man is to computer programmer as woman is to homemaker, Bolukbasi et al. (2016) propose algorithms to debias these embeddings. Through this approach, neutral words, such as computer programmer or homemaker, become independent of gender associations by making the distance between the word and each gender equal, while gender-appropriate words, such as king and queen, retain the correct embedded meaning. Adjusting the features entered into the model could be of value to psychology as it can potentially overcome the need to remove features in the model associated with bias, reducing the risk of altering the construct being measured.

In contrast to the previous approaches, which have all been pre-processing, Kamishima et al. (2011) propose a regulariser that can be applied during the training phase of a classification algorithm. The regularisation parameter, eta (η), becomes larger when more sensitive information is used to make a prediction, resulting in the influence of sensitive attributes on the final prediction being reduced. This approach, therefore, enforces independence between the classifier and sensitive or protected attributes. However, since the regulariser is designed to enforce independence by sacrificing correctness in prediction, there is still a trade-off between accuracy and fairness, where the fairer algorithm is less accurate. A recent approach by Rottman et al. (2023) aims to overcome the diversity-validity trade-off in their comparison between iterative predictor removal and their bias penalisation technique. While the iterative process removes predictors associated with high subgroup differences and low predictive validity, the bias penalisation technique adds a bias penalty to the model optimisation, preserving high model accuracy while minimising bias.

Achieving independence – Incompatible with recruitment

Other approaches are less compatible with I-O psychology. For example, Feldman et al. (2015) proposed an alternative method for transforming the input data in a way that preserves unprotected characteristics while removing associations with protected attributes.

Echoing the within-group percentile approach – which is illegal to use in recruitment in the US under the Civil Rights Act of 1991 that made it unlawful to adjust scores or use different cutoff scores based on protected attributes – this approach proposes that the data can be repaired by ranking individuals in each subgroup and moving these rankings towards a median distribution for all groups. Their repair process can be implemented fully, or a partial repair can be carried out to balance accuracy and fairness.

Likewise, Calders et al.'s (2009) massaging and reweighing approaches lack alignment with I-O psychology. The former is used to change the labels of the dataset so that individuals given a negative outcome (e.g., those who are not recommended for interview) by the algorithm but are ranked highly (were among the best performing of the rejected) are then given a positive label (recommended for interview). The opposite approach is used for the lowest-ranking individuals of the positive class (the lowest performing among those who are recommended to be hired), where these individuals are given a negative label. This procedure is applied for each group of the protected characteristic (i.e., for males and females) to ensure that there are equal proportions of successes and failures in each group. The models are then retrained on the relabelled data. The latter approach, reweighing, gives individuals who are classified as having a positive outcome (being recommended for interview) a higher weighting than individuals with a negative outcome (not being recommended for interview) in each group. By reweighing the datapoints in each group of the protected characteristic, the approach aims to balance the number of positive and negative classifications for each group by creating a balanced dataset through sampling the original training data and replacing it according to the assigned weights. However, both approaches could constitute disparate treatment and violate the Civil Rights Act.

In subsequent work, Kamiran and Calders (2012) propose a sampling technique to balance the training data, either removing or duplicating datapoints. Through uniform

sampling, all of the data in the same group – based on classification and subgroup membership – has the same chance of being duplicated or removed. Through preferential sampling, datapoints closer to the boundary have a greater chance of being duplicated or removed. Here, datapoints in the majority subgroup (White) with a negative classification and those in the minority subgroup (Black) with a positive classification that are closest to the border are duplicated, while majority group members with a positive classification and minority with a negative classification may be removed. This approach aims to balance the training data so that there is a more balanced distribution of the number of individuals in each subgroup receiving positive and negative classifications. However, manipulating the data in this way is not good practice in psychology since the duplication of datapoints makes the dataset more homogenous, meaning that the model could lack generalisability beyond the training set.

Achieving separation

Recall that separation occurs when the predicted value R and the sensitive attribute A are independent conditional on the target variable Y. While many of the approaches to achieving independence occur in the pre-processing phase, with the input data being adjusted to be independent of the protected attributes, proposed methods to achieve separation are less constrained to this phase of development. Indeed, Hardt et al. 2016) propose a post-processing approach known as thresholding, which is used to discretise the predicted score. By examining the receiver operating curve, which represents the accuracy or true and false positive rates of a model, different thresholds can be established to categorise individuals into the positive and negative classes for each subgroup to ensure that the true and false positive rates for each subgroup are the same. In other words, the threshold that determines whether an individual is classified positively or negatively is adjusted for each group, resulting in sufficiency since the predicted score and protected attribute are independent, conditional on the predicted score. While this is a simple approach to ensuring equalised odds and

separation, this approach does not follow best practices in I-O psychology since subgroups should receive equal treatment and therefore not be judged by different criteria. However, in line with the limitations of the approach recognised by Hardt et al. (2016), Woodworth et al. (2017) argued that this approach is sub-optimal and proposed an alternative two-step framework for achieving sufficiency; they suggest combining a training constraint and a post-processing approach, where the post-processing approach further reduces bias and corrects unfair outcomes not addressed by the training constraint. They recommend using one portion of the training data in the training phase and the second portion of data in the post-processing phase, where a randomised predictor is created to impose fairness.

Achieving sufficiency

Since sufficiency is close to the concept of calibration by group, where R represents the probability of being assigned to the positive class of Y and this rule applies to all subgroups, one suggested approach to achieving independence is through Platt scaling. This is a post-processing approach where a sigmoid (s-shaped) function is fitted against the uncalibrated score to minimise the log loss and calibrate the score by group (Barocas et al., 2023; Barocas & Hardt, 2017). However, this lacks compatibility with I-O psychology since subgroups should receive equal treatment and there should not be different calibration or other transformation strategies applied for different groups (Tay et al., 2022).

Table 2

Pre-processing, training, and post-processing approaches to achieving independence, separation and sufficiency.

Technique	Phase	Fairness definition	Summary	Recommended for Recruitment
Disparate Impact Remover (Feldman et al., 2015)	Pre-processing	Independence	Scores are transformed so an individual's ranking in their subgroup is preserved for their ranking in scores overall	No – disparate treatment
Learning Fair Representations (Zemel et al., 2013)	Pre-processing	Independence	An intermediate representation of the data is created, using clusters of datapoints, that minimises similarity to protected characteristics while maximizing similarity to the original data.	Yes
Variational Fair Autoencoder (Louizos et al., 2016)	Pre-processing	Independence	An autoencoder and penalization parameter are used to create a new representation of the data that minimises similarity to protected characteristics while maintaining the unrelated information	Yes
Statistical Framework for Fair Predictive Algorithms (Lum & Johndrow, 2016)	Pre-processing	Independence	A new representation of an arbitrary number of predictors is created that is independent of the protected attributes using regression approaches	Yes
Debiasing Word Embeddings (Bolukbasi et al., 2016)	In-processing	Independence	Debias word embeddings to remove the gender association of gender-neutral words	Yes
Prejudice Remover Regularizer (Kamishima et al., 2011)	In-processing	Independence	A discrimination-aware regularization term eta is added to the model and becomes larger when more sensitive information is used in a prediction, lessening its influence	Yes
Multipenalty Optimization (Rottman et al., 2023)	In-processing	Independence	Bias penalty added to the model optimisation, to preserve high model accuracy and minimise bias	Yes

Classification with Independence Constraints (Calders et al., 2009)	In-processing	Independence	a) Individuals in each group of the protected attribute passing and failing are ranked and those who are borderline are allocated to the other classification (e.g., the lowest performers in the positive group are reallocated to the negative group) b) Datapoints are reweighed according to the protected attribute membership group and whether they were a positive or negative classification	No – disparate treatment
Reweighing (Kamiran & Calders, 2012)	Pre-processing	Independence	Datapoints in the training data are duplicated or removed to have a more balanced distribution of individuals receiving a positive and negative classification in each subgroup. The model is trained on this altered data.	No – leads to homogenous datasets
Learning Non-discriminatory Predictors (Woodworth et al., 2017)	In- and post-processing	Separation	Builds on the approach of (Hardt et al., 2016), one portion of the data is used to impose a fairness constraint in training and the other portion is used to generate a randomised predictor to impose fairness	Yes
Equalized Odds (Hardt et al., 2016)	Post-processing	Separation	Different thresholds are used for different subgroups to allocate them to the positive and negative classes to balance true and false negative rates between subgroups.	No – disparate treatment
Platt Scaling (Barocas et al., 2023; Barocas & Hardt, 2017)	Post-processing	Sufficiency	A sigmoid function is fit to an uncalibrated score so that scores are calibrated by group	No – disparate treatment

Bias mitigation worked example

To investigate the interdisciplinary approach to bias mitigation in algorithmic recruitment tools in practice, a pre-processing and in-processing approach to achieve independence – namely Learning Fair Representations (Zemel et al., 2013) and Prejudice Remover Regularizer (Kamishima et al., 2011) – were applied to data collected during the validation of an image-based assessment of personality that is scored using machine learning based predictive algorithms. Given that independence approaches are typically pre- or in-processing, Equalized Odds (Hardt et al., 2016) was selected as the post-processing approach despite being designed to enforce separation since it is a widely cited and well-known mitigation in computer science. The aim of this worked example was to i) investigate their impact on model performance, and ii) examine their effectiveness at reducing subgroup differences.

Data

431 respondents were recruited using Prolific Academic. The majority (n = 222) of respondents were female and most (n = 356) were under the age of 40. 209 were White, 73 were Black, 66 were Asian, 56 Hispanic, and 14 were of Mixed Race. Respondents completed the questionnaire-based measure of personality IPIP-NEO-120 and a 150-item image-based assessment of personality designed for use in selection that is described in greater detail in Chapter 3.

Scoring models

While the assessment, in practice, is scored using linear regression to generate continuous scores from binary image choices, since the computer science mitigation approaches were created for use with categorical outcomes, logistic regression was used for the purpose of this example. Here, the training data was binarised such that scoring above the median score for that trait was considered as passing and scoring below as failing. This is in line with the metric for continuous scores defined by the enforcement rules for New York

City Local Law 144, which requires bias audits of automated employment decision tools (DCWP, 2023; The New York City Council, 2021). The training data was examined for subgroup differences based on age, gender, and ethnicity using the four-fifths rule as it converges most closely with the notion of independence. This analysis indicated that there was a lower pass rate for males compared to females for emotional stability and Mixed ethnicity test-takers compared to Black test-takers for conscientiousness. Specifically, in the training data, males had an impact ratio of .77 and Mixed ethnicity .37, allowing the effectiveness of the mitigations to be examined for different severities of subgroup differences.

Mitigations

Learning Fair Representations (Zemel et al., 2013) is a pre-processing approach that transforms the training data to reduce the influence of protected attributes on outcomes, meaning that the data fed into the model using this mitigation is not identical to the raw training data. In this example, the raw training data represents binary image choices, where a value of 1 indicates that the image was selected by the test-taker and 0 indicates that it was not. As a result of applying Learning Fair Representations, this data was transformed such that the data became floats, or continuous values.

Prejudice Remover Regularizer (Kamishima et al., 2011) is an in-processing approach that places constraints on the model itself to reduce the influence of protected attributes on predictions, meaning the unchanged raw training data (binary image choices) is fed into the model and used to make predictions.

Finally, Equalized Odds (Hardt et al., 2016) is a post-processing approach that changes the predicted scores based on subgroup membership to make the outcomes more similar across groups in an attempt to reduce adverse impact.

Appendix A and Appendix B demonstrate how the input data, model coefficients, and outputs are changed for the baseline and mitigated model for the pre-processing, in-processing, and post-processing approaches, respectively.

Results

As can be seen in Table 3, the baseline model reflected the subgroup differences in the training data, resulting in an impact ratio of .36 for conscientiousness and .70 for emotional stability for the aforementioned groups. However, the baseline models performed well, exceeding .90 for all metrics (accuracy, balanced accuracy, precision, recall, and F1-Score). For the conscientiousness model, the pre-processing and post-processing mitigation approaches result in the mitigated model performing similarly to the baseline model. On the other hand, the in-processing negatively impacted the accuracy of the model, but the model can still be said to perform well. Similarly, for the emotional stability model, the post-processing resulted in a similar level of accuracy to the baseline model and the in-processing approach resulted in a dip in performance. However, the pre-processing approach considerably reduced model performance.

Both the pre-processing and in-processing approaches for the conscientiousness and emotional stability models resulted in an impact ratio above the .80 threshold for the four-fifths rule, indicating that subgroup differences were mitigated. On the other hand, the post-processing approval did not result in impact ratios above the .80 threshold, although the impact ratio did improve slightly for the conscientiousness model.

Table 3

Performance metrics and adverse impact ratio for the baseline model and pre-, in- and post-processing mitigation approaches.

Metric	Baseline model	Pre-processing	In-processing	Post-processing
Conscientiousness				
Accuracy	.92	.92	.85	.93
Balanced accuracy	.92	.92	.85	.93
Precision	.90	.96	.87	.90
Recall	.93	.89	.83	.95
F1-Score	.91	.92	.85	.92
Impact ratio	.37	.89	.99	.40
Emotional stability				
Accuracy	.90	.57	.85	.91
Balanced accuracy	.90	.57	.85	.91
Precision	.89	.54	.88	.91
Recall	.91	.57	.83	.92
F1-Score	.90	.55	.85	.91
Impact ratio	.77	.94	1.00	.75

Note. The adverse impact ratio for conscientiousness is for Mixed ethnicity test-takers compared to Black for conscientiousness and males compared to females for emotional stability.

Appropriateness of mitigations for recruitment tools

Although the in-processing approach resulted in a slight dip in the performance of the model in terms of how accurately personality was predicted, it did result in the subgroup differences being mitigated for both traits. While the pre-processing approach performed well for conscientiousness and resolved the subgroup differences for both traits, its suitability for the present example is limited since the data fed into the model no longer represented binary image choices due to the transformation of the data into floats. Finally, the post-processing approach was not compatible with reducing subgroup differences as measured by the four-fifths rule, and may also lack compatibility with equal opportunity laws, particularly in the US. Indeed, the changing of outcomes based on subgroup membership may be considered unlawful under the Civil Rights Act of 1991. As such, for this worked example, the computer

science in-processing approach to bias mitigation was the most compatible with algorithmic recruitment tools that use binary data to make predictions.

Increasing the compatibility of I-O psychology and computer science approaches to bias

A major difference between the approaches to bias of I-O psychology and computer science is that in psychology, definitions are generally agreed upon (Society for Industrial and Organizational Psychology, 2018), meaning that practitioners do not have to deliberate between multiple, incompatible definitions. On the other hand, in computer science, the notion of fairness that is abided by is subject to the discretion of those developing the model and trade-offs will be necessary to come to a decision about which approach to adopt since definitions are often incompatible; independence, sufficiency, and separation are mutually exclusive, except in degenerate cases (Barocas & Hardt, 2017). Further, notions of bias in I-O psychology typically focus on linear models instead of categorical outcomes, while the majority of the computer science literature on fairness focuses on categorical outcomes since many definitions examine true and false positive rates (Verma & Rubin, 2018). While pre-employment tests inevitably do result in a classification, whether a candidate is hired or not, the process of getting to this decision is often based on linear models. As a result, definitions of and approaches to mitigating bias in I-O psychology focus on score distribution and the way the scores are used or allocated to outcomes. For example, cognitive ability scores are usually continuous and are then converted to a binary outcome through cutoff scores, where those below the cutoff are considered qualified and those below unqualified.

Moreover, the computer science approach to bias is focused on structural relationships, with less attention being paid to how well constructs are being measured since optimising the model can be prioritised over social impact, signalled by the interchangeability of the terms fairness and bias. Indeed, computer science does not differentiate between bias in measurement and bias in structural relationships, therefore taking a more technical approach

to addressing bias. The reason for this is that machine learning has traditionally been applied in more objective contexts (Mullainathan & Obermeyer, 2017). For example, in the application of machine learning algorithms to a self-driving car, it is easy to label road markings, pedestrians and other obstacles since their presence is objective. As a consequence, measurement bias has previously not been a concern for computer science. However, in more subjective contexts, such as recruitment, the model is trained on human ratings, which lack objectivity (Mullainathan & Obermeyer, 2017), meaning that social factors must also be considered when identifying and mitigating bias. Indeed, while computer science can offer some valuable technical approaches to addressing bias that go beyond just the removal of features, it has been argued that such approaches fail to consider the influence of behavioural responses. This can be problematic as even when disparities in predictions are removed, there may still be underlying disparities in behaviour or ability that result in differential predictions (Shimao et al., 2021). As such, the computer science approach to bias is more of an optimisation and technical issue, rather than a social one since the reason for group differences is not acknowledged or investigated.

In response to this, there have been calls for a socio-technical approach to addressing algorithmic bias, where human and systematic biases are also considered, alongside technical sources of bias (R. Schwartz et al., 2022). For example, some have recommended additional training for I-O psychologists in machine learning techniques (Oswald et al., 2020; Oswald & Putka, 2020) and updates to the curriculum of I-O psychology programs to support this (Aiken & Hanges, 2015). Indeed, it is no longer adequate to just be trained in I-O psychology or machine learning when working in interdisciplinary teams using algorithms in the context of pre-employment tests. Whereas I-O psychology approaches focus on outcome parity and suitability of measures used, they lack sophistication in using modern statistical and machine learning methods to improve datasets and prediction or scoring models to reduce bias. On the

other hand, machine learning approaches are highly sophisticated and provide a useful way for reducing bias in the training data and outcome data but are not always compatible with psychology definitions or equal opportunity laws. As such, computer scientists working in the domain of pre-employment testing would also likely benefit from becoming familiar with the laws and principles that I-O psychologists are bound by since approaches that are common in computer science, such as reweighing variables or training multiple models for different populations, can be considered a source of MLMB in psychology (Tay et al., 2022). Psychologists should also be exposed to more sophisticated techniques for mitigating bias and improving predictions from machine learning, which is particularly useful since psychology has previously been grounded in explanation rather than prediction (Yarkoni & Westfall, 2017).

Some progress has been made towards providing the necessary guidance for the fields to converge, with publications released concerning how Big Data can be used within psychology. While these publications have made useful contributions to the field, including recommending that the collection and use of Big Data should be theory-driven (Landers et al., 2016), and providing guides on the management, processing and analysis of Big Data (Chen & Wojcik, 2016) and how to deal with privacy and consent issues (Guzzo et al., 2015), there is still no formal guidance available for I-O psychologists. Indeed, despite Guzzo et al.'s (2015) publication being recognised by the SIOP executive board, they note that their recommendations are not formal standards. Further, not all machine learning applications in psychology rely on Big Data; machine learning can be applied to smaller datasets to predict constructs based on psychometric measures that use purposely collected data. Consequently, guides on Big Data will not be fully applicable to these methods. While there are some publications that are intended to guide I-O psychologists using machine learning models not based on Big Data in their research (Putka et al., 2018), they are limited to statistical

approaches to analysis and do not give guidance on other practical considerations such as bias and adhering to equal opportunity laws.

Overall, the MLMB Framework (Tay et al., 2022) is a step in the right direction towards formalising how bias is defined and dealt with in algorithmic recruitment, but greater efforts are needed to provide better guidance to I-O psychologists and data scientists working in interdisciplinary teams to design, validate, and implements selection tools grounded in machine learning. As such, some key recommendations are:

- **Formal guidance for psychologists** – Actionable and more comprehensive guidance is needed from governing bodies such as SIOP and the British Psychological Society to formalise the definitions of and tests for bias in algorithms, and how best to mitigate bias while ensuring algorithmic assessments are valid and comply with equal opportunity laws. For example, although SIOP has published guidelines on AI-based assessments (SIOP, 2023), additional guidance could specify which of the methods listed in Table 2 are suitable to use for recruitment algorithms. Computer scientists would also benefit from a governing body to provide similar guidance.
- **Updated legislation** - The field would benefit from updated equal opportunity laws that address the last 40 years of advancements in measurement and selection. In a world that is increasingly automating tasks (McKinsey & Company, 2017), it is important that laws are up-to-date and compatible with the applications of technology. While there has been some progress towards this, particularly in the US, there is still a way to go since the existing legislation focuses on specific aspects of the use of automated recruitment tools.
- **Education and training for policymakers** – in order to ensure that the updated legislation is actionable and appropriate for algorithmic tools, policymakers must

be educated on the technology and seek input from relevant, interdisciplinary stakeholders on best practices. The development of the enforcement rules for New York City local law 144, demonstrated this need. While the metric for conducting bias audits is largely based on the four-fifths rule, the metric for continuous systems has caused controversy due to concerns about how suitable it is for automated tools and different data distributions (Filippi et al., 2023; Groves et al., 2024). As such, policymakers require additional training and input to ensure that laws targeting algorithmic tools are in line with perspectives from both I/O psychology and computer science.

Conclusion

The combination of psychological theory and machine learning has resulted in powerful tools being developed that can predict job performance and other psychometric characteristics from just a short interaction with candidates. This convergence has revolutionised hiring and has opened the market for innovative tools to assess applicants but also presents unique challenges for ensuring that pre-employment tests are free from bias. Indeed, algorithmic recruitment tools offer more opportunities for both bias and bias mitigation compared to traditional approaches. The outcome data, training data, and the algorithm training process are all potential sources of bias and at the same time offer opportunities for bias mitigation using machine learning and statistical approaches.

In this ethically critical domain, it is important to have robust strategies to identify and mitigate bias that draw on practices from both fields. This review of the approaches to identifying and mitigating bias in I-O psychology and computer science demonstrates that there can be a lack of alignment, with machine learning focusing on optimisation and model performance, while I-O psychology takes a social approach, examining the source of group differences and whether they are genuine. Despite this misalignment, IO psychologists and

computer scientists are increasingly working together in the context of recruitment, and progress is being made towards a more socio-technical approach due to new methods such as Rottman et al.'s (2023) bias penalisation approach or Bolukbasi et al.'s (2016) approach to debiasing word embeddings, which can overcome the need to remove potentially biased features from the model by adjusting them.

However, the unsuitability of some of the techniques proposed in the computer science field, such as Kamiran and Calders' (2012) proposed method of duplicating and deleting data, signals that further reconciliation of the two fields is needed to allow them to learn from each other. We propose that this reconciliation can be aided by increased education of both psychologists and engineers, greater formalised guidance, and updated legislation that is in line with current scientific knowledge and practices. The reconciliation of these fields will benefit all involved parties, allowing each field to learn from the other and increasing candidates' trust in these tools.

Chapter 3. Studies One and Two – Creating and validating an image-based assessment of personality

Measuring Personality Through Images: Validating a Novel Image-Based Assessment of the Big Five Personality Traits

Airlie Hilliard,[1,2,*] Emre Kazim,[2,3] Theodoros Bitsakis,[4] Franziska Leutner[1,4]
[1]Institute of Management Studies, Goldsmiths, University of London, New Cross, London SE14 6NW, UK
[2] Holistic AI, 18 Soho Square, London W1D 3QH, UK
[3]Department of Computer Science, University College London, Gower St, London WC1E 6EA, UK
[4]HireVue, London, UK

Abstract

Selection methods are commonly used in talent acquisition to predict future job performance and find the best candidates. Questionnaire-based assessments can be lengthy and lead to candidate fatigue and poor engagement, affecting completion rates and producing poor data. The gamification of assessments can mitigate some of these issues through greater engagement and shorter testing times. One avenue of gamification is image-based tests, and, although they are starting to gain traction commercially, there are few studies describing their validity and psychometric properties. As such, this study explores the potential of a five-minute, forced-choice, commercially-developed, image-based assessment of the Big Five personality traits that is scored using machine learning to be used in selection. Study One describes the creation of the item bank of image pairs and the selection of the 150 best-performing items based on a sample of 300 respondents. Study Two describes the creation of machine-learning-based scoring algorithms and an examination of their convergent and discriminant validity and subgroup differences based on a sample of 431 respondents. All models showed good levels of convergent validity with the IPIP-NEO-120 (openness $r = .71$, conscientiousness $r = .70$, extraversion $r = .78$, agreeableness $r = .60$, and emotional stability $r = .70$) and were largely free from potential adverse impact, echoing subgroup differences present in the training data. The implications for recruitment policy and practice and the need for further validation are discussed.

Introduction

This article explores the potential for measuring the personality of job applicants through their image choices using a novel measure whereby respondents are presented with pairs of images and asked to indicate which image in the pair is more like them. We report on two studies, where the first describes the creation and refinement of the measure and the second creates scoring algorithms and validates them by assessing convergent and discriminant validity and the potential for adverse impact. Although some image-based personality assessments exist commercially (e.g., Traitify and RedBull Wingfinder), there is little literature in this field. Since applicant perceptions of the selection process can influence whether an applicant is likely to accept a job offer (Hausknecht et al., 2004), applicant experience is pertinent for both applicants and hiring teams. Image-based formats might increase engagement and thereby improve applicant perceptions as they increase satisfaction and elicit stronger responses compared to questionnaire-based measures (Downes-Le Guin et al., 2012; Meissner & Rothermund, 2015). The purpose of this article is to address this lack of validation in the research regarding image-based assessments of personality, particularly those created for use in selection. Our findings provide preliminary evidence that assessments of this type could be a valid and fairer alternative to questionnaire-based selection assessments that use Likert scales.

The article begins with an overview of selection assessments, particularly those measuring cognitive ability or personality, followed by evidence in favour of the use of game- and image-based assessments, such as their shorter testing times (Atkins et al., 2014; Leutner et al., 2021). As will be highlighted below, much of the research into gamification has focused on cognitive ability, but there is evidence that image choices can be used to measure personality (Krainikovsky et al., 2019; Leutner et al., 2017). Notwithstanding this, a validated Big Five personality measure created for use in selection has not been described in

peer-reviewed research. The reported study, therefore, aims to contribute towards the lack of evidence addressing the potential for soft skills, such as personality, to be measured through gamified assessments, particularly those using an image-based format. We do so through an interdisciplinary approach, drawing from psychology and machine learning to create and validate the measure. This feasibility study found that all five traits can be accurately measured through an image-based format, with convergent validity similar to that of other traditional measures of personality. While we note that further investigation is needed into the assessment, our preliminary findings demonstrate that a forced-choice, image-based assessment has the potential to be an accurate and fair way of measuring the personality of applicants following further validation. We discuss the implications of this for practitioners and suggest areas for future research.

Assessments in selection

Selection methods have been used in recruitment for over 100 years to evaluate candidate suitability and predict future job performance (Ryan & Ployhart, 2014; Schmidt et al., 2016a; Schmidt & Hunter, 1998), with around 40 million assessments being completed globally by candidates each year (Chamorro-Premuzic, 2017). The most valid predictor of job performance is cognitive ability (N. Schmitt, 2014), with validity estimates of $r = .51$ (Schmidt & Hunter, 1998), a value that increases when combined with integrity tests ($r = .65$), work sample tests ($r = .63$), structured interviews ($r = .63$), or tests of conscientiousness ($r = .60$; Schmidt & Hunter, 1998). However, personality traits are not only useful when combined with cognitive ability; the Big Five personality traits – openness to experience, conscientiousness, extraversion, agreeableness, and emotional stability – are also predictive of job performance on their own.

While conscientiousness is the most valid trait for predicting personality across multiple job contexts (Barrick & Mount, 1991; Higgins et al., 2007; Judge et al., 1999; Kuncel et al., 2010; N. Schmitt, 2014), other personality traits are also useful for predicting

performance for jobs that require specific skills (Kuncel et al., 2010). For example, extraversion predicts the job and training proficiency of sales and managerial roles while openness and extraversion both predict training proficiency across occupations including police officers, sales staff, professionals, managers, and skilled/semi-skilled workers (Barrick & Mount, 1991). For those working in pharmaceutics, openness, extraversion, and conscientiousness predict task performance, while agreeableness and openness predict creativity (Rothmann & Coetzer, 2003). Nevertheless, the strongest predictor is the combination of cognitive ability and personality, which demonstrate incremental validity over each other and predict distinct areas of performance (Leutner & Chamorro-Premuzic, 2018). However, assessments of cognitive ability are often associated with adverse impact (Hausdorf et al., 2003), describing differences in selection rates between different groups (De Corte et al., 2007). Personality assessments, on the other hand, have little to no adverse impact so are a fairer way of assessing candidates while still being predictive of job performance (Hogan et al., 1996).

Game and Image-Based Assessments

While personality is often assessed via self-report methods using Likert scales, a recent trend in selection is gamification (Chamorro-Premuzic et al., 2017; Winsborough & Chamorro-Premuzic, 2016) where elements of game are added to traditional assessments to increase engagement (Armstrong, Ferrell, et al., 2016), including progress bars and visual and audio feedback (Landers, Armstrong, et al., 2022). Common issues with self-reported assessments are poor response quality (Krosnick, 1991) and lack of completion (Yan et al., 2011) due to the lengthy nature of scales, resulting in poor data. Game-based assessments can overcome these issues by offering a more immersive (Georgiou & Nikolaou, 2020) and engaging experience for participants (Lieberoth, 2015) and shorter testing times (Atkins et al., 2014; Leutner et al., 2021) compared to traditional assessments. Game-based assessments also elicit less test-taking anxiety (Mavridis & Tsiatsos, 2017; Smits & Charlier, 2011), are

less prone to social desirability bias than questionnaire-based measures, and can be more predictive of future job performance if they are designed to measure job-related behaviours (Armstrong, Landers, et al., 2016).

To date, much of the focus of research into gamification has been on using games to assess cognitive ability (e.g., Quiroga et al., 2015, 2016), with less of a focus on the measurement of soft skills such as personality. However, there is emerging research into how image preferences can be used to infer personality, where an image-based format can reinforce the benefits of game-based assessments and has the additional advantage of being language-neutral, meaning assessments can easily be adapted for use in other languages (Paunonen et al., 1990) unlike questionnaire-based measures, which often need to be redeveloped in the target language (H. Zhang et al., 2017). Language neutrality also has the benefit of increasing the accessibility of the assessment to those who have a preference for visual processing, including individuals who are neurodivergent (De Beer et al., 2014; Fassbender & Schweitzer, 2006), reducing the barriers that may prevent them from securing employment if their underlying ability is obscured by language processing difficulties that cause them to perform poorly on verbal assessments.

For example, Krainikovsky et al. (2019) predicted scores using algorithms based on image preferences from a selection of images tagged with information relating to objects, behaviour, emotions, and scenery. The convergent validity of the assessment with the NEO-PI-R (Costa & McCrae, 2008) ranged from $r = .06$ for neuroticism to $r = .28$ for agreeableness (Krainikovsky et al., 2019). Although this assessment had low convergent validity, it was not designed for use in selection. On the other hand, Leutner et al. (2017) created and validated image-based measure of creativity for use in selection, finding that scores on traditional measures were accurately predicted, with good concurrent validity with the traditional scales for all three creativity constructs, namely curiosity ($r = .35$), cognitive

flexibility (r = .50) and openness to experience (r = .50). The measure was more purposefully designed than Krainikovsky et al.'s (2019) and presented respondents with images and asked them to indicate which was most like them. However, Leutner et al.'s (2017) study only examined the feasibility of measuring openness to experience and did not extend the assessment to the remaining Big Five traits. Therefore, to build on these findings, the reported study describes an image-based assessment of all five personality traits with a similar format to that of Leutner and colleagues' (2017) creativity measure, with respondents being asked to indicate which image in a pair is more like them. The feasibility of using an image-based assessment of the Big Five personality traits for use in selection is explored through the creation of an image bank, the development of machine learning based predictive scoring algorithms, and examining the assessment's adverse impact and convergent validity between scores on the image-based assessment and the questionnaire-based IPIP-NEO-120 (J. A. Johnson, 2014). Based on the findings that image preferences can be used to predict personality (Krainikovsky et al., 2019) and that openness to experience can be predicted through image choices (Leutner et al., 2017), we hypothesise that:

H1: Machine learning based algorithms can be used to score an image-based assessment of the Big Five with strong convergent and discriminant validity for each trait.

Method

In this section, we outline the methods used in the two studies summarised below. We draw upon methods from industrial-organisational psychology and machine learning to develop the measure and create and validate the machine-learning-based scoring algorithms used by the assessment. Overall, the aim of these feasibility studies was to explore the potential for creating a valid image-based measure of personality to be used in talent recruitment in the future. An image-based format was chosen as previous findings have

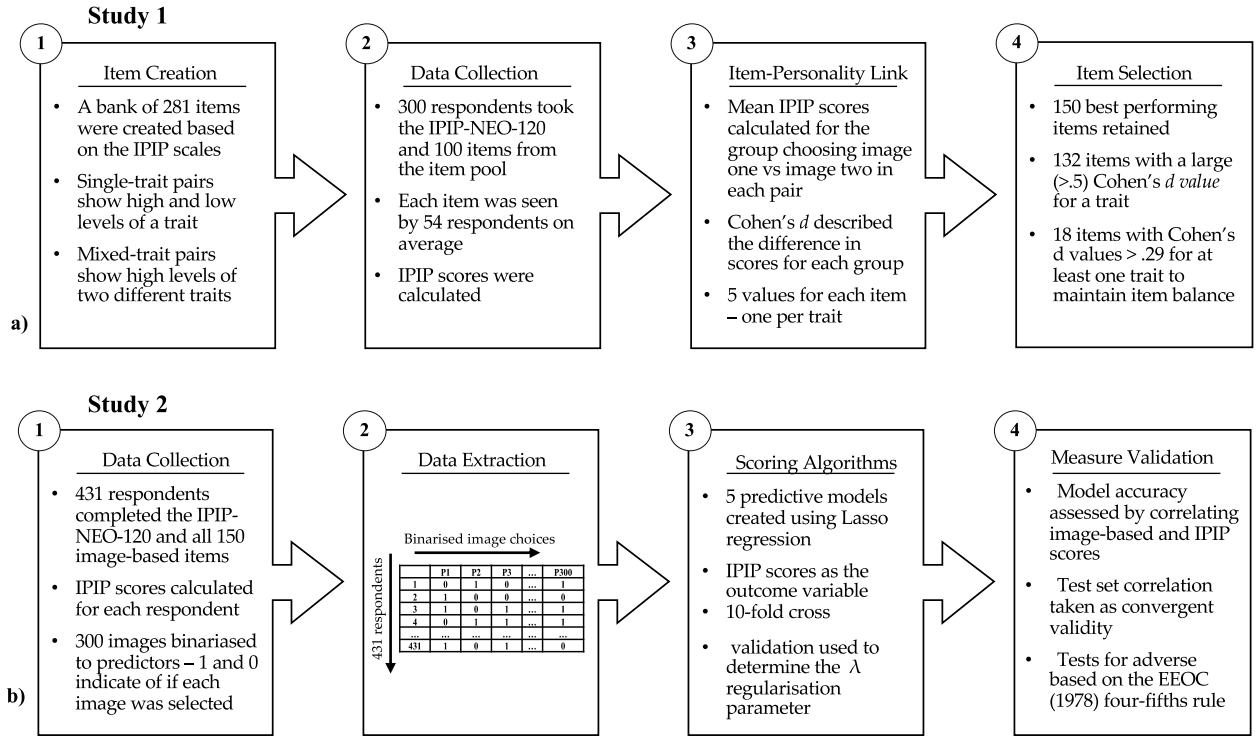
indicated personality can be measured through image choices (Krainikovsky et al., 2019; Leutner et al., 2017). The two studies conducted were:

- **Study One**: **Item Bank Creation** – Study One describes the creation of an item pool of image pairs, along with the selection of the 150 best-performing items, and the mapping of these items to the Big Five traits.
- **Study Two: Measure Validation** – Study Two describes the development of predictive machine-learning-based scoring algorithms based on a panel of respondents. This approach, where algorithms are developed that predict outcomes on traditional assessments, is common practice in predictive measures of personality (e.g., Bachrach et al., 2012; Kosinski et al., 2013; Krainikovsky et al., 2019; Leutner et al., 2017; Schwartz et al., 2013), as they convert binary choices to a more interpretable output that is more reflective of the continuous nature of the Big Five traits. Study Two also describes the validation of the assessment through measuring convergent and discriminant validity with Johnson's (2014) questionnaire-based IPIP-NEO-120 and tests for potential adverse impact.

An overview of the studies, which we expand upon below, can be seen in Figure 1.

Figure 1

Study One overview: Item creation and selection of the best-performing items for the image-based Big Five measure. (b) Study Two overview: Creation of scoring algorithms and tests of convergent validity with the questionnaire-based measure and adverse impact.



Study One: Item Bank Creation

Image-Based Measure

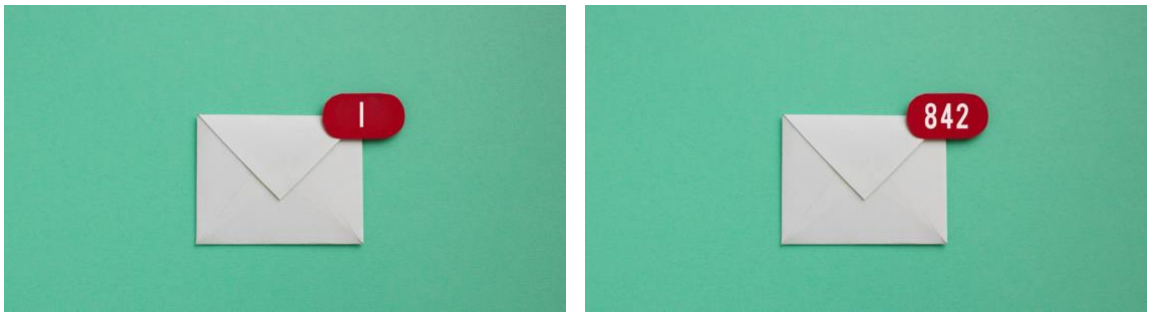
The creation of the image bank was guided by statements from Goldberg's (1992) International Personality Item Pool (IPIP) to ensure a good representation of the traits. To do so, these statements were used as a starting point to create visual representations of each of the traits using stock images sourced from Shutterstock. For example, the statement "I like to tidy up", was conceptualised as a messy versus tidy email inbox – an inbox with several unread emails versus an inbox that has few unread emails and is kept on top of (Figure 2). To find the images matching the conceptualisations of the team of industrial-organisational psychologists that created the measure, the image database was searched using keywords related to these concepts (e.g., searching for 'email notification'). The image pairs, or items, were created consciously and were designed to represent multiple ethnicities, age groups, and

genders. Moreover, the facets of neuroticism associated with mental health (anxiety and depression) were not included in the measure. Although it could be argued that the removal of these facets could alter the structure of the emotional stability construct, this action was taken to prevent discrimination against respondents with mental health issues, particularly since the measure is forced-choice, meaning this could be interpreted as asking respondents whether they have a mental health condition or not. Caution was also taken to ensure that the images would be suitable for professional use, with scenes featuring alcohol, parties, and inappropriately dressed models being avoided. Some images were edited to remove unnecessary text, which would have prevented language neutrality, to remove undesirable items such as cigarettes, or to allow the image to be cropped more effectively.

Figure 2

Examples of single-trait pairs. (a) is designed to measure the "I like to tidy up" statement from the orderliness facet of conscientiousness. (b) is designed to measure the "I look at the bright side of life" statement from the cheerfulness facet of extraversion.

a)



b)

The image pairs were either designed to be single-trait, where one image represented high levels of the trait and the other low levels of the trait, or mixed-trait, where the images reflect high levels of two different traits to determine the trait the respondent identified with

most. Some pairs that were more ambiguous were presented with adjectives to aid understanding. An example of two single-trait pairs can be seen in Figure 2 and an example of mixed-trait pairs in Figure 3. Once the images had been edited as required and cropped to a 1:1 aspect ratio, they were uploaded to the game development platform for the creation of a functional assessment. The measure is primarily designed to be completed on a smartphone device, although it can be completed on a computer, and presents image pairs one at a time along with the statement "I am more like...", prompting respondents to select the image they identify with most in the pair, thus being forced-choice. Audio and visual feedback were added to gamify the measure (Landers, Armstrong, et al., 2022), including a progress bar at the top and sound effects when an image was selected, as well as a pause button to allow respondents to pause and resume the assessment.

Figure 3

Examples of mixed-trait image pairs. (a) is designed to map onto the "I love to help others" statement from the altruism facet of agreeableness (left) and the "I feel comfortable around others" statement from the friendliness facet of extraversion (right). (b) is designed to be mapped onto the "I act comfortably around others" statement from the friendliness facet of extraversion (left) and the "I believe in the importance of art" statement from the artistic interests facet of openness (right).

a)

b)

Questionnaire-Based Measure

The IPIP-NEO-120 (J. A. Johnson, 2014) measures each trait through 24 questions using a five-point Likert scale, with a maximum score of 120 for each trait. Each trait is divided into six facets, with four questions measuring each, e.g., the cheerfulness facet of extraversion is measured by statements like “I radiate joy” and “I love life”. Items measuring neuroticism were reversed to measure emotional stability. This data served as the ground truth measure of personality.

Participants

300 compensated respondents were recruited through the online participant pool Prolific Academic (M_{age} = 31.14, SD = 9.26, 69% female) by an industry partner. Respondents completed the questionnaire-based measure along with 100 items from the image-based measure to avoid test-taking fatigue, resulting in each item being completed by an average of 54 participants (95% CI [38, 68]).

Item Selection

To select the best-performing items and reduce the length of the assessment, Cohen’s *d* values were used to quantify the difference in mean IPIP scores for the group of respondents choosing image one versus image two in each pair. This was calculated for each trait. Items that had large Cohen’s *d* values, indicating a large difference in personality scores between those selecting image one versus image two, were considered to perform well. Based on these values, 150 items, or 300 images, were selected to be retained: 132 items with moderate to large values (>.5 for a trait), and 18 items that showed small to moderate differences (>.29 on at least one trait) to maintain a balance in the items for each trait. The trait corresponding to the highest Cohen’s *d* value for that image pair was also used to identify the trait that the image measured best. In some cases, this trait differed from the trait that the item was designed to measure, but, as can be seen in Appendix C, almost two-thirds (60%) of 300 images were mapped onto the trait that they were designed to measure. The 150

items included in the assessment had a mean highest Cohen's *d* value of .77 (*SD* = .25; range: .29–1.77), suggesting that there were considerable differences in the personalities of individuals selecting each image in the pairs.

Study Two: Measure Validation

Participants

A second sample of 431 compensated respondents was recruited using Prolific Academic by an industry partner. Respondents completed the IPIP-NEO-120 and the full 150-item image-based assessment from Study One. The majority (*n* = 222) of respondents were female and most (*n* = 356) were under the age of 40. 209 were White, 73 were Black, 66 were Asian, 56 were Hispanic, and 14 were of Mixed Race. The IPIP scores once again served as the ground truth personality scores and were used to train the scoring algorithms described below.

Scoring algorithm creation

A machine learning based predictive model was created for each of the five traits, where scores for the relevant trait on the questionnaire-based measure served as the outcome or target variable. To represent image choices, the 150 pairs were used to create 300 binary dummy variables that indicated whether each image was selected or not by each candidate. These dummy variables served as predictors in least absolute shrinkage and selection operator (Lasso; Tibshirani, 1996) regression models that were created using a train-test split, where 70% of the data was used for training and the remaining 30% was held out as an unseen sample, allowing the generalisability of the models beyond the training dataset to be examined (Jacobucci et al., 2016). Lasso regression was favoured over ordinary least squares (OLS) regression, which is commonly used in behavioural sciences, as it is prone to overfitting and inflating R^2 values and consequently a lack of generalisability due to variance between datasets (McNeish, 2015). Lasso regression, however, reduces the effects of variance by adding some bias to the model and introduces a regularisation parameter, known as λ,

which decreases the size of all of the coefficients by an equal amount. As a result of λ, some coefficients are reduced to zero (McNeish, 2015) and removed from the model, creating a more interpretable model with fewer variables (Tibshirani, 1996). The removal of predictors also enabled the investigation of whether there was a crossover in the predictors retained by each model, as well as whether only image pairs mapped to that trait had predictive power. To determine the most appropriate hyperparameters for the models, 10-fold cross-validation was used. This iterative process divided the data into 10 folds, where each fold was predicted by the nine other folds fitted by a specified λ value. Mean squared error was calculated to measure fit and the process was repeated with different values of λ, and the average mean squared error for each λ value was compared, where the value corresponding to the smallest average mean squared error selected for the model (McNeish, 2015). Once the models had been trained, they were then applied to the test set data to predict the personality of these respondents.

Analysis

Typically, part of the validation of psychometric scales involves an examination of the internal consistency of items (Fenn et al., 2020), as measured by Cronbach's alpha (Cronbach, 1951), which provides an estimate of the amount of variance in the data that can be accounted for by a common construct. Items measuring the same underlying construct should have a similar pattern of responses and items that reduce the value of the alpha statistic can be removed to increase consistency (Taber, 2018). However, given that the image-based assessment is scored using machine learning, the items in the assessment are not used in the same way as with traditional measures that use a scoring key since each item will contribute to the overall score differently. However, because Lasso only retains the most meaningful predictors, this is similar to how items that reduce the alpha coefficient are removed from the measure. Consequently, the focus of the validation was how accurately the assessment measured personality, calculated by correlating the predicted personality scores

with actual scores on the IPIP (Cui & Gong, 2018). Correlations for the test set were also used to determine convergent validity. Discriminant validity was determined using a multi-trait multi-method approach (Campbell & Fiske, 1959), that represented correlations between the questionnaire- and image-based scores across the five traits. Here, a measure has discriminant validity if the convergent correlations are stronger than the correlations of two constructs that are theoretically distinct (D. J. Hughes, 2017). In other words, conscientiousness scores for the two assessments should have a stronger correlation than conscientiousness scores on the image-based assessment and extraversion scores on the questionnaire-based assessment, for example. If this was not the case, it would indicate that the scoring algorithm for conscientiousness was in fact measuring extraversion.

As well as the validity of the algorithms, the potential for adverse impact was determined for age, gender, and ethnicity. To do so, subgroup differences in scores were examined based on pass or fail criteria, where the median score for each trait was used as the passing threshold, in line with the metric for continuous regression models required under Local Law 144 (DCWP, 2023; The New York City Council, 2021). Subgroup differences were examined using the four-fifths rule, the two standard deviations rule, and Cohen's *d* effect sizes. According to the four-fifths rule, if the hiring rate of a group is less than four-fifths of the hiring rate of the group with the highest rate, adverse impact is occurring (Equal Employment Opportunity Commission, 1978). According to the two standard deviations rule, also known as the Z-test (S. B. Morris & Lobsenz, 2000), if the disparity between the expected and observed pass rates is greater than two standard deviations, adverse impact is occurring (Office of Federal Contract Compliance Programs, 2020). Finally, Cohen's *d* can be used to determine the effect size of the difference between the mean scores of two groups, where $d > |.20|$ indicates a small effect size, $d > |.50|$ indicates a medium effect size, and $d > |.80|$ indicates a large effect (Cohen, 1992). All three types of analysis were used to more

robustly test for group differences since the sample size is relatively small. However, group differences in scores are not always indicative of adverse impact and could instead reflect genuine group differences in ability (Society for Industrial and Organizational Psychology, 2018).

Results

In this section, we evaluate the performance of the scoring algorithms created in Study Two. We first present descriptive statistics for both the questionnaire-based measure and the novel image-based measure and subsequently present the metrics used to determine the performance of the models. We assess the convergent and discriminant validity between the questionnaire and image-based measures and test for potential adverse impact.

Descriptive Statistics

The descriptive statistics for scores on the IPIP-NEO-120 and image-based measure can be seen in Table 4. While the skewness and kurtosis values for these scores indicate that there may be a slight divergence from a normal distribution, the values are below the thresholds (two for skewness and seven for kurtosis) to be considered as substantially deviating from a normal distribution (Kim, 2013; West et al., 1995). Internal consistency of the questionnaire-based measure, determined by Cronbach's alpha (Cronbach, 1951), was high, ranging from .83 for openness to experience to .93 for emotional stability (see Table 5). This range is consistent with that reported by Johnson (2014), which ranged from .83 for openness to experience to .90 for emotional stability. The descriptive statistics for both measures are similar, suggesting that there is a similar distribution of scores.

\

Table 4

Descriptive statistics for the questionnaire- and image-based measures.

Trait	Mean	SD	Min	Max	Range	Skewness	Kurtosis
Questionnaire-based measure							
Openness	82.84	12.02	34.00	114.00	80.00	−.32	−.32
Conscientiousness	86.31	15.14	16.00	120.00	104.00	−.54	−.54
Extraversion	75.98	15.12	11.00	114.00	103.00	−.30	−.30
Agreeableness	90.32	13.64	13.00	119.00	106.00	−1.11	−1.11
Emotional stability	76.10	18.69	10.00	120.00	110.00	−.27	−.27
Image-based measure							
Openness	82.89	7.19	59.16	102.31	43.16	−.05	.03
Conscientiousness	86.79	10.68	52.17	110.84	58.67	−.51	−.12
Extraversion	76.07	11.77	44.20	101.97	57.77	−.11	−.65
Agreeableness	90.20	8.46	59.80	110.54	50.74	−.54	.51
Emotional stability	75.71	12.67	42.35	100.84	58.49	−.37	−.49

Although the Big Five traits are five different constructs, they intercorrelate (Chang et al., 2012). The intercorrelations for scores on the questionnaire-based measure, as seen in Table 5, concurred with intercorrelations that would usually be reported, ranging from .09 between openness to experience and emotional stability to .61 between conscientiousness and emotional stability. One reason for the high level of intercorrelation between emotional stability and conscientiousness could be because of the removal of some facets from emotional stability, which would leave the sub-scales for emotional stability that might be closer to conscientiousness (fewer neurotic behaviours).

Table 5

Correlation matrix for the questionnaire-based measure (Sample 2; N = 431).

Trait	1	2	3	4	5
1. Openness	.83				
2. Conscientiousness	.12*	.91			
3. Extraversion	.33**	.43 **	.90		
4. Agreeableness	.34**	.52 **	.24 **	.89	
5. Emotional stability	.09	.61 **	.60 **	.35 **	.93

Note. Diagonal values represent Cronbach's alpha coefficient. * $p < 05$. ** $p < 001$.

Model performance

The performance for each of the scoring algorithms created for the image-based assessment can be seen in Table 6, where correlations between scores on the image- and questionnaire-based assessments indicate model accuracy (Cui & Gong, 2018). While

correlations for all models were stronger for the training set, the test set correlations remained strong and significant, suggesting that the models can be generalised to unseen datasets (Jacobucci et al., 2016). This is important when creating scoring algorithms since the models will be applied to datasets other than the ones they were trained on when the assessment is deployed in practice.

Table 6
Model performance for the image-based assessment.

Trait	Training (n = 323)					Test (n = 108)				
	r	**R^2**	**MAE**	**MSE**	**RMSE**	***r***	**R^2**	**MAE**	**MSE**	**RMSE**
Openness	.77**	.56	6.24	65.63	8.10	.71**	.50	6.55	64.46	8.03
Conscientiousness	.82**	.66	6.88	82.62	9.09	.70**	.47	7.50	97.26	9.86
Extraversion	.86**	.74	5.78	61.23	7.82	.78**	.61	6.82	82.28	9.07
Agreeableness	.77**	.56	6.79	84.97	9.22	.60**	.34	8.06	103.03	10.15
Emotional stability	.80**	.63	8.92	131.35	11.46	.70**	.47	10.70	175.29	13.24

Note. *r* = Pearson correlation coefficient for actual and predicted scores; R^2 = proportion of variance explained; MAE = mean absolute error; MSE = mean squared error; RMSE = root mean squared error.

The test set correlations were also used to assess convergent validity, which ranged from .60 for agreeableness to .78 for extraversion, indicating that the image-based format can be used to measure personality in a similar way to traditional, questionnaire-based formats. Moreover, as can be seen in Table 7, in the majority of cases, the discriminant correlations were smaller than the convergent. While for emotional stability in particular, the discriminant correlations were relatively high, the same pattern is seen in Table 5. This result could also be explained by the removal of the anxiety and depression facets from emotional stability since the remaining facets are closer to those of other traits, such as conscientiousness. As such, H1 was supported for all five traits.

Table 7
Multitrait-multimethod matrix of the image- and questionnaire-based measures (Test set of sample 2; n = 108).

Trait	1	2	3	4	5	6	7	8	9	10
Questionnaire-based										
1. Openness	1.00									
2. Conscientiousness	.10	1.00								
3. Extraversion	.29**	.35**	1.00							
4. Agreeableness	.31**	.50**	.12	1.00						
5. Emotional stability	.01	.63**	.54**	.30**	1.00					
Image-based										
6. Openness	.71**	.05	.39**	.31**	.06	1.00				
7. Conscientiousness	-.04	.70**	.21*	.32**	.54**	.00	1.00			
8. Extraversion	.26**	.30**	.78**	.12	.52**	.48**	.32**	1.00		
9. Agreeableness	.18	.42**	.13	.60**	.23*	.46**	.54**	.22*	1.00	
10. Emotional stability	.08	.51**	.57**	.21*	.70**	.19	.69**	.71**	.38**	1.00

Further, the number of predictors retained in the models ranged from 13 for openness to 32 for extraversion, suggesting that personality could be measured through shorter assessments and that they can offer similar insights into personality as longer assessments. Indeed, only 68 of the 150 items were retained across the five models, suggesting that the majority of the items could be removed from the assessment without impacting its validity. This indicates that the Big Five could be rapidly measured through a small number of images in around two minutes.

Subgroup differences

Given that the metrics used to measure subgroup differences, the four-fifths rule, 2SD rule, and Cohen's *d,* can result in discrepant findings (S. B. Morris & Lobsenz, 2000), the potential for adverse impact was flagged when there were exceptions to two or more metrics. Specifically, impact ratios below 80, Cohen's *d* values greater than .20, and values greater than +/−2 standard deviations indicated that there were group differences. Based on these measures, potential for adverse impact was found against male and Asian respondents for the openness model, against Hispanic, Mixed, and Other ethnicity respondents for the conscientiousness model, against males, Asians, and Hispanics and Other ethnicity test-takers

for the agreeableness model, and females, Mixed and Other ethnicity test-takers for emotional stability. The results of the adverse impact analysis for these groups can be seen in Table 8 (see Appendix D for the full adverse impact analysis).

Table 8

Subgroup differences in scores for the image-based measure where two or more metrics were violated (Accepted Adverse Impact Ratio: >.8. Accepted Cohen's D: < |.20|. Accepted 2 SD: <|2|).

Trait	Subgroup	Group Size	*n* passing	Pass rate	Impact ratio	Cohen's *d*	2SD
Openness	Male	205	83	.40	.69	-.37	-3.73
Openness	Asian	66	19	.29	.47	-.68	-3.72
Conscientiousness	Hispanic	56	23	.41	.70	-.36	-1.41
Conscientiousness	Mixed	14	3	.21	.36	-.81	-2.16
Conscientiousness	Other	13	4	.31	.52	-.58	-1.40
Agreeableness	Male	205	79	.39	.63	-.47	-4.69
Agreeableness	Asian	66	28	.42	.74	-.29	-1.32
Agreeableness	Hispanic	56	21	.38	.66	-.39	-1.99
Agreeableness	Other	13	5	.38	.67	-.37	-.84
Emotional stability	Female	222	98	.44	.78	.25	-2.57
Emotional stability	Mixed	14	5	.36	.62	-.44	-1.08
Emotional stability	Other	13	5	.38	.67	-.38	-.84

To examine whether these subgroup differences resulted from the scoring algorithms or whether they could represent genuine subgroup differences, adverse impact analysis was also conducted for the IPIP-NEO-120. As can be seen in Table 9, the subgroup differences found for the image-based assessment echo those of the questionnaire-based assessment, suggesting that the subgroup differences identified in the image-based measure were due to subgroup differences in scores on the questionnaire-based measure and not due to the image-based format. This highlights the need to examine subgroup differences in the training data since machine learning algorithms can amplify this bias (Tay et al., 2022). The subgroup differences may be due to measurement bias in the questionnaire-based assessment or could reflect genuine differences in ability since group differences are not always indicative of bias (Society for Industrial and Organizational Psychology, 2018).

Table 9

Subgroup differences in scores for the questionnaire-based measure where two or more metrics were violated (Accepted Adverse Impact Ratio: >.8. Accepted Cohen's D: < |.20|. Accepted 2 SD: <|2|).

Trait	Subgroup	Group Size	*n* passing	Pass rate	Impact ratio	Cohen's *d*	2sd
Openness	Age 40 or older	75	29	.39	.75	-.26	-2.05
Openness	Male	205	86	.42	.75	-.28	-2.87
Openness	Asian	66	21	.32	.51	-.64	-3.11
Openness	White	209	102	.49	.78	-.28	-.25
Conscientiousness	Male	205	87	.42	.79	-.22	-2.31
Conscientiousness	Asian	66	30	.45	.77	-.27	-.41
Conscientiousness	Hispanic	56	24	.43	.73	-.32	-.79
Conscientiousness	Mixed	14	3	.21	.36	-.81	-2.01
Conscientiousness	Other	13	5	.38	.65	-.41	-.68
Agreeableness	Male	205	81	.40	.68	-.38	-3.84
Agreeableness	Asian	66	24	.36	.61	-.48	-2.18
Agreeableness	Black	73	24	.33	.55	-.56	-2.97
Agreeableness	Hispanic	56	26	.46	.78	-.27	-.37
Agreeableness	Mixed	14	5	.36	.60	-.49	-.99
Agreeableness	Other	13	6	.46	.77	-.27	-.19
Emotional stability	Female	222	92	.41	.70	.36	-3.63

Since group differences were observed for the questionnaire-based measure, measurement bias was investigated by examining whether convergence varied by subgroup (Tay et al., 2022). As can be seen in Table 10, there are differences in the convergence for subgroups, and these differences echo the group differences in scores for both the image- and questionnaire-based measures. For example, the convergence for Black and Asian respondents for agreeableness is significantly lower than that of White and Hispanic respondents, with group differences being found in their scores. This may be because these groups are underrepresented in the data.

Table 10

The convergent correlations by subgroup.

Demographic	O	C	E	A	ES
Gender					
Male	.65**	.76**	.80**	.69**	.72**
Female	.76**	.58**	.76**	.42**	.69**
Age					
Under 40 years old	.67**	.68**	.80**	.50**	.68**
Age 40 or older	.78**	.76**	.68**	.86**	.77**
Ethnicity					
White	.77**	.75**	.83**	.62**	.83**
Black	.69**	-.01	.86**	.05	.73**
Asian	.65**	.72**	.85**	.08	.45
Hispanic	.56**	.72**	.83**	.83**	.57*

Note. O, C, E, A, and ES refer to openness, conscientiousness, extraversion, agreeableness, and emotional stability, respectively.

Discussion

This study aimed to create scoring algorithms for and validate a novel, image-based measure of the Big Five personality traits to explore the potential for such a measure to be used in selection. Study One described the creation of an item bank and the selection of the 150 best-performing items. Study Two described the development of a predictive machine learning based scoring algorithm for each trait and the validation of the image-based measure by measuring convergent validity with a validated, questionnaire-based measure and testing for potential adverse impact.

In this section, we discuss the performance of the scoring algorithms created for the reported image-based assessment of personality and the possible limitations that could result from the relatively small sample used in this study. Specifically, we discuss the performance of the models and methodological considerations. We also suggest some areas for further research before this assessment can be deployed in practice and discuss the implications that our preliminary findings may have for the use of image-based measures of personality in selection.

Model performance

The findings of this study provide preliminary evidence that all five personality traits can be accurately measured via image choices in a similar way to traditional measures of personality. Models were trained on 70% of the data and then cross-validated with the remaining test data to assess generalisability. The model's accuracy was assessed by correlating scores on the image- and questionnaire-based measures. Across all five traits, correlations were strong for both the training and test data, indicating good model accuracy and generalisability to unseen data (Jacobucci et al., 2016) The convergent validity between the image- and questionnaire-based measures, determined by correlations for the test set, ranged from .60 for agreeableness to .78 for extraversion. These values exceed those reported by previous non-verbal personality measures; correlations between the Nonverbal Personality Questionnaire and the NEO-FFI (McCrae & Costa, 1985) ranged from .45 for emotional stability to .59 for agreeableness (Paunonen et al., 2001). The convergent validity range for this measure is comparable to the convergent validity between different questionnaire-based measures of personality, with correlations between scores on the IPIP (Goldberg, 1992) and NEO-FFI (McCrae & Costa, 1985), ranging from .50 for agreeableness to .76 for emotional stability (Lim & Ployhart, 2006). However, since some discriminant correlations were of a relatively similar magnitude to convergent correlations, this limits the conclusions that can be made about the validity of the assessment and highlights the need for further investigation. Nevertheless, H1 – that machine learning based scoring would result in strong convergent and discriminant validity – was generally supported.

Limitations and future directions

While this assessment performed well in terms of accuracy and convergent validity, the sample size was small relative to other studies that describe predictive personality measures (Bachrach et al., 2012; Kosinski et al., 2013; Leutner et al., 2017; A. H. Schwartz et al., 2013). Although this could have implications for the model's performance (Raudys &

Jain, 1991; Vabalas et al., 2019), the use of a train/test split is likely to have reduced the potential for the sample size to negatively impact the models (Vabalas et al., 2019). Additionally, although the size of the sample was relatively small, the range of personality scores was large, suggesting that the sample represents a range of personalities. Nevertheless, a larger sample size would likely have been beneficial, with the potential for creating more robust models.

Furthermore, while the majority of items were mapped onto the trait that they were designed to measure in Study One, some were not, suggesting that it is difficult to perfectly align text- and image-based measures. This may be because image-based measures rely on personal interpretation of meaning which may vary between people. Despite this, all items were included as predictors for each model to examine which items were retained in the models and whether they aligned with the mapping.

Finally, although the subgroup differences in the scores on the image-based assessment reflected differences in scores on the questionnaire-based measure, indicating that they could be genuine differences in personality, this could not be investigated in the current study due to a lack of performance data to examine whether these differences represented genuine differences in performance (Society for Industrial and Organizational Psychology, 2018). Further, White participants were overrepresented in the data, which could explain why the algorithms do not perform as well for certain subgroups because the algorithms were not well-optimised to evaluate them (Buolamwini & Gebru, 2018).

Given that this study was the first step in the validation process for the assessment, before it can be used in practice, further validation is needed to more robustly explore the construct validity of the measurement, bias, generalisability, and how comparable this assessment is to other questionnaire-based measures. To do so, we suggest the following:

- **User experience:** To better understand how respondents engage with the measure, future studies could examine user experience, including how engaging respondents found the measure to be. It could also investigate whether the meaning assigned to the items by the team of designers converges with that of the respondents by asking a sample of respondents to assign their own adjectives to the items. This would allow further refinements to be made to the measure which may strengthen its performance.
- **The potential for shorter measures:** Since only a small number of predictors were retained by each model and there was some crossover in the predictors retained by the models, future studies should investigate how shorter versions of the assessment could take advantage of this by examining how effective different item combinations are at measuring the traits. This would result in even shorter testing times for candidates, reducing the time it takes to complete the overall battery of selection assessments.
- **Bias and transparency:** Group differences in scores do not always indicate bias and can instead be reflective of genuine differences in latent levels of traits for different subgroups (SIOP, 2018). However, even when group differences in scores are not due to differences in ability, they do not always lead to adverse impact, especially when the analysis is based on a small sample (EEOC, 1978). Therefore, further validation is needed with a larger sample to more robustly determine whether the reported group differences could result in adverse impact, particularly since the importance of transparency and fairness in the algorithms used in hiring is increasingly a point of concern (Kazim et al., 2021; Raghavan et al., 2020);

- **Mitigating subgroup differences**: While the potential for adverse impact from this assessment echoes concerns about the fairness of conventional selection assessments (Hough et al., 2001), there are a number of avenues for investigating and mitigating bias in algorithms (Tay et al., 2022). Further research exploring the potential for mitigating group differences in the algorithms used by this assessment is needed, particularly since there is evidence of potential measurement bias in the questionnaire-based measure used to construct and validate the algorithms. Follow-up studies are, therefore, required to investigate the best way to mitigate this.
- **Generalisability:** The samples used in this study may be limited if they did not represent a diverse group of respondents. For example, data referring to the occupation of respondents were not collected, meaning the generalisability of the findings could be limited to a particular industry if respondents are from a similar background. To address this, a future study should recruit an additional sample from another source such as MTurk to validate the algorithm in a different population of respondents who are likely to have different attributes to those in the current samples.
- **Cultural appropriateness:** As only English-speaking respondents were included in this study, a variation in the interpretation of the items was not investigated across multiple cultures or languages. Whilst it is suggested that non-verbal assessments can be applied to any language without redevelopment (Paunonen et al., 1990), it is still important to ascertain whether the images included in this assessment are appropriate in other cultures. The findings of this study indicate that there are potential differences in the interpretation of the images for different subgroups, with convergence being null on some traits for Asian and Black

respondents. Therefore, future studies should take a cross-cultural approach to investigate the performance of the measure in different cultures and ethnicities.

- **Score inflation:** Job application contexts have higher stakes as they can affect career-related opportunities (Stobart & Eggen, 2012). Since there is evidence for the inflation of personality scores in high-stakes contexts (Arthur et al., 2010), a future study could investigate score inflation on this novel assessment in a high-stakes context. The forced-choice image-based format might decrease candidates' ability to fake their responses compared to questionnaire-based tests.
- **Measure validity:** While this study used Cohen's *d* values to investigate the mapping of the images to traits, the construct validity of the measure was not robustly investigated through unsupervised methods such as factor analysis as a first step, particularly to inform how items should be used in the models when they contribute to several traits (Speer & Delacruz, 2021). The importance of construct validity is emphasised by (F. Y. Wu et al. (2022)'s game-based assessment that measured cognitive ability instead of conscientiousness, as intended. Moreover, the lack of performance data did not allow for the predictive validity and therefore utility of the assessment to be examined. Future studies should draw on psychology best practices in assessment and scale construction to a greater extent to provide additional evidence of the validity of the measure.
- **Measure reliability:** Respondents only took the measure once, meaning that response stability and consistency (test-retest reliability) could not be examined. Thus, it is not known whether respondents are likely to make the same image choices and therefore have similar personality scores each time they take the assessment. Further, performance data was not collected, meaning that the predictive validity of the assessment could not be determined. Consequently,

additional validation is needed to determine the test-retest reliability of this assessment (Fenn et al., 2020). Further, the internal reliability of the scale could be investigated through traditional approaches such as Cronbach's alpha, or through test information function (Samajima, 1994) to examine the contribution of the assessment items to Big Five scores.

Implications

The reported study contributed towards addressing the lack of validated gamified assessments of personality, particularly assessments using an image-based format. Since game-based assessments are reportedly more engaging (Lieberoth, 2015), result in greater satisfaction (Downes-Le Guin et al., 2012; Leutner et al., 2021), and have shorter testing times (Atkins et al., 2014; Leutner et al., 2021) than traditional assessments, this measure could offer a more positive experience for applicants than traditional psychometric assessments. As applicants who view the selection process of an organisation more positively reportedly have more favourable perceptions of the employer and are more likely to accept a job offer (Hausknecht et al., 2004), this has implications for businesses as attractive selection methods can avoid offer rejections from talented candidates. The findings of this validation study provide preliminary evidence that all five personality traits can be measured through image choices, with the novel assessment showing promise for use in selection following further validation.

Conclusion

Overall, this study found that image-based measures of personality that combine expertise from psychology and computer science may be a valid and fair alternative form of assessment that could be used in place of traditional assessments using Likert scales. Using predictive scoring algorithms, the image-based assessment of personality described in this study demonstrates convergent validity with a validated, questionnaire-based measure comparable with the convergence between other questionnaire-based personality measures,

suggesting that the reported assessment measures the Big Five traits in a similar way to traditional measures. Furthermore, this study found that the image-based measure is largely free from group differences which could potentially lead to adverse impact; however, further studies are needed using larger samples to test this more robustly. The measure also needs to be further validated to assess test-retest reliability and score inflation, as well as in other languages and cultures. Further studies could also compare user experience for the image-based assessment and a questionnaire-based measure. These preliminary findings have positive implications for the use of this assessment in selection; however, there is scope for further validation before this measure can be used in practice.

Chapter 4. Study Three – comparing scoring approaches for a forced-choice, image-based personality assessment

Scoring a Forced-Choice Image-Based Assessment of Personality: A Comparison of Machine Learning, Regression, and Summative Approaches

Airlie Hilliard,[1,2,*] Emre Kazim,[2,3] Theodoros Bitsakis,[4] Franziska Leutner[1,4]
[1]Institute of Management Studies, Goldsmiths, University of London, New Cross, London SE14 6NW, UK
[2] Holistic AI, 18 Soho Square, London W1D 3QH, UK
[3]Department of Computer Science, University College London, Gower St, London WC1E 6EA, UK
[4]HireVue, London, UK

Abstract

Recent years have seen rapid advancements in the way that personality is measured, resulting in a number of innovative predictive measures being proposed, including using features extracted from videos and social media profiles. In the context of selection, game- and image-based assessments, which can overcome issues like social desirability bias, lack of engagement and low response rates that are associated with traditional self-report measures, are increasingly being used. Forced-choice formats, where respondents are asked to rank responses, can also mitigate issues such as acquiescence and social desirability bias. Chapter 3 reports on the development of a gamified, forced-choice image-based assessment of the Big Five personality traits created for use in selection. While the assessment was scored using Lasso regression in Study Two, in this study, we compare the machine learning based Lasso and Ridge approaches to ordinary least squares regression, as well as the summative approach that is typical of forced-choice formats. We find that the Ridge and Lasso models perform best in terms of generalisability and convergent validity. Moreover, Lasso performs similarly to Ridge but with fewer predictors and thus provides an opportunity to reduce the assessment length. We recommend the use of predictive Lasso regression models for scoring forced-choice image-based measures of personality over the other approaches. Potential further studies are suggested.

Introduction

This study compares machine learning based, ordinary least squares, and summative approaches to scoring the forced-choice image-based assessment of personality that was described in Studies One and Two in Chapter 3. While in recent years, new ways of scoring forced-choice assessments have been developed that can overcome issues associated with traditional forced-choice scoring approaches (Brown & Maydeu-Olivares, 2011, 2013), these are typically for multidimensional measures. Since the measure used in this study has a combination of unidimensional (single-trait) and multidimensional (mixed-trait) items, these methods have limited applicability. As such, Study Two uses machine learning based scoring algorithms to overcome these challenges. Here we extend this work, examining how the use of different predictor combinations in different models impacts the validity of the measure.

We begin by examining the significance of personality and how it is measured, both using traditional and more contemporary approaches, before narrowing our focus to image-based and forced-choice measures. We then describe the development of the models and evaluate their performance in terms of convergent and discriminant validity with the IPIP-NEO-120 and generalisability from the training to test data. In line with prior research (Speer & Delacruz, 2021), we conclude that machine learning based approaches outperform other scoring approaches and that they are a viable alternative option for scoring forced-choice assessments.

Measuring Personality

An individual's personality has significant implications for many aspects of their life, including their wellbeing, social relationships, health, and career success (B. W. Roberts et al., 2007; Soldz & Vaillant, 1999). Indeed, the Big Five personality traits (openness to experience, conscientiousness, extraversion, agreeableness and neuroticism or emotional stability) are routinely tested in pre-employment screenings due to their ability to predict future job performance (Barrick & Mount, 1991; Kuncel et al., 2010; Pletzer et al., 2021;

Rothmann & Coetzer, 2003; Schmidt & Hunter, 1998; N. Schmitt, 2014). While self-report methods, such as the International Personality Item Pool (IPIP; Goldberg, 1992) scales and the NEO-PI R (Costa & McCrae, 2008), have been the default method of assessing personality until recently, self-report scales are associated with poor response quality (Krosnick, 1991) and can lead to incomplete responses due to respondent attrition (Yan et al., 2011), particularly if scales are lengthy. Self-reported measures of personality are also associated with social desirability bias or faking (van de Mortel, 2008), especially in high-stakes contexts, where respondents inflate their scores more compared to respondents completing the assessment in low-stakes contexts (Arthur et al., 2010). This has implications for the use of personality assessments in high-stakes contexts like recruitment, where candidates may attempt to inflate their scores to appear more favourably (Le et al., 2011).

Alternative Ways of Measuring Personality

To overcome some of the issues associated with self-report measures, some have proposed using adjective-based daily diary measures of specific personality traits to avoid issues with one-time measurements (Di Sarno et al., 2020) while others have proposed a number of alternative data sources that can be used to measure personality through predictive models. For example, personality has been inferred from facial expressions in YouTube videos (Biel et al., 2012) and video interviews (H.-Y. Suen et al., 2019). However, facial recognition analysis is a controversial approach due to concerns about how it might impact individuals with disabilities who do not display typical expressions (e.g., Electronic Privacy Information Center, 2019). Fortunately, it is not just facial expressions that can be used to infer personality; audio (speaking activity and prosody) and non-verbal cues (looking activity, pose and body movements) in videos can also be used (Biel & Gatica-Perez, 2013). Indeed, features such as speaking activity, prosody, head motion, and overall motion observed in video interviews have also been used to infer personality (Nguyen & Gatica-Perez, 2016).

Moving away from video analysis, personality has also been predicted from various other data sources such as Facebook Likes (Kosinski et al., 2013), eye movement while running errands (Hoppe et al., 2018), and mobile phone data, including calls and text frequency, GPS data and text response rate (de Montjoye et al., 2013). Others have explored a text-based approach, using language to predict personality. While this is nothing new given that the Big Five model of personality was derived from language analysis (Digman, 1990), contemporary approaches use non-traditional sources of language such as social media posts and combine them with natural language processing computational techniques. For example, based on the frequency of word use and clusters of topics mentioned in Facebook status updates, personality has been predicted using latent Dirichlet allocation, a natural language processing technique used to cluster words into related topics (Park et al., 2015). Such approaches, therefore, move away from the need for self-report, reducing the influence of faking and allowing personality to be measured automatically (Park et al., 2015), although impression management on social media is not uncommon (Schlosser, 2020), meaning that social media based measures could still be affected by social desirability bias.

While many of these ways of assessing personality were not designed for use in selection, purposely created image-based assessments of personality are emerging and increasingly being offered by commercial providers (e.g., HireVue and Traitify). Image-based formats offer a number of benefits such as their language neutrality, which means that they can more readily be used in other languages compared to questionnaire-based assessments (Paunonen et al., 1990; H. Zhang et al., 2017). In addition, if the image-based measures are gamified by adding features like sound effects and progress bars (Landers, Armstrong, et al., 2022), as with the measure described in Study One, this can have additional benefits since game-based assessments elicit less test-taking anxiety (Mavridis & Tsiatsos, 2017; Smits & Charlier, 2011), can be more engaging for respondents (Lieberoth, 2015). and are generally

quicker to complete (Georgiou & Nikolaou, 2020; Leutner et al., 2021). Machine learning based scoring can also support a shorter completion time if the same items are used in multiple scoring algorithms since fewer items will be needed. Assessments in this format can, therefore, overcome some of the issues associated with traditional self-report measures. This is particularly beneficial in the domain of recruitment since applicant perceptions of a selection process can influence how likely they are to accept a job offer (Hausknecht et al., 2004), which can have implications both for candidates and hiring managers who could potentially miss out on top talent if they have an unengaging recruitment process.

Forced-Choice Assessments

Forced-choice assessments vary from traditional self-report assessments in that they ask respondents to indicate the responses that are most and least like them, instead of asking them where they lie along a scale. Assessments of this type typically have either two or four response options, where the most common format is multidimensional i.e., showing blocks of response options with statements from multiple constructs (Hontangas et al., 2015). In the context of personality, a respondent could be shown statements relating to multiple traits and asked to identify the statements they identify with most and least – in either a text or image-based format. Assessments of this format can prevent central tendency and extreme response styles since there is no midpoint or extreme response options (Brown & Maydeu-Olivares, 2013), as well as acquiescence responses, where respondents select both positive and negative statements, since they are not able to endorse all of the statements presented to them (Brown & Maydeu-Olivares, 2013). Further, they are more resistant to faking than traditional measures (Salgado & Táuriz, 2014) since it is often more difficult to identify response options that may be more desirable to employers compared to Likert-based scales, which is particularly important in high-stakes contexts like recruitment.

A simple way of scoring two-item forced-choice measures is by summing the number of times an item that represents a high level of a certain trait is selected (Hontangas et al.,

2015). When the measure is multidimensional, featuring positive statements about two different traits, the selected statement is given a score of 1 and the unselected a score of 0. Accordingly, the score associated with the trait the selected image represents increases by one point. When the blocks have more than two statements and respondents are asked to indicate the most and least like them, the item selected for most is given 2 points, the unselected 1 point, and the least favoured 0 points. Again, scores for each construct can be calculated by summing the number of points relating to each trait (Hontangas et al., 2015). However, multidimensional forced-choice measures result in ipsative scales, where the score on one dimension is relative to another dimension and the total score for each respondent across the constructs is the same. As such, concerns have been raised about how well individuals can be compared since it is impossible to score above or below the mean score for all constructs (Brown & Maydeu-Olivares, 2011). On the other hand, mixed-dimensional forced-choice formats, or those combining multidimensional and unidimensional items, are not prone to ipsative scoring in the same way as fully multidimensional measures as they behave more like a traditional questionnaire-based method.

To combat the issue of ipsative scores with multidimensional measures, alternative ways of scoring forced-choice measures have been proposed based on Item Response Theory (IRT), which models how latent constructs such as personality manifest through item responses (Harvey & Hammer, 1999), and Thurstone's framework for comparative data (Thurstone, 1994), which views individuals' choices as probabilistic (Brown & Maydeu-Olivares, 2011, 2013). These models, which are estimated by structured equation modelling using MPlus, are more similar to traditional models that would be seen with Likert scale based measures, with structured factor loadings and uniqueness (Brown & Maydeu-Olivares, 2011), allowing the scores to be better compared between individuals (Brown & Maydeu-Olivares, 2011, 2013). Indeed, comparisons of traditional and IRT-based scoring of forced-

choice measures found the IRT scoring to perform better across multiple types of forced-choice measures in terms of the correlation between the true score and that estimated by the traditional and IRT approaches (Hontangas et al., 2015). However, such approaches are typically focused on fully multidimensional measures due to their ipsative nature, with fewer efforts being focused on scoring approaches for measures that use solely unidimensional items or a combination of unidimensional and multidimensional items, such as in the current study, where mixed-trait pairs represent multidimensional items and single-trait pairs represent unidimensional items.

Predictive Scoring

In contrast to questionnaire-based measures, many contemporary measures of personality use predictive scoring algorithms that make use of non-traditional and unstructured data to predict personality scores on traditional measures (Biel et al., 2012; Hilliard, Kazim, et al., 2022a; Kosinski et al., 2013; Leutner et al., 2017; Park et al., 2015). Models of this type, which predict a specified outcome, are trained using supervised learning (Nasteski, 2017). This is in contrast to unsupervised learning, where algorithms are used to identify clusters in the data, with no specified target variable (Rosenbusch et al., 2021). Since the personality scores are known, these approaches, therefore, use supervised learning.

While it is possible to use ordinary least squares regression (OLS) to predict scores, predictive measures typically have a large number of predictors relative to the number of participants, meaning there is a small n/p ratio (Putka et al., 2018). As a result of OLS being designed to minimise the sum of the squared difference between the actual score and predicted score, this can lead to overfitting of the model to the data it was trained on, particularly with small n/p ratios (McNeish, 2015). Since the assessment will be taken by individuals other than those the model was trained on, this can result in the model performing poorly when applied to other samples, limiting its usefulness as a scoring algorithm. However, machine learning approaches to prediction can help to overcome this.

Indeed, Lasso regression (Tibshirani, 1996) introduces bias into the model, therefore reducing the impact of variance between datasets on the performance of the model due to the bias-variance trade-off (McNeish, 2015). Consequently, compared to OLS regression, Lasso produces a model that is more generalisable to datasets other than the one it was trained on, being more suitable for a scoring algorithm. Another advantage of Lasso regression is that as a result of the regularisation parameter lambda (λ), which places a constraint on the absolute sum of coefficients and shrinks all coefficients by an equal amount, the coefficient of some predictors is reduced to zero, removing them from the model (McNeish, 2015). Consequently, only the predictors that are most powerful are retained in the model, reducing the complexity of the model and increasing its interpretability (Tibshirani, 1996). However, this could also be at the expense of important predictors with weaker relationships than others being removed from the model or patterns in the data being overlooked.

Similar to Lasso regression is Ridge regression, which uses L2 regularisation instead of the L1 regularisation that is used by Lasso regression. Here, a constraint is placed on the sum of squared coefficients to penalise the loss function. This results in coefficients being shrunk in a way that is proportional to the size of the coefficient without the possibility of predictors being removed from the model (McNeish, 2015). As such, highly influential predictors are made to have less of an effect, and all items are retained in the model, overcoming the potential for important predictors to be removed that is associated with Lasso regression. However, despite the regularisation that is characteristic of Lasso and Ridge helping to overcome issues with overfitting

Study Two reported on the use of Lasso regression to create scoring algorithms for a forced-choice image-based assessment of personality, which presented respondents with pairs of images (items) designed to measure the Big Five personality traits and asked them which image in the pair is most like them. After refinement, 150 item pairs (300 images) were

retained in the assessment, with these predictors being binarised to represent whether a respondent selected the image or not. During the validation study, Cohen's *d* values were used to identify the trait that image measured best. All 300 predictors, regardless of which trait they were designed to measure or mapped to in the validation, were entered into the models and used to predict Big Five scores on the IPIP-NEO-120 (J. A. Johnson, 2014). This study aims to build on the findings of Study Two, comparing the convergent and discriminant validity, generalisability, and adverse impact of multiple approaches to scoring the image-based assessment:

a) Lasso regression using all 300 images to predict each trait.

b) Lasso regression using only the items intended to measure each trait.

c) Lasso regression using the images mapped to each trait by Cohen's *d.*

d) Ridge regression using all 300 images to predict each trait.

e) Ridge regression using only the items intended to measure each trait.

f) Ridge regression using only the items mapped to each trait by Cohen's *d*.

g) OLS regression using all 300 predictors.

h) OLS regression using only the items intended to measure each trait.

i) OLS regression using items mapped to each trait by Cohen's *d*.

i) A summative approach using images designed to measure each trait.

These models are elaborated on below. Convergent validity was measured as the correlation between the image-based score and the score on the traditional personality test and discriminant validity was measured as the inter-correlations between the five personality traits and compared to inter-correlations on the traditional personality test. Generalisability was determined by comparing the correlation between the training and test sets as this can be used to establish how well the model can be applied to unseen data (Jacobucci et al., 2016). Finally, subgroup differences were calculated using adverse impact metrics. The purpose of

this study was to examine whether using machine learning based scoring was the most appropriate approach for the image-based assessment. Given that typical regression approaches can result in overfitting (McNeish, 2015; Putka et al., 2018) and that the summative approach solely used expert opinion and was not informed by a data-driven approach, we hypothesise that:

H2: The machine learning scoring approaches will have stronger convergent and discriminant validity than the manual or OLS approaches.

Method

The image-based measure of personality described in Study One presents respondents with pairs of images and asks them to indicate which image in the pair is more like them, where the image pairs, or items, were intended to map onto the statements from the IPIP scales (Goldberg, 1992). Items are either single trait (unidimensional), featuring high and low levels of a single trait, or mixed-trait (multi-dimensional), with the images showing high levels of two different traits. The measure, therefore, contains both unidimensional and multidimensional items. Study Two examined the validity of the measure based on scoring algorithms created using Lasso regression, where all 300 images (150 pairs) were entered as predictors into the model for each trait. This data-driven approach was chosen to maximise the predictive validity of the measure. In addition, supervised machine learning approaches can reflect variance from items that contribute to but are not designed to measure a trait since some facets from different traits can be similar (Speer & Delacruz, 2021) (e.g. excitement-seeking from extraversion and adventurousness from openness to experience). This study extends the previous findings, comparing the performance of multiple scoring approaches, including the initial scoring algorithms, in terms of convergent and divergent validity with the IPIP-NEO-120 (J. A. Johnson, 2014) and generalisability beyond the training data.

Participants

The same sample as for Study Two was used for this study, namely 431 compensated respondents recruited using Prolific Academic (222 female; 356 under 40 years old; 209 White, 73 Black, 66 Asian, 56 Hispanic, 14 Mixed Race). Respondents took the 150-item image-based assessment along with the IPIP-NEO-120 (J. A. Johnson, 2014). The 150 items described in this study were previously selected from a larger item pool based on a sample of 300 compensated respondents (Mage = 31.14, SD = 9.26, 69% female) who took the IPIP-NEO-120 along with 100 of the image-based items, as described in Study One. Based on this sample, the 150 best-performing items were selected based on Cohen's *d* values, which represented the difference in the Big Five scores of respondents selecting image one versus image two, where items with higher Cohen's *d* values performed better and were therefore selected to be retained. These values also enabled the mapping of items to the trait the data indicated they measured, instead of what they were intended to measure, by assigning the item to the trait corresponding to the highest Cohen's *d* value (Hilliard, Kazim, et al., 2022a)

Scoring models

In Chapter 3, a separate scoring algorithm was created for each Big Five trait, where neuroticism was reversed to emotional stability. Specifically, binarised responses to all 300 images were used as the predictor variables and IPIP-NEO-120 scores for the relevant trait as the outcome variable in Lasso models. This approach was selected as the regularisation parameter results in some predictors being removed from the model, therefore producing a model with fewer predictors that is more interpretable (McNeish, 2015; Tibshirani, 1996), allowing it to be examined whether the items intended to measure each trait were indeed the most predictive of that trait. In the current study, we compare this approach with other forms of regression and predictor combinations:

a) **Lasso all** – Lasso regression with all 300 binarised predictors entered in the model for each trait, with a separate model being created to predict each Big Five trait. Due to

regularisation, the final models each had less than 300 predictors retained. These models served as the baseline for comparison of the alternative scoring approaches examined in this study.

b) **Lasso intended** - Lasso regression using only the items designed to measure each trait during the development of the measure as predictors for each model.

c) **Lasso mapped** – Lasso regression using the items mapped to each trait using Cohen's *d* as predictors for each model.

d) **Ridge all** – Ridge regression using all 300 predictors for each model. No predictors were removed from the model by regularisation, resulting in a higher n/p ratio than a).

e) **Ridge intended** – Ridge regression using only the items designed to measure each trait during the development of the measure as predictors for each model. No predictors were removed from the model, resulting in a higher n/p ratio than b).

f) **Ridge mapped** – Ridge regression using the items mapped to each trait using Cohen's *d* as predictors for each model. No predictors were removed from the model, resulting in a higher n/p ratio than c).

g) **OLS all** – OLS regression using all 300 predictors for each model. No predictors were removed from the model by regularisation, resulting in a higher n/p ratio than a).

h) **OLS intended** – OLS regression using only the items designed to measure each trait during the development of the measure as predictors for each model. No predictors were removed from the model, resulting in a higher n/p ratio than b) but the same ratio as e).

i) **OLS mapped** - OLS regression using only the items mapped to each trait using Cohen's *d* as predictors for each model. No predictors were removed from the model, resulting in a higher n/p ratio than c) but the same ratio as f).

j) **Summative** – summed the number of times an image intended to measure a trait, as informed by expert opinion, was selected from the mixed-trait pairs and the number of times an image representing high levels of that trait was selected from the single-trait pairs. This was repeated for all five traits. The summative approach was not investigated for the mapped trait as the Cohen's *d* values did not indicate which image measured high levels of the trait – just that there was a large difference in scores of individuals selecting image one versus image two.

Analysis

Model accuracy was determined by correlating the actual and predicted scores (Cui & Gong, 2018) for the training set, while convergent validity with the IPIP-NEO-120 was determined using the correlation for the test set (30% of the data). Although the summative approach is not trained in the same way that a predictive approach is, scores were still grouped into training and test sets to allow for better comparison with the regression models. Further, the generalisability of the model was observed by examining the disparity between the correlation for the training and test sets since this can give insight into how generalisable the models are beyond the training data (Jacobucci et al., 2016). Finally, subgroup differences were examined for age, race/ethnicity, and gender using the four-fifths rule, 2 standard deviations rule, and Cohen's *d*.

Results

The descriptive statistics for each model can be seen in Table 11, where all predictive models result in a similar distribution of scores regardless of the predictors included in or retained in the model and have a similar mean value compared to the IPIP-NEO-120 scores. Since the number of items designed (openness: 49, conscientiousness: 75, extraversion: 67, agreeableness: 62, emotional stability: 47) and mapped to each trait (openness: 57, conscientiousness: 61, extraversion: 63, agreeableness: 60, and emotional stability: 59) varies, the maximum score for each trait differs accordingly for the summative scoring

approach, resulting in lower mean scores and smaller ranges for the summative models compared to the predictive models. Further, given that ipsative measures result in all test-takers having the same total score and that the total score across all five traits for the summative approach ranged from 98 to 140 (M = 121.59, SD = 8.50) and was not the same across test-takers, this demonstrates that the measure is not vulnerable to ipsative scoring as with fully multidimensional scales.

Table 11

Descriptive statistics for the questionnaire-based assessment and image-based for each scoring approach (N = 431).

	Mean	SD	Min	Max	Range	Skewness	Kurtosis
Questionnaire-based							
Openness	82.84	12.02	34	114	80	-.33	.59
Conscientiousness	86.31	15.14	16	120	104	-.54	1.05
Extraversion	75.98	15.12	11	114	103	-.30	.33
Agreeableness	90.32	13.64	13	119	106	-1.11	3.41
Emotional stability	76.10	18.69	10	120	110	-.27	-.15
a) Lasso all							
Openness	82.96	7.42	59.2	102.3	43.2	.06	-.12
Conscientiousness	86.45	10.74	52.2	110.8	58.7	-.42	-.21
Extraversion	75.81	11.54	44.2	102.0	57.8	-.14	-.54
Agreeableness	90.20	8.54	59.8	110.5	50.7	-.55	.34
Emotional stability	75.38	12.42	41.0	100.8	59.8	-.37	-.43
b) Lasso intended							
Openness	82.89	4.92	68.9	94.7	25.8	-.31	-.29
Conscientiousness	86.57	8.89	63.2	105.1	41.8	-.40	-.54
Extraversion	75.95	9.77	54.0	97.5	43.6	-.05	-.85
Agreeableness	90.41	6.75	68.2	104.9	36.7	-.58	.02
Emotional stability	75.33	10.57	50.8	94.7	43.9	-.42	-.87
c) Lasso mapped							
Openness	82.95	4.85	68.1	94.2	26.1	-.23	-.33
Conscientiousness	86.53	9.27	62.9	104.8	41.9	-.44	-.61
Extraversion	75.87	9.78	55.3	97.0	41.7	.00	-.89
Agreeableness	90.28	6.49	67.7	103.0	35.3	-.49	-.06
Emotional stability	75.33	10.52	50.0	97.5	47.5	-.37	-.90
d) Ridge all							
Openness	82.96	7.42	59.2	102.3	43.2	.06	-.12
Conscientiousness	86.45	10.74	52.2	110.8	58.7	-.42	-.21

Extraversion	75.81	11.54	44.2	102.0	57.8	-.14	-.54
Agreeableness	90.20	8.54	59.8	110.5	50.7	-.55	.34
Emotional stability	75.38	12.42	41.0	100.8	59.8	-.37	-.43
e) Ridge intended							
Openness	82.89	4.92	68.9	94.7	25.8	-.31	-.29
Conscientiousness	86.57	8.89	63.2	105.1	41.8	-.40	-.54
Extraversion	75.95	9.77	54.0	97.5	43.6	-.05	-.85
Agreeableness	90.41	6.75	68.2	104.9	36.7	-.58	.02
Emotional stability	75.33	10.57	50.8	94.7	43.9	-.42	-.87
f) Ridge mapped							
Openness	82.95	4.85	68.1	94.2	26.1	-.23	-.33
Conscientiousness	86.53	9.27	62.9	104.8	41.9	-.44	-.61
Extraversion	75.87	9.78	55.3	97.0	41.7	0.00	-.89
Agreeableness	90.28	6.49	67.7	103.0	35.3	-.49	-.06
Emotional stability	75.33	10.52	50.0	97.5	47.5	-.37	-.90
g) OLS all							
Openness	82.93	11.09	46.8	116.9	70.2	.01	.11
Conscientiousness	86.71	14.70	33.0	135.7	102.7	-.17	.25
Extraversion	75.84	14.02	34.0	109.7	75.7	-.27	-.21
Agreeableness	90.25	11.95	38.7	118.6	79.9	-.66	.84
Emotional stability	75.70	16.73	25.7	113.9	88.2	-.42	-.17
h) OLS intended							
Openness	82.92	8.10	60.4	102.9	42.5	-.23	-.28
Conscientiousness	86.59	11.77	52.3	113.7	61.4	-.31	-.26
Extraversion	75.98	12.18	45.3	106.5	61.2	-.16	-.46
Agreeableness	90.49	9.89	52.5	113.4	60.9	-.69	.55
Emotional stability	75.29	13.11	38.0	103.3	65.3	-.35	-.68
i) OLS mapped							
Openness	82.95	7.85	60.3	102.5	42.2	-.24	-.35
Conscientiousness	86.59	11.83	54.3	113.9	59.6	-.37	-.49
Extraversion	75.78	11.65	49.8	104.9	55.1	.04	-.76
Agreeableness	90.29	9.76	56.7	112.9	56.3	-.53	.03
Emotional stability	75.36	13.29	38.0	107.3	69.3	-.24	-.63
j) Summative							
Openness	20.17	6.63	3	37	34	-.10	-.32
Conscientiousness	32.35	7.00	9	50	41	-.30	-.05
Extraversion	22.37	9.23	3	46	43	.05	-.71
Agreeableness	29.58	6.17	9	43	34	-.33	.00
Emotional stability	17.12	4.09	5	30	25	-.13	-.05

Scoring algorithm comparison

The performance of each model, for both the training and test set, can be seen in Table 12. Since the summative model does not use a regression line, the error statistics and R^2 statistics cannot be calculated for this scoring approach. In general, the machine learning based models had the highest convergent validity with scores on the IPIP-NEO-120 compared to the OLS and summative approaches, supporting H2. Convergent validity with the IPIP-NEO-120, as measured by test set correlations, for model a (Lasso all) ranged from .60 for agreeableness to .78 for extraversion, model b (Lasso intended) ranged from .58 for agreeableness to .77 for extraversion, and model c (Lasso mapped) ranged from .59 for agreeableness to .76 for extraversion. Similarly, models d, e, and f (Ridge all, intended, and mapped) ranged from .60 to .78, .58 to .77, and .59 to .76, respectively, for agreeableness and extraversion. In contrast, the convergence for models g, h, and i (OLS all, intended, and mapped) ranged from .45 to .58, .52 to .72, and .58 to .68 for agreeableness and extraversion, respectively. Finally, convergence for model j (summative) ranged from .41 for emotional stability to .73 for extraversion.

Table 12

Performance of each model and predictor combination (N = 431).

Trait	Training					Test				
	r	**R^2**	**MAE**	**MSE**	**RMSE**	**Corr**	**R^2**	**MAE**	**MSE**	**RMSE**
a) Lasso all										
Openness	.77**	.56	6.24	65.63	8.10	.71**	.50	6.55	64.46	8.03
Conscientiousness	.82**	.66	6.88	82.62	9.09	.70**	.47	7.50	97.26	9.86
Extraversion	.86**	.74	5.78	61.23	7.82	.78**	.61	6.82	82.28	9.07
Agreeableness	.77**	.56	6.79	84.97	9.22	.60**	.34	8.06	103.03	10.15
Emotional stability	.80**	.63	8.92	131.35	11.46	.70**	.47	10.70	175.29	13.24
b) Lasso intended										
Openness	.60**	.31	8.02	102.74	10.14	.68**	.42	6.62	73.93	8.60
Conscientiousness	.71**	.49	8.56	124.05	11.14	.70**	.48	7.41	95.81	9.79
Extraversion	.77**	.58	7.35	98.44	9.92	.77**	.58	7.31	89.52	9.46
Agreeableness	.64**	.38	7.95	121.25	11.01	.58**	.34	8.10	103.10	10.15
Emotional stability	.66**	.43	11.03	201.86	14.21	.69**	.44	11.09	185.42	13.62
c) Lasso mapped										
Openness	.59**	.31	7.93	103.40	10.17	.68**	.40	6.80	76.80	8.76
Conscientiousness	.71**	.49	8.24	123.18	11.10	.68**	.46	7.54	98.53	9.93
Extraversion	.75**	.55	7.45	105.13	10.25	.76**	.56	7.56	93.78	9.68
Agreeableness	.64**	.37	8.13	123.12	11.10	.59**	.35	8.11	101.27	10.06
Emotional stability	.67**	.44	10.98	198.78	14.10	.67**	.41	11.52	194.88	13.96
d) Ridge all										
Openness	.77**	.56	6.24	65.63	8.10	.71**	.50	6.55	64.46	8.03
Conscientiousness	.82**	.66	6.88	82.62	9.09	.70**	.47	7.50	97.26	9.86
Extraversion	.86**	.74	5.78	61.23	7.82	.78**	.61	6.82	82.28	9.07
Agreeableness	.77**	.56	6.79	84.97	9.22	.60**	.34	8.06	103.03	10.15
Emotional stability	.80**	.63	8.92	131.35	11.46	.70**	.47	10.70	175.29	13.24
e) Ridge intended										
Openness	.60**	.31	8.02	102.74	10.14	.68**	.42	6.62	73.93	8.60
Conscientiousness	.71**	.49	8.56	124.05	11.14	.70**	.48	7.41	95.81	9.79
Extraversion	.77**	.58	7.35	98.44	9.92	.77**	.58	7.31	89.52	9.46
Agreeableness	.64**	.38	7.95	121.25	11.01	.58**	.34	8.10	103.10	10.15
Emotional stability	.66**	.43	11.03	201.86	14.21	.69**	.44	11.09	185.42	13.62
f) Ridge mapped										
Openness	.59**	.31	7.93	103.40	10.17	.68**	.40	6.80	76.80	8.76
Conscientiousness	.71**	.49	8.24	123.18	11.10	.68**	.46	7.54	98.53	9.93
Extraversion	.75**	.55	7.45	105.13	10.25	.76**	.56	7.56	93.78	9.68
Agreeableness	.64**	.37	8.13	123.12	11.10	.59**	.35	8.11	101.27	10.06
Emotional stability	.67**	.44	10.98	198.78	14.10	.67**	.41	11.52	194.88	13.96

g) OLS All										
Openness	.83**	.69	5.27	45.91	6.78	.55**	-.10	9.39	141.01	11.87
Conscientiousness	.88**	.78	5.65	53.45	7.31	.46**	-.46	12.41	267.97	16.37
Extraversion	.90**	.82	5.08	42.65	6.53	.58**	.16	9.84	177.97	13.34
Agreeableness	.82**	.67	6.15	64.04	8.00	.45**	-.18	10.34	183.95	13.56
Emotional stability	.85**	.73	7.68	96.08	9.80	.52**	.01	13.82	327.20	18.09
h) OLS intended										
Openness	.63**	.40	7.66	89.81	9.48	.59**	.30	7.64	89.95	9.48
Conscientiousness	.75**	.56	7.95	106.98	10.34	.66**	.37	8.20	115.11	10.73
Extraversion	.80**	.64	7.13	84.17	9.17	.72**	.50	7.77	104.99	10.25
Agreeableness	.67**	.45	7.71	107.98	10.39	.52**	.12	8.73	137.59	11.73
Emotional stability	.69**	.47	10.63	187.59	13.70	.66**	.40	11.24	196.51	14.02
i) OLS Mapped										
Openness	.62**	.39	7.46	91.48	9.56	.56**	.27	7.74	92.82	9.63
Conscientiousness	.74**	.55	7.65	110.31	10.50	.64**	.32	8.81	124.57	11.16
Extraversion	.77**	.59	7.23	95.86	9.79	.68**	.46	8.08	114.53	10.70
Agreeableness	.67**	.45	7.64	108.12	10.40	.58**	.25	8.42	117.53	10.84
Emotional stability	.70**	.49	10.49	181.88	13.49	.63**	.35	11.58	213.49	14.61
j) Summative										
Openness	.34**	-	-	-	-	.53**	-	-	-	-
Conscientiousness	.46**	-	-	-	-	.52**	-	-	-	-
Extraversion	.66**	-	-	-	-	.73**	-	-	-	-
Agreeableness	.52**	-	-	-	-	.54**	-	-	-	-
Emotional Stability	.43**	-	-	-	-	.41**	-	-	-	-

Note. *r* = Pearson correlation coefficient for actual and predicted scores*;* R^2 = proportion of variance explained; MAE = mean absolute error; MSE = mean squared error; RMSE = root mean squared error.

Further, the model performance for the training and test sets was more similar for the machine learning approaches compared to the OLS approach, indicating greater generalisability of the machine learning models to unseen data – likely due to the OLS models overfitting as a result of the small n/p ratios. Interestingly, although the summative approach (model j) did not use regression so cannot have generalisability in the same way as other approaches, there is still some disparity between the training and test set correlations.

Discriminant validity measures the extent to which an assessment of one construct measures a different construct, where an assessment of one construct should not be strongly

related to another construct if they are theoretically distinct (D. J. Hughes, 2017) and can be measured using the multitrait-multimethod approach (Campbell & Fiske, 1959), in this case by correlating scores generated by each model with IPIP-NEO-120 scores.

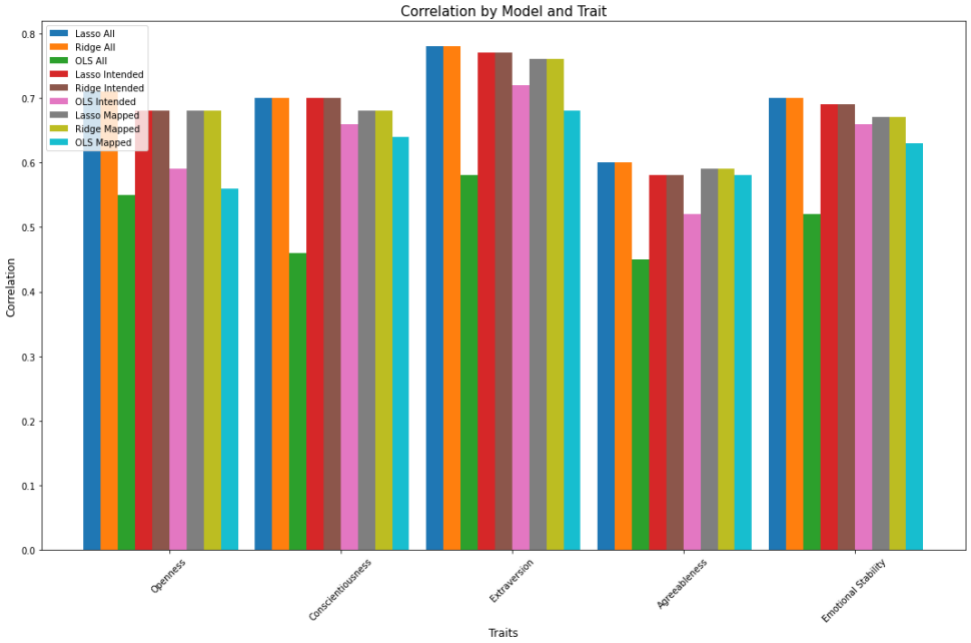
Across all predictor combinations and scoring approaches, heterotrait-heteromethod correlations are lower than convergent correlations, indicating good convergent validity of the scoring algorithms and assessment. Interestingly, heterotrait-heteromethod correlations are often higher for the mapped models (models c, f, and i) than the all or intended models, although they do not exceed the convergent correlations.

Predictor comparison

As can be seen in Figure 4**Error! Reference source not found.**, across all traits, the machine learning models that included all 300 images as predictors for each trait generally had the strongest convergent validity, followed by the models using images intended to measure each trait and images mapped to each trait. However, for the OLS models, the convergent validity for those using all 300 images was lower than the mapped and intended images, which had fewer predictors, likely due to overfitting caused by a lower n/p ratio for the 300 predictor models.

Figure 4

Test-set convergent validity for different all, intended, and mapped images for each trait.



Interestingly, the Lasso and Ridge models had similar performance across all predictor combinations and traits despite the fact that the regularisation parameter λ resulted in some predictors being removed from the Lasso models, where Table 13 shows the number of predictors retained by each model. Since the Lasso models used the fewest predictors but had the highest convergent validity compared to the OLS and summative approaches and were more generalisable, this demonstrates the benefits of using machine learning based approaches in predictive measures when measuring personality through alternative formats. Moreover, the similar performance of Ridge and Lasso highlights that by using a machine learning, data-driven approach, similar or better performance can be obtained using fewer predictors.

Indeed, for the Lasso all approach (a), some of the predictors were retained by multiple models. As such, personality can be rapidly measured through a small number of items that are relevant to multiple traits, meaning that there is potential to reduce the assessment length. Further, models contained a mixture of both multidimensional and unidimensional images (see Appendix C), retaining the structure of the assessment even when some images were removed from the models and therefore avoiding ipsative scores.

Table 13

Predictors retained by each predictive model and completion time estimates for an assessment with the respective number of (unique) items.

Model	O	C	E	A	ES	Total (unique)	Estimated completion Time (Mins)
a) Lasso all	13	26	32	23	30	102	2.5
b) Lasso intended	5	10	21	17	13	66	1.25
c) Lasso mapped	6	15	20	17	12	70	1.5
d) Ridge all	300	300	300	300	300	300	5
e) Ridge intended	49	75	67	62	47	300	5
f) Ridge mapped	57	61	63	60	59	300	5
g) OLS all	300	300	300	300	300	300	5
h) OLS intended	49	75	67	62	47	300	5
i) OLS mapped	57	61	63	60	59	300	5
j) Summative	49	75	67	62	47	300	5

Note. O, C, E, A, and ES refer to openness, conscientiousness, extraversion, agreeableness, and emotional stability, respectively.

Subgroup differences

To investigate whether scoring algorithm type and predictor combination affected subgroup differences, adverse impact analysis was carried out, to examine if there were differences in selection rates for different subgroups based on age, gender and race/ethnicity (De Corte et al., 2007). To examine the potential for adverse impact, differences in scores

based on age (binarised into below/above 40 in line with the Age Discrimination in Employment Act), gender, and race/ethnicity were examined using three widely used metrics:

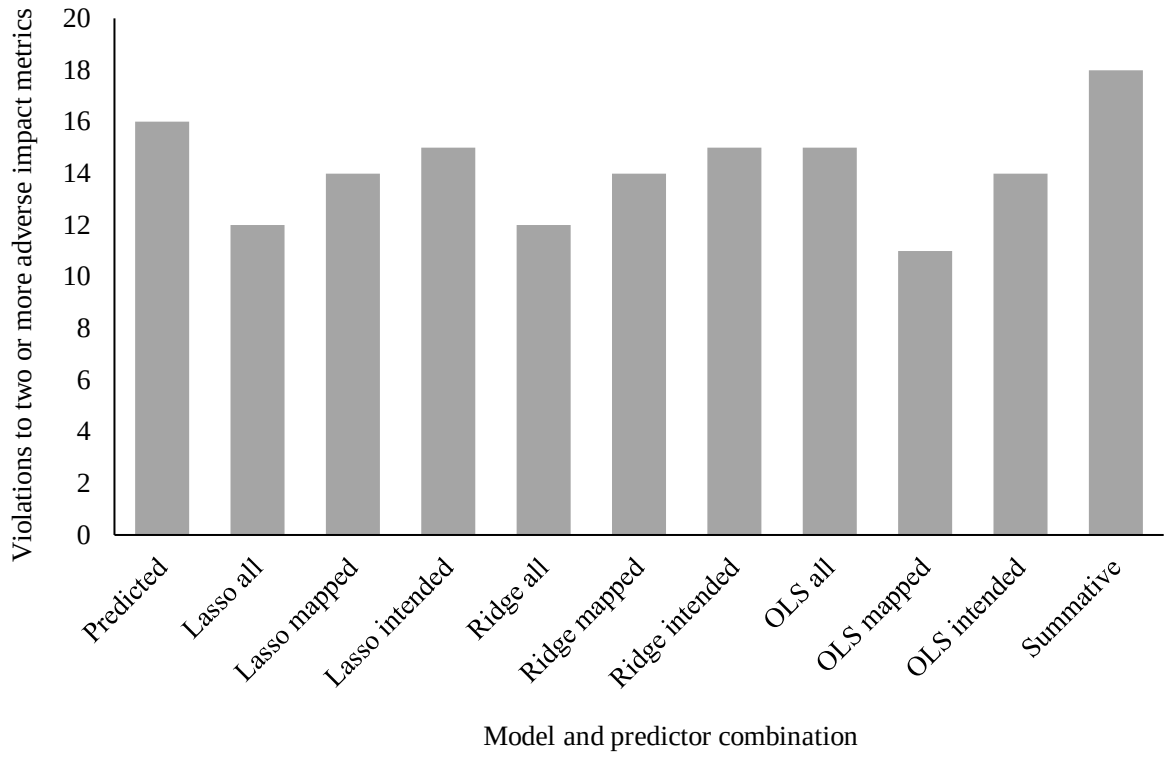
- **Four-fifths rule** – compares the pass rates of subgroups to the group with the highest rate to calculate an impact ratio, where ratios below .80 can indicate adverse impact (EEOC, 1978)
- **Two standard deviations rule** (also known as the z-test) – compares the expected and observed pass rates of each group based on the proportion of data that each subgroup represents, where values >2 indicate that there is a statistically significant discrepancy in expected and observed pass rates (D. Morgan, 2010; S. B. Morris & Lobsenz, 2000). Overpowered when sample size exceeds 100.
- **Cohen's *d*** – a measure of effect size of the difference between means, where values above .20, .50, and .80 indicate small, medium, and large effect sizes, respectively (Cohen, 1992). The current study used a threshold of +/-.30 as indicative of group differences.

Since the scores generated by the algorithms were continuous, scores for each trait were binarised into pass/fail using the median score for each trait, meaning 50% of respondents passed for each trait, in line with the enforcement rules for New York City Local Law 144 (DCWP, 2023; The New York City Council, 2021) .

Due to the fact that the three metrics can result in discrepant findings (Hilliard, Kazim, et al., 2022a; S. B. Morris & Lobsenz, 2000) , exceptions were flagged if two or more metrics were violated (Accepted Adverse Impact Ratio: >.8., accepted Cohen's D: < |.20|. accepted 2 SD: <|2|). The full adverse impact analysis can be seen in Appendix E. As can be seen in Figure 5, the summative approach resulted in the most violations while the OLS mapped resulted in the least, although this could represent the minimisation of genuine

subgroup differences. In contrast, the Lasso and Ridge mapped and intended models resulted in a similar number of violations to the questionnaire-based measure.

Figure 5
Number of times two or more adverse impact metrics were violated across all traits per model.



Discussion

This study aimed to compare the performance of different approaches to scoring a forced-choice, image-based assessment of personality. Specifically, it compared machine learning based (Lasso and Ridge regression), OLS regression based and summative

approaches using different predictor combinations to explore i) how the type of model impacted the performance of the assessment, ii) how different predictor combinations within these model types impacted model performance, and iii) the effect of model type and predictor combination on subgroup differences. Performance was defined as the convergent validity between the predicted scores and scores on the questionnaire-based IPIP-NEO-120, as well as the disparity between correlations between the actual and predicted scores for the data the model was trained on and data held out as an unseen sample. In this section, we compare the performance of each model and predictor combination and then provide some potential areas for further research, such as an examination of the predictive validity of the different models for predicting job performance.

Model Evaluation

In this study, we compared the performance of:

a) Lasso regression using all 300 images in the assessment as predictors for each trait.

b) Lasso regression using only the images intended to measure each trait.

c) Lasso regression using the images mapped to each trait during the assessment validation.

d) Ridge regression using all 300 images to predict each trait.

e) Ridge regression using only the items intended to measure each trait

f) Ridge regression using only the items mapped to each trait

g) OLS regression using all 300 predictors for each trait

h) OLS regression using only the items intended to measure each trait

i) OLS regression using items mapped to each trait

i) A summative approach using images designed to measure each trait.

A separate model was created for each of the Big Five personality traits. Comparing the different models, the machine learning based approaches had the greatest convergent

validity with questionnaire-based scores compared to the OLS regression based approach across predictor combinations, as well as the summative approach, supporting H2. Moreover, while the summative approach appears to have a large range of convergence values, where the upper end is more in line with the convergence for the other models, this is due to a particularly high correlation for extraversion (r = .73). In contrast, the correlation for emotional stability is particularly low at r = .41 and for the remaining traits, convergent validity is around .51.

The machine learning models also demonstrated greater generalisability, determined by comparing correlations for the training and test set since the test set acts as an unseen sample (Jacobucci et al., 2016). Indeed, for the machine learning models, there were smaller disparities in the performance of the training data and test data for the machine learning models compared to the OLS models, likely due to the OLS models overfitting the data as a result of the small n/p ratio (McNeish, 2015).

In terms of discriminant validity, none of the models had higher correlations with other traits than the traits they were intended to measure, indicating that using the same predictor across multiple models does not cause the resulting scores to be too similar. Furthermore, the machine learning models that included all 300 images as predictors for each trait generally had the strongest convergent validity compared to other scoring approaches and predictor combinations. This was followed by the machine learning models using images intended to measure each trait and images mapped to each trait.

While the convergent validity of the Ridge and Lasso models were similar, due to Lasso removing predictors from the model, the Lasso models resulted in the same performance but with fewer predictors compared to Ridge. For the OLS models, the convergent validity for those using all 300 images was lower than the mapped and intended images, which had fewer predictors, likely due to overfitting caused by a lower n/p ratio for the 300 predictor models.

Finally, subgroup differences were investigated using the four-fifths rule, Cohen's *d*, and two standard deviations rule, where a violation of two or more metrics was flagged as a subgroup difference. The scores generated by the Lasso and Ridge mapped and intended models resulted in a similar number of exceptions to scores on the IPIP-NEO-120, indicating that these differences may be genuine and not an artefact of the assessment format or scoring algorithm. The summative approach resulted in the most violations, where differences may not reflect genuine differences in personality. The OLS mapped model had the fewest violations. While this could be a positive finding in terms of fairness, it could also be the case that the models are diminishing genuine subgroup differences.

Due to the generally better performance of the machine learning based scoring models, we recommend that a machine learning based approach to scoring forced-choice measures of personality, particularly those of an image-based format, is a viable option. This finding is in line with previous findings that the convergent validity of forced-choice personality measures is stronger when they are scored using supervised machine learning approaches, as opposed to typical forced-choice scoring approaches (Speer & Delacruz, 2021). As well as the better performance of the machine learning models in this study, Lasso has the additional benefit of removing predictors from the models, leaving only those with the greatest predictive power and resulting in a more interpretable model (Tibshirani, 1996), as well as allowing shorter measures to be derived. This therefore has implications for personality measurement, where more complex assessments based on non-traditional formats or data sources that traditional scoring approaches are not sophisticated enough for can be created and scored using machine learning techniques increasing opportunities for innovation.

Limitations and future directions

While this study demonstrates the potential of machine learning based scoring to be used to score psychometric assessments, particularly those with a forced-choice, image-based

format, there are limitations that must be considered. For example, this data did not include any measures of job performance, meaning that the predictive validity of the different scoring approaches could not be compared. Consequently, it is not known which approach is most optimised for predicting future job performance, which is the aim of selection assessments (Barrick & Mount, 1991; Ryan & Ployhart, 2014; Schmidt et al., 2016b; Schmidt & Hunter, 1998). To address this, future research could build on the foundations laid by this study and compare how different scoring approaches affect the predictive validity of algorithmic recruitment tools.

Moreover, this study only compared two machine learning based approaches – Lasso and Ridge regression. However, other machine learning approaches have been suggested to be appropriate for scoring a forced-choice personality measure, including elastic net regression, deep neural networks, and random forest (Speer & Delacruz, 2021). Accordingly, future studies could compare different machine learning approaches to examine if the performance of the models can be improved further. Alternatively, future research could move away from machine learning based approaches and instead seek to develop an IRT-based scoring approach for unidimensional or mixed-dimensional measures since current efforts have focused on IRT approaches for fully multidimensional measures (Brown & Maydeu-Olivares, 2011, 2013).

Future research could also address one of the major limitations of the current study, where the measure was only investigated in relation to English-speaking respondents. Although it is claimed that image-based measures do not need to be redeveloped in the target language (Paunonen et al., 1990), the interpretation of image meaning may vary between cultures, meaning that the models may perform differently in other cultures. This could have a particular implication for the Lasso models since the images retained in the models for one language or culture may be less predictive of personality in another. Consequently, future

research could use a cross-cultural approach to examine the performance of the measure and each model in different cultures. Moreover, since the Lasso models reduce the number of predictors retained (Tibshirani, 1996), further research could examine how these items can be combined to create a measure that rapidly, and accurately, measures personality in around one minute.

Finally, a more purposefully-created manual approach to scoring would likely provide more robust insights into whether machine learning approaches offer advantages over manual scoring approaches. Specifically, the manual approach in this study was not underpinned by any data-driven insights and was purely based on expert opinion. As such, investigating the construct validity and factor structure of the measure to develop a data-driven manual scoring approach would help to provide a more equivalent basis for comparisons of the performance of machine learning versus manual scoring approaches.

Conclusion

This study supports the use of machine learning based scoring models for forced-choice personality assessments, particularly those designed for high-stakes contexts like selection. We found that the machine learning based Lasso models performed the best in terms of generalisability, convergent validity, and subgroup differences. Although the Ridge models performed comparatively in terms of convergent validity and generalisability, they did so with more predictors, being less conducive to a shorter assessment. The OLS models had acceptable performance but resulted in less interpretable models that retained all predictors and were less generalisable to unseen data, likely due to overfitting (McNeish, 2015), particularly with the model using all 300 predictors due to the small n/p ratio. The summative approach performed the least well, although it does not have issues with generalisability as predictive approaches can.

Based on these findings, we recommend that the best approach to scoring forced-choice personality measures, particularly if they have an image-based format, is through machine learning based predictive scoring algorithms. Specifically, we recommend Lasso regression for forced-choice assessments over OLS or summative approaches since machine learning algorithms can maximise the accuracy of the model and generalisability of models to unseen data. Moreover, through machine learning, shorter measures can be developed, allowing personality to be measured rapidly through forced-choice statements, something that is particularly true for Lasso regression. While we did not examine whether machine learning can maximise the predictive validity of a measure, our findings show promise for machine learning as a viable scoring method for forced-choice assessments of personality and highlight the possibility for innovative measures of personality to be developed and scored by machine learning.

Chapter 5. Study Four – Interviews with neurodivergent adults on experiences with pre-employment tests

"Gamification makes it feel a little bit less like a test": The potential of algorithmic recruitment tools to improve experiences of neurodivergent job applicants

Airlie Hilliard,[1,2] Franziska Leutner[1]
[1]Institute of Management Studies, Goldsmiths, University of London, New Cross, London SE14 6NW, UK
[2] Holistic AI, 18 Soho Square, London W1D 3QH, UK

Abstract

Despite representing around 20% of the population, the experiences of neurodiverse job applicants are poorly understood. Instead, the majority of literature on neurodiversity in the workplace focuses on accommodations for neurodivergent adults who have achieved employment. However, the effectiveness of technology-based training interventions for neurodivergent job seekers indicates that there is potential for technology to make pre-employment tests more accessible through novel, algorithmic formats such as game- and image-based assessments. As such, semi-structured interviews were conducted with adults with a diagnosis of ADHD, dyslexia, or autism on their experiences with pre-employment tests. Interviews distinguished between algorithmic and traditional tools, focusing on the barriers presented by each format and potential adjustments that can be made to overcome these barriers. Barriers typically revolved around sources of stress such as time pressures, fear of judgement, and issues with the compatibility of the graphics and display and result in performance being impacted and feelings of frustration. Identified adjustments focused on reducing sources of stress, such as relaxing time constraints, taking wider skills into account, gamification to make the experience less test-like, and providing feedback, potentially using artificial intelligence. Moreover, neurodivergent applicants experience unique vulnerabilities relating to the disclosure or exposure of their condition and associated bias and stigma that may impact their chances of being hired. Overall, findings indicate that there is potential for algorithmic formats to overcome barriers by including built-in accommodations, considering universal design, and making the experience more gameful.

Introduction

This study explores the experiences of neurodivergent adults with pre-employment tests, distinguishing between traditional and algorithmic formats. Here, traditional pre-employment tests include face-to-face or online interviews and questionnaire-based psychometric scales while algorithmic pre-employment tests include more novel formats such as video interviews and game- and image-based assessments scored by algorithms. Specifically, we report the findings of semi-structured interviews with 12 adults with a diagnosis of at least one of autism, attention deficit hyperactivity disorder (ADHD), or dyslexia. Although an underexplored area compared to supporting neurodiversity in education, technology-based interventions for workplace support and job seeker training have been indicated to be promising (Bozgeyikli et al., 2017; Burke et al., 2018; Rosales & Whitlow, 2019; Smith et al., 2021).

Moreover, preliminary findings indicate that neurodivergent job applicants, namely those who are autistic, are unlikely to be disadvantaged by algorithmic recruitment tools (Willis et al., 2021). However, the potential for algorithmic recruitment tools to improve the test-taking experience has not yet been explored. As such, the purpose of this study is to provide initial insights into the barriers associated with pre-employment tests for neurodivergent job applicants, how they may vary between traditional and algorithmic formats, and opportunities to reduce these barriers through adjustments and design considerations.

The chapter begins with an overview of neurodiversity in the workplace and the support programs and interventions available. It then narrows to the potential benefits of algorithmic recruitment tools, informed by findings from other technology-based interventions for neurodivergent job seekers and employees before outlining the approach to the interviews used in the current study. Findings suggest that many aspects of pre-

employment tests, such as time constraints, ambiguity, and display/graphics can be a source of stress, which consequently can affect performance and result in associated frustration. However, providing feedback, gamification, and widening the focus on tests are suggested to improve the recruitment experience. Given the customisability of algorithmic formats, this means that built-in accommodations can be relatively easily added. Accordingly, this indicates the potential for algorithmic formats to reduce barriers to entry and stress, which should be investigated by future research.

Neurodiversity in the workplace

Neurodivergence is an umbrella term that describes differences in thinking and cognition (Doyle, 2020) where the presentation and symptomology of the conditions encompassed by the term can be thought of as a continuum of differences (British Psychological Society, 2021). It is estimated that around 20% of the global population is neurodivergent (Doyle, 2020), with around 15-20% having dyslexia (International Dyslexia Association, 2016), just under 1% having autism (Baxter et al., 2015), and around 5% having ADHD (Polanczyk et al., 2007). These conditions can also co-occur, where around 40% of autistic individuals have ADHD (Rong et al., 2021) and the chance of having dyslexia is increased fourfold in dyslexic individuals (Wagner et al., 2019). Given that between 20-30% of individuals with autism or a specific or severe learning difficulty are successfully in employment (Office for National Statistics, 2021b), neurodivergent employees make up a notable proportion of the workforce – whether this has been disclosed to employers or not.

However, much of the research into neurodiversity is centred around education and how learning can be supported at school, with much less focus on neurodivergence in adulthood (Leather & Kirwan, 2012). Indeed, until recently, there has been little investigation into how neurodivergent job applicants perceive the recruitment process and their performance relative to neurotypical applicants. Likewise, there is also a lack of research into accommodations that can be introduced during pre-employment tests to reduce the barriers to

employment that individuals might face. This is a particular issue since selection assessments have been identified as a barrier to the employment of neurodivergent adults, where assessments are perceived as being designed for neurotypical people and not being inclusive of the different ways of thinking of neurodivergent applicants (Vincent & Fabri, 2022), although recent research indicates that cognitive ability tests could be a fair way of assessing neurodivergent talent (Camden et al., 2024).

Universal design in recruitment tools

Although large-scale research investigating how the recruitment process can be enhanced to be more accessible for neurodivergent job seekers is scarce, anecdotal evidence suggests that even simple tweaks can improve the performance of job seekers on pre-employment tests due to the removal of barriers unrelated to their underlying ability. For example, providing clear instructions or providing examples or walk-throughs of how to approach questions can help those who struggle with comprehension, giving them a better opportunity to show their true ability during testing (Doyle, 2023). Indeed, the seven principles of universal design, which consider how products can be designed to be usable with as many individuals as possible without the need for accommodations (The Center for Universal Design, 1997), have been applied to contexts such as education (Black et al., 2015; Courey et al., 2013; Lombardi et al., 2011) and can also be applied to pre-employment testing (Doyle, 2023; Doyle & McDowall, 2022; Rickerson, 2009). Here, universal design means that assessments (Doyle, 2023; The Center for Universal Design, 1997):

- Are accessible to individuals with diverse abilities (equitable use).
- Are able to accommodate a range of needs (flexibility in use).
- Are easy to understand and lack ambiguity (simple and intuitive use).
- Communicate necessary information in a way that is sensory-friendly (perceptible information).

- Tolerate errors and mistakes (tolerance of error).
- Result in the least amount of fatigue possible (low physical effort).
- Consider the appropriate location for testing that accounts for differences in needs (size and space for approach and use).

Specifically, in the recruitment context, flexibility in use could be implemented in the form of giving instructions multiple times over the course of the assessment in multiple formats (e.g., verbal and written; Doyle, 2023) or provided in a video-based rather than text-based format to reduce cognitive demands (Tippins, 2009). Applicants could also be allowed to take breaks or have control over the speed and order of testing (Doyle, 2023), for example.

Support for autistic job seekers

With this in mind, there is an emerging body of research examining how neurodivergent job seekers' particular needs can be supported, although much of this research is focused on supporting those who are autistic, with fewer initiatives for adults with ADHD or dyslexia. For example, autistic job seekers can benefit from a collaborative approach, where experts work with employers and autistic job seekers to adapt job descriptions and create customised positions to suit the strengths of the individual (Wehman et al., 2016). On a larger scale, programs have been set up to promote the recruitment and employment of autistic individuals, such as the Autism at Work program established in the United States by the software company SAP in 2013 to promote the employment of autistic individuals and support them in the application process (Woo, 2019). Since starting, a number of employers including Microsoft, Salesforce, JP Morgan Chase, and EY have signed up to join the scheme (Bernick, 2021; Doyle et al., 2022). However, it is estimated that only 1500 autistic employees have been hired as a result of the Autism at Work Scheme and other similar hiring initiatives (Bernick, 2021) and that there are higher rates of autistic males participating in

autism-specific hiring schemes than females (Doyle et al., 2022), questioning the effectiveness of such schemes.

Other approaches target the skills of autistic applicants, with a number of coaching interventions trialled to support the development of interview ability, focusing on how to answer interview questions (V. D. Hutchinson et al., 2019; L. Morgan et al., 2014; K. Roberts et al., 2021; Stocco et al., 2017). Such interventions can be particularly useful for autistic job seekers since interview questions can often cause anxiety due to difficulty knowing how to answer the question directly (Müller et al., 2003). This can be due to the fact that interview questions are often not direct and autistic individuals can have difficulty understanding implied meaning from indirect language (A. C. Wilson & Bishop, 2021), resulting in autistic adults being rated less favourably than neurotypical adults during job interview simulations (Maras et al., 2021; Sasson & Morrison, 2019). However, a simple solution to this could be changing how questions are asked; a study of 25 autistic and non-autistic adults found that interview performance across both groups improved when questions were revised so that interviewers first provided some context about the question and then asked specific and direct questions. The gap in ratings between autistic and neurotypical interviewees was also decreased compared to baseline performance ratings (Maras et al., 2021).

Moreover, there is emerging research examining how technology can be used to support autism-focused employment interventions, which could support autistic job seekers regardless of whether potential employers have autism-specific employment schemes. For example, a small-scale targeted intervention with six autistic participants that provided written instructions on how to prepare for interviews combined with a series of mock interviews using the video feedback tool InterviewStream and human-led feedback saw improvements in ratings of performance that were maintained at follow-up (Rosales & Whitlow, 2019). Similarly, online video interviews with real-time feedback based on the

appropriateness of answers have been used to teach effective interview skills to autistic youth, with a study of 48 transition-aged autistic adolescents finding that virtual interview training improves interview skills incrementally over traditional school-based pre-employment services such as job shadowing and workplace readiness training (Smith et al., 2021). Further, a study of the effectiveness of a virtual training agent that provided hierarchical job interview practices in areas such as greetings, small talk, and closing with virtual human interviewers found improved outcomes for face-to-face interviews compared to baseline ratings (Burke et al., 2018). Others have used video-based interventions to support email skills training for use in contacting hiring managers (Fontechia et al., 2019) and virtual-reality systems in vocational rehabilitation to teach autistic job seekers transferrable skills and conversational abilities (Bozgeyikli et al., 2017), highlighting the range of skills that can be targeted through such interventions. These technology-based interventions are especially likely to benefit younger job seekers, particularly those from generations X and Z, who are typically proficient with technology and have integrated it into a large proportion of their lives (K. R. Johnson et al., 2020).

The potential benefits of algorithmic recruitment tools for neurodivergent applicants

Using technology to support job application training also reflects the shift towards technology-enhanced and AI-driven recruitment processes in recent years. Indeed, the popularity of game-based assessments has grown in recent years (Chamorro-Premuzic et al., 2017; Winsborough & Chamorro-Premuzic, 2016), along with other alternative assessment formats such as image-based assessments (Hilliard, Kazim, et al., 2022a; Leutner et al., 2017), something that was accelerated by the pandemic (Strazzulla, 2020). The fact that these formats can reduce test-taking anxiety (Mavridis & Tsiatsos, 2017; Smits & Charlier, 2011) may be particularly beneficial for neurodivergent applicants since individuals with dyslexia and ADHD are more prone to test-taking anxiety than neurotypical populations (Lewandowski et al., 2015; Nelson et al., 2014, 2015). Further, a non-verbal format can

potentially make assessments more accessible to neurodiverse applicants, particularly those with dyslexia, since the language element is removed (De Beer et al., 2014). Although research on game-based assessments and other novel formats is still emerging, particularly in relation to neurodivergent applicants, early research indicates that alternative formats do not result in unequal outcomes for autistic job seekers. Indeed, by using two packages of games designed to measure cognitive ability – a visuospatial task combined with a number sequence or memory and maths-based game – and comparing autistic and non-autistic job seekers, Willis et al. (2021) found mixed results. While there was no difference in the performance of autistic participants and general participants for the package containing the maths and memory game, the general population scored higher on the package containing the sequencing game compared to autistic participants (Willis et al., 2021), which could reflect the spiky profile that neurodivergent individuals can display on measures of cognitive abilities. However, this study only examined test outcomes and did not examine test-taking experience.

Nevertheless, this area of research is also particularly pertinent in light of the Equal Employment Opportunity Commission's (EEOC) launch of an Artificial Intelligence (AI) and Algorithmic Fairness Initiative (EEOC, 2021) to ensure that the use of AI in hiring and employment decisions is compliant with equal opportunity laws. As part of this, the EEOC (2022) has also issued a technical assistance document on the use of AI in hiring decisions in relation to the Americans with Disabilities Act (ADA) that outlines how the use of AI, algorithms, or other software might lead to violations of the ADA and how employers can take steps to ensure that this does not happen. This can include training staff to recognise requests for accommodations, even if the term reasonable accommodation is not explicitly used, and to develop or procure alternative measures when necessary. The document also provides guidance on ensuring that the competencies measured by the assessment are job-

relevant. Adding to this guidance, the Society for Industrial and Organizational Psychology (SIOP; 2022) has also issued guidance on the use of AI in hiring, including how alternative assessment formats could impact those with disabilities and how greater transparency about the data collected by assessments could aid decisions about seeking accommodations for those who may need them. These documents serve as a reminder to employers and vendors that existing equal opportunity laws apply to algorithmic and AI-driven tools; their non-traditional format does not exempt them from existing laws and best practices governing employment decisions. Instead, algorithmic formats can require additional considerations and accommodations compared to traditional formats to ensure they are not discriminatory and as accessible to different needs as possible (EEOC 2022). As such, there is a clear need to understand the factors that could impact the accessibility of pre-employment tests, particularly those using an alternative format, how accommodations could be used to increase accessibility, and general user experience and perceptions, especially among neurodivergent adults, particularly as the use of these assessments becomes more widespread.

The current study, therefore, explores the perceptions of pre-employment tests among neurodivergent adults, with a focus on adults with a diagnosis of ADHD, dyslexia, or autism. Given that there is potential for algorithmic formats to reduce test-taking anxiety (Mavridis & Tsiatsos, 2017; Smits & Charlier, 2011) and mental ill health is an acquired neurominority condition linked to anxiety disorder (Doyle, 2020), individuals with anxiety disorder were also eligible for participation. To gain as many insights as possible, this study investigates perceptions of traditional and algorithmic pre-employment tests in general rather than focusing on a particular type of procedure. Here, a pre-employment test is defined as any selection procedure that is used to make employment decisions (EEOC, 1978; Tippins et al., 2021), including interviews, psychometric assessments, video interviews, game-based assessments, and algorithmic CV screening tools. With the aim of providing preliminary

insights into perceptions of pre-employment tests among neurodivergent applicants, this study investigates barriers to performing well in pre-employment tests and how they could be removed or accommodated, both with traditional pre-employment tests and algorithmic or AI-driven assessments.

Method

Participants

Interviewees for this study were recruited by sharing posts in LinkedIn groups for neurodivergent professionals – namely Neurodiversity @ Work, Neurodiversity in Education and the Workplace, Institute Of Neurodiversity ION Global Members, Occupational Psychology, Neurodiversity and Employment, Campaign for Disability Employment, The NeuroDiversity GiFTS (NDGiFTS), Autism Forum, and Professionals with dyslexia, dysgraphia, dyspraxia, dyscalculia – as well as on the researcher's personal LinkedIn profile. LinkedIn was selected to recruit interviewees due to a desire to gain insights from professionals, or so-called 'high-functioning' individuals, who have experience with completing pre-employment tests. Eligibility criteria to complete the interviews were: have a diagnosis of dyslexia, ADHD, or autism, and have been employed in the past five years. To maximise the potential interviewee pool and enable exploration of both reactions and perceptions, participation was not limited to those with experience with algorithmic tools. The posts shared on LinkedIn invited eligible adults to take part in a 30-minute interview about perceptions of automated recruitment tools for a doctoral research study in return for a 10-pound Amazon voucher.

Those interested in taking part in this study were able to sign up to do so by following a Calendly link in the post, which allowed them to select a time and date of their choosing that was aligned with the researcher's availability for ease of scheduling. Once a suitable time had been selected, interviewees were asked to indicate which condition(s) they had a diagnosis of and were given the option to provide additional information or request a specific

accommodation during/prior to the interview. In order to make the experience as inclusive and accessible to different needs and communication preferences as possible, and because phone calls can be particularly demanding for neurodivergent individuals (Müller et al., 2003), interviewees were able to select whether they would prefer a live video format, live voice chat, or live messaging (Cummins et al., 2020; Nicolaidis et al., 2015, 2019; Romualdez, Walker, et al., 2021). Further, interviewees were asked to provide demographic information relating to age range, gender, ethnicity, and employment status to determine the representativeness of the interview sample, with previous interviews with neurodivergent individuals collecting and reporting similar information (Hand, 2023; Romualdez, Walker, et al., 2021).

12 interviewees signed up to complete an interview (eight via video chat, two via live chat, and two asynchronously via email). All synchronous interviews were conducted via Microsoft Teams and recorded for transcription. Eight had an ADHD neurotype, four had a dyslexia neurotype, three had an autism neurotype, and two had anxiety disorder. Half (n = 6) of the interviewees who signed up had co-morbid conditions, where ADHD was the condition most commonly co-occurring. Interestingly, autism always co-occurred with ADHD in the interviewees. In addition to the targeted diagnoses, dyspraxia and dysgraphia were also reported. There was an equal proportion of males and females, and the respondents represented multiple ethnic groups. Most interviewees were between 25 and 44 (n = 9). Except for one interviewee who was a full-time student, all interviewees were either employed or self-employed at the time of the interview. See Appendix F for a full demographic breakdown.

Interview Design and Procedure

A semi-structured design was used to conduct the interviews in order to provide some standardisation to allow responses to be compared while also allowing follow-up questions to be asked to prompt interviewees to elaborate on particular aspects of their responses and

provide richer insights (Kallio et al., 2016). Although thematic analysis of the transcripts was conducted using an inductive approach guided by the data, the interview questions sought to achieve insights on:

- Barriers associated with pre-employment tests of any format.
- Whether algorithmic formats would alleviate or worsen barriers associated with traditional formats.
- Whether algorithmic formats posed additional barriers compared to traditional formats.
- Accommodations to overcome barriers.
- Whether algorithmic formats would facilitate accommodations better compared to traditional formats.

Indeed, the interview questions were designed to elicit responses touching on each of these points so that themes could be developed in terms of recurring barriers and accommodations, as well as more general reactions to or perceptions of each format (Braun & Clarke, 2006, 2022). Interviews began with a short introduction to the researcher, the doctoral research project, and how the findings from the study would be used to inform subsequent studies. Although this was specified in the information provided during sign-up, interviewees were reminded that the interview would be recorded for transcription but that they would not be shared and would be saved securely, as well as their right to withdraw at any point for any reason, including after completing the interview. One interviewee requested that the interview questions be sent ahead of time, and during the live video interviews, as well as being given verbally, questions asked were also posted in the chat to refer back to. To ensure alignment, the terms pre-employment test and algorithmic pre-employment tests were defined at the start of the interviews and examples of each were provided. These were also pasted into the chat during the video interviews.

Interviewees were then asked four main questions, starting with general perceptions or feelings about pre-employment tests before narrowing the focus to algorithmic pre-employment tests and whether views of the two formats differed. Interviewees without any personal experience with algorithmic recruitment tools were asked about their perceptions based on the definition given at the start of the interview and any other knowledge they might have, whereas those who had previously been exposed to an algorithmic format were asked to share the positive and negative aspects of their experience. Finally, interviewees were asked about any specific barriers they could identify with pre-employment tests, for either a traditional or algorithmic format. Follow-up questions were used to explore whether any identified barriers were unique to either format (traditional vs algorithmic) as well as how these barriers might be overcome. In particular, interviewees were asked whether algorithmic formats might help to overcome these barriers. At the end of the interview, interviewees were asked whether there were any additional thoughts or experiences they would like to share that they were not able to in response to the questions asked. The full interview schedule and provided definitions of (algorithmic) pre-employment tests can be seen in Appendix G. Due to the time taken to type responses, the most detailed discussion occurred via video interview, followed by email, with live chat resulting in the smallest volume of discussion.

Analysis

Video interviews were recorded and then transcribed using Microsoft's automatic transcription. Transcripts were checked for accuracy by the researcher by listening back to the recordings, and any relevant edits were made. Messages from the live chats and email interviews were used to form a transcript. Transcripts were analysed using thematic analysis to identify patterns in responses (Braun & Clarke, 2006), using an inductive approach to generate themes. Following the phases of thematic analysis outlined by Braun and Clarke (2006), the transcripts were first read thoroughly, and preliminary codes were developed. Codes were then collated into themes, which were then reviewed for consistency. Transcripts

were reviewed again to verify and refine the initial coding before the themes were named and defined.

Results

Across the interviews, a rich range of experiences emerged, with insights relating to barriers that result in stress and anxiety, associated accommodations or ways to reduce stress and anxiety and neurodivergent vulnerabilities. Indeed, although only five interviewees had first-hand experience with automated recruitment tools to the best of their knowledge, and those who had exposure to them had experienced different types of tools, all interviewees identified barriers to performance and potential accommodations or adjustments. Thematic analysis resulted in 18 codes across three themes, as can be seen in Table 14. Specifically, there were multiple sources of stress during pre-employment tests, where neurodivergent vulnerabilities are a subset of drivers of stress that were directly related to being neurodivergent.

Table 14

Themes and associated codes mapped to neurotypes and test format.

Theme	Description	Relevant interviewees	Neurotypes affected	Format affected
Source of stressful experience				
Test-taking anxiety	Pre-employment tests are a source of stress and anxiety, which can impact performance	BD, DA, EH, LC, MH, NA, RC, SS, SN, SZ	Dyspraxia, Dyslexia, Anxiety, ADHD, Autism, Dysgraphia	Both
Time constraints	Time constraints during pre-employment tests can create stress	BD, EH, MH, SS, SN, SZ	Dyspraxia, Anxiety, Autism, ADHD, Dyslexia	Both
Neurodivergent vulnerabilities				
Bias	Pre-employment tests can be biased against particular subgroups and lack inclusivity	DA, LC, MH, MK, NA, SN, SZ, TM	ADHD, Anxiety, Autism, Dyslexia, Dyspraxia	Both
Graphics	Graphics and user interface can be problematic for individuals with sensory differences when taking a pre-employment test on a device	MH, NA	ADHD, Autism, Dyslexia, Dyspraxia	Both if delivered via a screen
Ambiguity	Giving unclear or ambiguous instructions about the task or objective presents a barrier to performance	BD, LC, MH, SN	Dyspraxia, Anxiety, ADHD, Autism	Both
Narrow focus	Pre-employment tests are focussed on a narrow range of skills and can overlook additional skills that might contribute to success in the role	BD, DA, EH, LC, MH, MK, RC, SS, SN, TM	Dyspraxia, Dyslexia, Autism, Anxiety, ADHD, Dysgraphia	Both
Reluctance to disclose	Neurodivergent job seekers are often reluctant to disclose their diagnosis to their prospective employer during the assessment phase and requesting accommodations can create a dilemma	MH, MK, NA, SN	ADHD, Autism, Dyslexia, Dyspraxia	Both
Neurodivergent traits identified	The skills measured by pre-employment tests and associated areas of weakness can be associated with symptoms of neurodivergent conditions	MK, NA, RC, SN, TM	Dysgraphia, Dyslexia, Autism, ADHD, Dyspraxia	Both
Desired support				
Written communication	Providing instructions or interview questions in written form as well as providing them verbally can support performance	BD, EH, SN	Dyspraxia, Anxiety, ADHD	Both

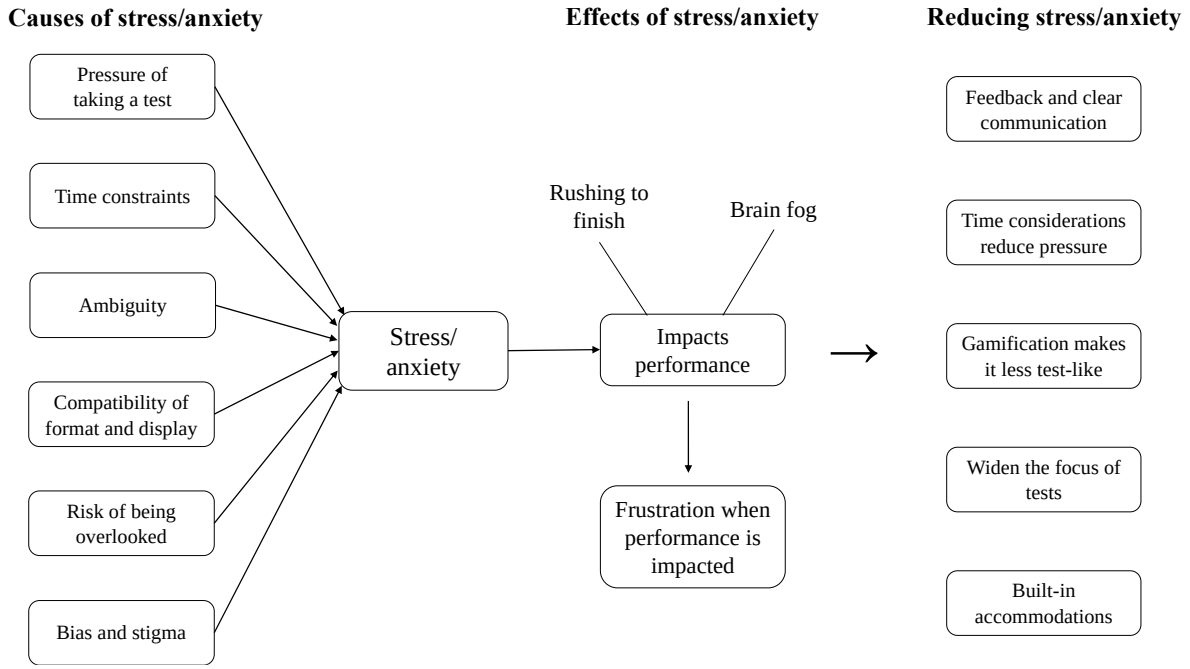
Reduce distractions	Reducing unnecessary external stimuli can reduce overstimulation and distractions to support performance	EH, MH, TM	Anxiety, Dyslexia, Autism, ADHD	Both
Time considerations	Being able to receive additional time or other time-related accommodations can reduce pressure and give time to process what is required	EH, MH, NA, SS, SN, SZ	Anxiety, Autism, ADHD, Dyslexia, Dyspraxia	Both
Preparation	Being able to prepare for selection procedures can improve applications and test performance	BD, LC, MH, MK, NA, SN	Dyspraxia, Anxiety, Autism, ADHD, Dyslexia	Both
Feedback	Receiving feedback on performance can facilitate improvement in the future	BD, LC, MK, SN	Dyspraxia, Anxiety, ADHD, Autism	Both but more potential with algorithms
Gamification	Pre-employment tests that are interactive and include game elements can make the process more enjoyable	SS, SZ, TM	ADHD, Dyslexia	Algorithmic
Widen the focus	Focusing on a wider range of skills or assessing applicants over an extended period of time can give applicants more opportunities to demonstrate their abilities	BD, DA, MH, MK, RC, SS, SN, TM, EH, LC	ADHD, Dyspraxia, Autism, Dysgraphia, Anxiety, Dyslexia	Both
Human presence in online tests				
Lack of judgement	The presence of a human during pre-employment tests can be off-putting and a source of anxiety	DA, EH, MH, NA	Dyslexia, Anxiety, ADHD, Autism, Dyspraxia	Algorithmic
Lack of support	The absence of humans when completing algorithmic tests can support performance and provide opportunities to ask for clarification	BD, LC, MK, NA, RC, SN, SZ	Dyspraxia, Anxiety, ADHD, Autism, Dysgraphia, Dyslexia	Algorithmic
Format preference				
Use in combination	Neither prefer traditional nor algorithmic pre-employment tests and instead recommend that they could be used in combination with each other	BD, MH, MK, SN, SZ, TM	ADHD, Autism, Anxiety, Dysgraphia, Dyslexia	-
Prefer algorithmic tests	Preference for algorithmic pre-employment tests over traditional formats	DA, EH, SS	Dyslexia, Anxiety, ADHD	-
Prefer traditional tests	Preference for traditional pre-employment tests over algorithmic tools	LC, NA, RC	Anxiety, Dyslexia, Autism, Dyspraxia, ADHD, Dysgraphia	-

All of the identified barriers/stressors, accommodations/relievers and vulnerabilities applied to multiple neurotypes, suggesting many are not unique to a single diagnosis group. Moreover, for the most part, these barriers, accommodations, and neurodivergent vulnerabilities largely applied to both algorithmic and traditional formats, although some barriers, such as graphics, were only applicable to non-algorithmic tests delivered via a computer screen. The main distinction between the two formats concerned human presence, where many types of traditional pre-employment tests are typically delivered in the presence of a recruiter or hiring manager, while algorithmic formats are typically taken without a human being present, usually from a location of the applicant's choosing.

A key finding from the analysis was that many of the themes and associated codes were centralised around stress and anxiety (see Figure 6). As such, the remainder of this section presents the key drivers, effects, and alleviators of stress and anxiety identified by interviewees.

Figure 6

Thematic map centralised around causes and effects of and ways to reduce stress and anxiety.



Causes of Anxiety

Pressure of taking an assessment

Pre-employment tests can be a significant source of distress, where the pure thought of taking a test can evoke stress and anxiety: “It’s a psychological barrier – anyone that’s saying you’re going to have a test” (RC). Another interviewee added, “It’s a stressful situation as well when under assessment” (SZ). Interviewees also noted that this stress and anxiety from the thought and experience of taking an assessment can be exacerbated by their neurotype: “If you’re thrown into an unfamiliar environment and there’s a lot of tension, anxiety only multiplies that” (LC).

Time constraints

Outside of general test-taking anxiety, a major driver of anxiety while completing pre-employment tests can be time pressures associated with taking psychometric assessments. Such assessments are usually completed within an allotted amount of time, or as one interviewee put it, “there’s a short amount of time, so it’s not something that I can do in my own time” (SZ). Indeed, if short periods of time are given to complete assessments, this can be perceived as insufficient, resulting in high levels of perceived pressure: “I just couldn’t do it because I was under so much pressure. When you think about it and always just not enough time” (SN). Another interviewee offered an alternative point of view, where having to do multiple tasks in short bursts of time can be easier compared to a single extended period of time in which multiple tasks are delivered with breaks in between: “If there is like a break, a little time between different things, like for most people it’s helpful, but for me, I think it would be actually detrimental because it’s like I lose the focus in there in the break” (SZ).

Ambiguity

The ambiguity surrounding the stimulation that may be encountered while completing selection procedures can also result in stress and anxiety, particularly when it comes to in-person interviews: “I don’t think we can go away with the fact that an autistic person going

into an interview [will have] a high level of anxiety because we don't know how what social cues we are going to encounter" (MH).

Other than potential sources of stimulation, ambiguity can also be created from the outset by the use of jargon and so-called buzzwords, which can alienate those who are not familiar with them: "You have these buzzwords in the job spec or the pre-employment test […] those words a bit ambiguous, a bit not clear, but it's like I feel like people out there do understand them. But I just think, what is it actually?" (BD). Moreover, interview questions can be ambiguous in that the question being asked and the desired response can lack alignment: "Sometimes in interviews as well people don't say exactly the question that they want answered" (SN).

Further, there can be a lack of information provided on procedures when delivering pre-employment tests, where interviewees highlighted that the set-up and scenario is something that "isn't always super clear" (LC). This is something that can be more of an issue with novel assessment formats such as game-based assessments that do not use a typical or familiar scenario, particularly when little information is given about why the pre-employment test is being used or what candidates must do: "You have a game. OK. What's your objective? What am I trying to achieve here? You don't always get that information" (MH).

Compatibility of format and display

Something that can drive stress and anxiety in neurodivergent individuals in particular are concerns about how compatible the test format or display will be with their needs and how it will interact with any difficulties that result from their condition. For example, one interviewee shared, with respect to "program" or computer-based pre-employment tests, that "high contrast is really hard with dyslexia" (NA). Indeed, another interviewee shared that information delivered via a computer screen can be harder for those who are neurodivergent, where text is harder to or takes longer to read on a screen compared to on paper: "It does take

maybe some time to understand written stuff in the computer. I'm very good at reading from paper and pick it up a lot quicker, but when it's the computer I don't know if it is the brightness or what it is so and it's like I struggled a bit" (MH).

Risk of being overlooked

Another major source of concern among interviewees was the risk of being overlooked and/or alienated by pre-employment tests due to differences in thinking styles. There were two major components to these concerns: the focus on narrow skills in a narrow time frame and the lack of sophistication of algorithms to be able to infer strengths.

A number of interviewees expressed that being tested in a narrow time frame that focuses on isolated performance could put them at a disadvantage if their performance in this thin slice is not reflective of their general abilities:

> "I cannot imagine that actually is indicative of like me performing well in the role itself" (LC).
>
> "It doesn't represent the working capabilities […] Every exam is a picture of that moment and doesn't represent the before or after that moment" (RC).
>
> "Pre-interview tests under the time pressure, it's not realistic. Even if someone is doing that at their job, they probably wouldn't be doing that on their first day on the job. It's just not realistic" (SH).

A focus on narrow skills was also a driver of stress, with multiple interviewees sharing views that these assessments cannot form a complete picture of an individual or capture wider or more unique skills:

> "Some abilities might not be measured by paper tests" (DA).
>
> "It might not show other abilities we have" (TM).
>
> "It's never going to capture all of the skillsets that I might have" (SS).
>
> "I think they don't really give a picture about the person. It's just trying to make a psychological profile of the person, I guess" (RC).

This can be a more salient driver of stress for algorithmic tools, particularly CV screening tools that rely on keyword matches, or as one interviewee put it, “the main thing it’s done on is previous job titles” (MK). Accordingly, multiple interviewees shared concerns about the perceived lack of sophistication of algorithms used in these tools that can lead to their transferrable skills and technical abilities being overlooked: “I’ve got a new technical qualification so therefore you would have thought that a sophisticated NLP program would be able to pick this stuff up from what I write in my CV […] yet still I was sort of blocked right left and centre” (MK).

Specifically, those with a more technical background expressed distress about the perceived simplicity of algorithmic resume screening tools in comparison with the sophisticated abilities of AI and algorithms outside of this context: “The actual technology is brilliant, but as applied to these damn HR processes, they’re rubbish and they’re stuck 20-30 years ago in terms of looking at kind of rules-based keyword attribute” (MK); “Yeah, so they’re pure data-driven, it’s an Excel table […] basically just advanced statistics” (RC).

Given that neurodivergent individuals often have non-typical interests and can take non-traditional routes through education and early career paths, one interviewee shared that they felt that this disadvantaged them: “Unless you are a candidate who has is a very traditional candidate and been doing the job role that they are recruiting for the last however many years, then your CV is going to get spat out by these applicant tracking systems” (SN).

Bias and Stigma

Finally, a significant source of stress and anxiety surrounding pre-employment tests that was shared by interviewees was concerns about bias and stigma. Bias was a concern for both traditional and algorithmic formats, with one interviewee sharing that they perceived assessment centres to be “very biased towards being liked and a popularity contest” where they are “designed for 75, 80% of the population” and “generally speaking, inclusive of differences” (MH), which causes stress due differences in ways of thinking.

Although this lack of inclusivity is not always intentional, the fear of stigma can lead to a lack of disclosure: “I didn’t declare about my ADHD because sometimes you don’t know how it’s gonna be received” (SN). This can cause stress particularly when it comes to accessing accommodations: “It’s very hard to decide. Do you disclose, do you say something? […] when you say what you need still, there’s so many biases and preconceptions” (MH); “Unfortunately, the assessment may also force me to declare SPD which I really don’t want to do” (NA). As such, some interviewees shared that they might not disclose during pre-employment tests to access accommodation, which can see prolonged stress when they are forced to either mask or disclose down the line if they obtain the position: “Now I’ll have to spend six months showing I’ve got value and then I’ll have to sort of declare it, it will be awkward” (MK).

Moreover, although both human judgements and algorithmic judgements have the potential to be biased, intentionally or not, some pointed out that the types of bias displayed by human recruiters and algorithms can differ, where there could be more of an opportunity to find a human with a different set of biases that do not disadvantage them compared to an algorithm: “Every human being is gonna have a different set of biases, with bias not necessarily bad as you know, and a different set of associations and, you know, eventually I’m going to be in front of the right person at the right time and something’s gonna click if I get through enough human beings because that’s a human condition. Whereas AI I think it’s much less diverse” (MK). However, others believed that humans can be more biased than algorithms, with bias in algorithms being able to be mitigated more effectively than human biases: “I know of like the biases in AI, but I think like human beings are probably maybe even more biased, so I don’t know. I guess it depends. I think human beings are quite biased as well” (SZ). While there were mixed opinions about the most concerning source of bias,

algorithms or humans, interviewees were in consensus that due to their neurodivergence, they were at risk of being overlooked for positions, resulting in stress and anxiety.

Human presence

One of the major differences identified between algorithmic and traditional formats was the lack of human presence when using algorithmic formats, where traditional job interviews are conducted with humans. For some, the lack of human presence served as a source of stress and anxiety where it is perceived that there is a greater chance to perform when humans are present since they can "help clarify words and sentences a lot more" (BD), where the lack of being able to do this with an algorithm can be distressing. Another interviewee shared that with humans, there is the potential to demonstrate a wider range of abilities: "People like me the more and more they get to know me […] if I'm not having the ability to have a conversation with somebody then they aren't actually getting to see the full capabilities of what I can do" (LC). This was a sentiment shared by multiple interviewees, where algorithms can cause concern about not being able to get the full view of someone like humans can: "An algorithm cannot assess the atmosphere in a meeting. It cannot assess if somebody would fit in a team in a team or not. It cannot assess if somebody has a potential to grow" (RC). One interviewee even went as far as to express that an intern or someone with no training would be better at holistically evaluating a candidate compared to an algorithm: "Even somebody on work experience with a human brain is capable of making associations and looking at someone's story on a piece of paper and thinking as a human being, like relating it to their own story and people they know, about what values someone might have to bring in a way which an AI model […] just doesn't seem to do" (MK).

Others had mixed feelings about human versus algorithmic evaluations, with one interviewee sharing that it is "easier to control how I'm perceived when interacting with a human but that doesn't necessarily make the situation better" (NA). On a similar note, another interviewee shared that being aware of how they may be perceived by humans when

they are present can be a source of anxiety: “The biggest fear is not being able to have a positive impact on someone and feeling intimidated by looks or responses from others - the fear of the unknown of what they are thinking. Removing this and replacing it with it being on a screen removes that fear” (EH).

Another interviewee also shared this sentiment about preoccupation with reactions from others, but from the perspective that processing this information while trying to articulate a response can be overwhelming: “I’m not seeing your input I’m focusing on what so if you tell me the question in a clear way, I’d rather not see you or not having to take your input at the same time because then that distracts me from my thought process trying to elaborate my answer” (MH).

Effects of anxiety

Impact on performance

One of the major implications of experiencing stress and anxiety is that it can preoccupy the minds of applicants while they are completing pre-employment tests: “You cannot not think about it. Don’t think about a pink elephant. I can’t right. I can’t forget the fact that I have an amount of time” (MH). As a result, this can impact performance: “I just couldn’t do it because I was under so much pressure” (SN).

This links closely to time pressures, where rushing due to perceived or actual lack of time can impair their ability to think through responses: “I want to see the endpoint and know what I have to do to get there. I need to be able to have enough time to think things through properly” (NA). Indeed, the pressure of a pre-employment test can mean that applicants can struggle to develop cohesive thoughts, which can mean their true abilities are not captured. As one interviewee put it, “I also find it difficult to complete some assessments based on cognitive ability and perception when I am in a stressful situation as I cannot think properly, which then gives a false representation of my abilities” (EH).

This is particularly an issue for those who are neurodivergent: “Maybe it’s my ADHD, maybe it’s just stress, maybe it’s a combination of everything. My mind just goes blank and something that I would be able to do in an hour in a day that is not in an interview setting I just couldn’t do because I was under so much pressure” (SN); “I have anxiety so I think that, you know, that does come into play as far as like how well you can concentrate, how well you’re able to formulate opinions on the spot and create cohesive answers” (LC). As such, neurodivergent individuals can feel disadvantaged compared to others who may be able to process information quicker or perform better under stress: “In the moment, your brain is trying to really forget even trying to find the information, just process first the question and sometimes I just don’t have time and then I answer it and most of the time I think I don’t know if I’ve really answered that fully […] I am processing each word in a bit more time than probably someone else” (BD).

Frustration

As a result of performance being impacted by time pressures and the associated stress and anxiety, this can result in frustration, where under normal circumstances, applicants would be able to complete the task, but their ability becomes impaired by issues with processing and brain fog. For example, one interviewee shared that “I could do all the computations, I just couldn’t process the information fast enough within the allotted time” (SS). Similarly, another interviewee shared that it can be frustrating when some of the allotted time is taken up by trying to comprehend what is required despite having the skills to complete the task if not in a high-stress context: “They gave me 45 minutes for a game assessment, but I have to understand and process what was expected of me within those 45 minutes […] I’m good [at creative problem solving], but I just run out of time to do all the things I had to do” (MH).

Alleviating stress

Time considerations

Given that time considerations are a significant source of stress and anxiety that can affect performance, multiple interviewees suggested that relaxing time constraints could help ease anxiety and improve performance. This is something that could be introduced for both algorithmic and traditional test formats. For example, one interviewee shared that "having time to able to complete these tests that are not constricted tightly I think would help to accommodate for anxiety and periods of a 'brain fog' where as much as I try, I just can't think of what I need to say or do" (EH). Similarly, another interviewee shared that time taken to plan responses should be factored into completion times and that they "need to be able to have enough time to think things through properly" (NA). For some, just knowing that they have extra time can provide peace of mind and reduce anxiety and stress by reducing preoccupation with time constraints, even if it is not used: "I did an assessment last year where I was given extra time […], I probably took about the same amount of time responding on this recent assessment, but I knew I had time if I needed to dedicate more time" (MH). On the other hand, another interviewee spoke of time accommodations in a different direction, where removing breaks between tasks could help them sustain their focus over having to move in and out of several shorter focus periods: "I would rather go do it all in one rather than having to try and focus. Sitting down and focusing once is much easier than having to like pull this numerous times anyway" (SS).

Feedback and clear communication

In addition, given that time can be spent working out what is required and comprehending instructions, something that is likely to be exacerbated by ambiguity, several interviewees called for increased transparency when conducting pre-employment tests to reduce stress and anxiety, e.g., "I found basically the more transparent and […] the more planning I have, the better" (LC). Multiple interviewees proposed that to increase

transparency, the procedure could be explained to them “one day prior” (MH) to the procedure, and that “having the questions in advance or knowing what is involved is helpful” (NA). This can also have the added benefit of allowing applicants to prepare answers in advance, particularly in the case of interviews: “I just read off what I’ve revised basically you know, cause if I’m ready for that question, I’ll just say it as I’ve written it” (BD).

Moreover, interviewees shared that stress and anxiety can be reduced by ensuring that information and instructions are constantly available during the procedure. For example, in the context of video interviews, one interviewee shared that providing written questions instead of having them delivered verbally can support performance: “Having it be on a screen rather than verbal reduced my anxiety by removing the fear of talking and messing up my words and what I am trying to say. The format of it staying on the screen whilst I answer also helps so that I don’t forget the question being asked” (EH). This was a sentiment also shared by another interviewee: “The question stayed up as a written question on the screen that I could look to and I had 20 seconds to think about the question […] with the 20 seconds, you don’t have to start talking straight away and your brain just actually like slows down and you can refer to the question because it’s written there and so that’s very, very helpful” (SN). However, this is not unique to video interviews; this could also be useful for face-to-face interviews, with one interviewee suggesting that interviewers provide a subset of the interview questions on a sheet of paper: “Perhaps they asked the question […] and they actually give you it in writing as well and then you’re just given that time to process the question, to understand it, but in writing” (BD). Having this information written down can help to support processing difficulties that can occur when under stress: “[written questions give] time to think OK right the information they’re asking here it means this, so then I can find that information in my brain if I have it for that answer” (BD).

Gamification

Another suggested that gamification can lessen stress and anxiety and support performance as it can make pre-employment tests feel less like a test: "[gamification] makes it feel a little bit less like a test" (SS); "I think using things like games as a test can be fun and less tense" (TM). As such, gamification has the potential to mitigate some of the psychological barriers associated with test-taking. One interviewee also suggested that gamification can also help to support focus on the task at hand: "Maybe like game assessments could also be engaging […] the more passive something is, the harder it is generally speaking. So, if it's a game and it's like something quite quick and I have to like do then like that could be ok. I think in general, if something is more active and engaging, that's the best thing" (SZ).

Widen the focus of tests

A penultimate way that interviewees suggested that stress and anxiety about the (perceived risk of) being overlooked for positions due to differences in thinking and pathways due to their neurodivergence was to widen the focus of pre-employment tests to capture a wider range of skills. To support this, one interviewee expressed a desire for pre-employment tests to consider their application and skills more holistically and in the context of their main job responsibilities: "In an interview, no one would know how I am. When I write because I'm very detailed […] I think if I was just given another chance to show in a different way, I could have shown more" (BD). Likewise, another interviewee suggested that the time frame over which performance is evaluated could be widened to allow skills to be captured outside of a period of intense anxiety: "I did have another positive experience where I was given some tasks to complete, and I had about a week to do it. […] Having a week, you can't stay in that period of stress for a whole week" (SN).

Accounting for neurodiversity

Finally, interviewees expressed that pre-employment tests accounting for neurodiversity and different ways of thinking could provide some relief. In particular, one interviewee implied that greater flexibility in criteria could allow for neurodivergent applicants to show their full and authentic self: "As someone with ADHD that has different interests and so on, it can be a barrier because I think people want a very simple, straightforward story […] I don't know how to put things down without feeling like I'm cutting off like half of myself" (SN). This interviewee further highlighted that their differences could help set them apart from other applicants: "People do different things and actually, I think that's my strength" (SN).

Furthermore, two interviewees suggested that built-in accommodations could benefit neurodivergent applicants: "In the ideal world ADHD and dyslexia would somehow be accounted for in the pre-employment tests, but then again we don't always want to declare our SPDs up front" (NA). Moreover, having built-in accommodations available could reduce the need to declare conditions and the associated stigma. One interviewee implied that this could be as simple as toggles to change the interface: "I'd rather use dark background or dark mode when I'm reading in the in the screens, especially when I need to do deep thinking" (MH).

Such accommodations are likely to benefit everyone, even if they are not neurodivergent. They could also be particularly helpful to those who are still learning to navigate their condition or awaiting a diagnosis and may not know they can ask for accommodations or which ones to ask for: "I have ADHD, but, you know, I was recently diagnosed in 2020. So, before that, I would have found the thing helpful, but I would not have known that is something that I could ask for" (SN).

Discussion

This study sought to explore perceptions of pre-employment tests (i.e., CV screening interviews, and psychometric assessments) among neurodivergent applicants, considering both traditional approaches to pre-employment tests and more novel algorithmic assessment formats. Interviewees shared a variety of insights into the barriers that can present when taking pre-employment tests due to their neurodivergent way of thinking, as well as suggestions for how these barriers can be overcome. Thematic analysis of the interviews resulted in 18 codes across three themes: barriers that were the *source of a stressful experience*, desired support or *adjustments that reduced stress and anxiety*, and *human presence* in online tests. Interestingly, although interviewees spoke about general test-taking anxiety, concurring with prior research finding higher levels of test anxiety in neurodivergent individuals (Lewandowski et al., 2015; Nelson et al., 2014, 2015), many of the identified barriers associated with being a neurodivergent applicant also centred around the evocation of stress and anxiety, which could impact performance and accordingly result in frustration about performance being impacted, where some barriers were specifically identified as a result of being neurodivergent. Consequently, many of the identified accommodations served to lessen stress and anxiety. This finding was applicable across diagnosis groups, with each theme affecting interviewees representing multiple neurotypes, suggesting these concerns are not limited to particular conditions.

Differences between traditional and algorithmic formats

While a range of perceptions about pre-employment tests were shared, interviewees did not indicate that they had a strong preference for either format. Instead, many of the themes applied to both traditional and algorithmic formats with the exception of concerns about algorithms lacking sophistication, which could lead to applicants being overlooked since traditional career pathways can be more challenging for those who are neurodivergent (Flower et al., 2019; Verheul et al., 2016). It is important to note that some of the concerns

were more relevant to particular types of pre-employment tests, such as in the case of algorithms lacking sophistication, with these concerns raised in the context of CV screening tools. Additionally, concerns about the compatibility of the user interface and graphics with symptoms were relevant for both traditional and algorithmic assessments, although only traditional assessments delivered via a screen. Indeed, a large-scale study of over 171,000 participants found paper-based comprehension to be stronger than digital-based (Delgado et al., 2018), meaning that interface considerations are likely to benefit all test-takers, not just those who are neurodivergent.

Furthermore, one proposed way to alleviate some stress and anxiety, gamification, is mostly applicable to algorithmic formats that typically offer greater customisation compared to traditional formats. Indeed, a number of algorithmically scored game-based or gamified assessments have been developed in recent years, both commercially and for research purposes, including game-based assessments of cognitive ability scored using gameplay data (Auer et al., 2022; Leutner et al., 2023) and gamified image-based assessments of the Big Five scored using image choices (Hilliard, Kazim, et al., 2022a, 2022b). In line with the views expressed by interviewees, gamification is associated with several benefits including less test-taking anxiety (Leutner et al., 2023; Mavridis & Tsiatsos, 2017; Smits & Charlier, 2011), greater satisfaction with test-taking experience (Georgiou & Nikolaou, 2020), and greater immersion (Leutner et al., 2021) compared to non-gamified assessments, where simply framing a task as a game can increase interest and enjoyment (Lieberoth, 2015). Game-based assessments often typically have shorter test-taking times compared to traditional assessment formats (Atkins et al., 2014; Leutner et al., 2021), something that can be supported by using machine learning to score assessments since the same or better measurement performance can be achieved with fewer features (Hilliard, Kazim, et al.,

2022b), meaning that any anxiety experienced would be for a shorter period of time than with traditional measures.

Moreover, one of the major distinctions identified between algorithmic and traditional formats was human presence. For some, the lack of human presence with algorithmic formats was a source of stress and anxiety, while for others, human presence when completing pre-employment tests was anxiety-inducing. For the interviewees expressing the former, the lack of human presence was a concern due to less opportunity to ask questions and clarify requirements as well as less opportunity to present a more holistic picture of themselves. For the interviewees expressing the latter, human presence sourced as a distraction and elicited fear of judgement over performance. The former view concurs with the majority of the previous research into perceptions of algorithmic recruitment tools (Hilliard, Guenole, et al., 2022), particularly studies investigating perceived social presence, or the extent to which interpersonal warmth and empathy are perceived during an interaction (Langer et al., 2019). For example, Kaibel et al. (2019) found that simply informing individuals that an algorithm will screen an online application instead of a human can result in the process being perceived as less personable, despite no actual interaction in the human evaluation condition. Moreover, a small-scale study with 33 professionals found that human ratings of interviews are judged as higher in human connection and interaction compared to algorithmic evaluations and were consequently preferred despite recognising that algorithmic judgements are more objective and standardised compared to human judgements (Mirowska & Mesnet, 2021). Furthermore, in concurrence with views expressed by interviewees that algorithms are less able to get a full picture of an applicant compared to humans, Lee (2018) reported that algorithms are perceived as lacking human intuition, making judgements based on keywords over skills, experience, and merit, as was the case for human reviewers.

Universal design

Several of the identified barriers and associated accommodations can be mapped onto the concept of universal design (Doyle, 2023; The Center for Universal Design, 1997) , as can be seen in Table 15. While many of these considerations can be implemented for both algorithmic and traditional formats, such as providing verbal and written instructions, algorithmic assessments often provide more flexibility since they are typically taken on an applicant's own device in a location of their choosing, meaning settings such as font size and brightness can be optimised to an individual's needs by adjusting their device settings. On the other hand, adjusting the font size or contrast between the font and the colour of the paper may not be particularly feasible, especially if there was not an opportunity to request this ahead of time.

Table 15

Potential ways to overcome real and perceived barriers mapped to elements of universal design.

Element of universal design	Removal of barriers
Equitable use	Ability to change font size and compatibility with assistive technologies such as screen readers
Flexibility in use	Time considerations and other accommodations such as font size changes for those who need them
Simple and intuitive design	No unnecessary text, graphics, or other elements that may be confusing to navigate
Perceptible information	Provide instructions verbally and in writing, allowing candidates to refer back to them as needed Provide clear feedback or debriefs
Tolerance for error	Provide a back button or the option to retake questions
Low physical effort	Use of colours, fonts, and contrasts that are easy to read (i.e., dyslexia and colour-blind friendly)
Size and space for approach and use	Allowing applicants to take tests from a location of their choosing where possible

It is important to carefully consider the impact that accommodations and general design considerations could have on the validity of assessments. For example, removing time

constraints might be something that could be implemented for personality assessments, but for cognitive ability assessments that use completion time as part of the scoring or for branching (e.g., Landers et al., 2021; Leutner et al., 2023)Landers et al., 2021; Leutner et al., 2023), the assessment could need to be redeveloped and revalidated with a different design or scoring algorithm. As such, it is important to balance validity and job relevance with accessibility to maximise the utility of the pre-employment test.

Disclosure and stigma

As well as the focus on barriers to success and ways to overcome them, interviewees also spoke about the specific vulnerabilities that come with being a neurodivergent job seeker. In the first instance, the assessment itself could serve as an identifier of neurodivergence, with ADHD related to low levels of conscientiousness, for example (Nigg et al., 2002; J. D. A. Parker et al., 2004), due to inattention symptomology, which could be particularly problematic as conscientiousness is a strong predictor of performance (Barrick & Mount, 1991; Kuncel et al., 2010; Schmidt et al., 2016a; Schmidt & Hunter, 1998; N. Schmitt, 2014). Therefore, as pointed out by interviewees, outcomes from assessments could relate to neurodivergence and associated symptoms, which may put neurodivergent applicants at a disadvantage. This emphasises the need for using job analysis to inform selection processes to ensure that constructs being measured are job-relevant and that there is a range of knowledge, skills, and other abilities measured to reduce reliance on a single measurement. For example, given that those who are neurodivergent can have higher levels of creativity (Cope & Remington, 2022; McDowall et al., 2023) , measuring this for certain jobs for which creativity is important could level the playing field for neurodivergent applicants. Further, this highlights the need for greater efforts to identify adverse impact against neurodivergent applicants, particularly if certain traits have correlations with neurodivergence. This can not only help to ensure that pre-employment tests are not biased against those who are neurodivergent, but potentially could increase trust and reduce perceptions that pre-

employment tests are not designed with neurodivergent applicants in mind. However, it is important that such actions are informed by job analysis to ensure that the abilities being assessed are relevant to the job role.

Prior to even taking the pre-employment test, applicants may find themselves in a dilemma about whether to disclose their condition in order to access accommodations if they are not built in or able to be toggled on as needed. Indeed, while neurodivergent applicants may find accommodations useful, for some, the fear of stigma outweighs any potential benefits of accommodations, leading to a lack of disclosure (Bonaccio et al., 2020; Lindsay et al., 2021; Locke et al., 2017; McDowall et al., 2023). Interestingly, neurodivergent applicants can be more inclined to disclose their diagnosis during the application process or once having started a position than during an interview, with the fear of discrimination often driving a lack of disclosure (Romualdez, Heasman, et al., 2021). These concerns could be justified, with a field experiment of over 6000 accounting positions finding that disclosure of autism, or a spinal cord disability, during application led to 25% less employer interest compared to no disclosed condition, with this gap remaining for both junior and experienced CVs (Ameri et al., 2018). However, interestingly, disclosure of conditions can help to improve perceptions and judgements of neurodivergent individuals, as evidenced by several studies into autism disclosure. For example, Sasson and Morrison (2019) report that first impressions of autistic adults from videos of completing a social challenge task are less favourable than for neurotypical adults, but ratings significantly improve when they are labelled as being autistic, with similar findings also being found for ratings of individuals with ADHD based on vignettes (Jastrowski et al., 2007). Sasson and Morrison (2019) also found that those with greater knowledge of autism also rated autistic adults labelled as such higher than those with less knowledge. Similarly, Gillespie-Lynch et al. (2015) found that an online training course that educated college students on topics such as autism diagnostic criteria and processes,

prevalence, and stigma effectively increased autism knowledge and decreased stigma from baseline. As such, one way to decrease the stigma of neurodivergent applicants could be to educate hiring managers on the conditions and their associated strengths, potentially taking inspiration from programs such as Autism at Work (Woo, 2019).

Limitations and future directions

Although the interviewees represented a diverse group of people in terms of their ethnicity, gender, and age, and multiple neurotypes were present, the sample size was relatively small, and the recruitment of interviewees via LinkedIn and the offering of Amazon UK vouchers as compensation may have limited the diversity of insights of interviewees. Nevertheless, despite the small sample size, an interpretative point of saturation (Braun & Clarke, 2019) was reached towards the final interviews, where interviewees were providing similar views and there was a lack of new insights. Notwithstanding this, future research should aim to collect insights from a more diverse group, taking a cross-cultural approach and going beyond the diagnoses focused on in the current study.

Further, given that the current study aimed to establish preliminary insights about how neurodivergent adults perceive pre-employment tests and it was desirable to have as many eligible to take part as possible, the interviews asked about (algorithmic) pre-employment tests in general and did not focus on a specific type of test. To build on these findings and provide more actionable insights to inform the design of pre-employment tests, future research could focus on how experiences vary for a particular type of pre-employment test among neurodivergent adults, particularly since perceptions can vary depending on test type and where in the funnel it is used (Hilliard, Guenole, et al., 2022). Specifically, future studies could explore the impact of assessment format for neurodivergent individuals on outcomes such as test-taking anxiety, engagement, and motivation, as well as specifically explore how assessment format interacts with neurodivergent ways of thinking to ensure pre-employment tests are as accessible to as many different needs as possible. This could also be expanded by

exploring the relationship between these outcomes and performance to help ensure equality of outcomes for neurodivergent applicants.

Conclusion

This study provides some preliminary insights into how pre-employment tests are perceived among neurodivergent adults, an area that is very much under-explored in the literature. While significant headway has been made in recent years towards more inclusive education and workplace supports, there is still a long way to go to ensure that pre-employment tests do not pose unnecessary psychological or physical barriers for neurodivergent job applicants that result in an unpleasant experience or impact performance. It is vital that pre-employment tests are created with universal design in mind and that any accommodations are informed by research and best practices in order to maximise inclusivity and reduce sources of stress and anxiety.

Novel assessment formats present a unique opportunity to ensure that those with different ways of thinking do not become alienated or experience high levels of anxiety and stress due to incompatibility with different ways of thinking. This is particularly due to their highly customisable nature that facilitates a more gameful and immersive experience and allows accommodations to be built in by considering universal design. Further research is needed to explore specific considerations that can be implemented to ensure that pre-employment tests in any format are as accessible to different needs as possible and further investigate the effect of test format on the performance of neurodivergent applicants.

Chapter 6. Studies Five and Six – test-taking experience of neurodivergent test-takers with an image-based assessment of personality

Beyond Words: Exploring Neurodivergent Experiences with Image-Based Personality Assessments

Airlie Hilliard,[1,2,*] Franziska Leutner[1]
[1] Institute of Management Studies, Goldsmiths, University of London, New Cross, London SE14 6NW, UK
[2] Holistic AI, 18 Soho Square, London W1D 3QH, UK

Abstract

Despite their use rapidly growing in recent years and their potential to benefit neurodivergent job applicants, there is a lack of understanding of how applicants experience image-based selection assessments. As such, this study sought to provide some first data on the fairness of image-based selection assessments for neurodivergent test-takers. Study Five quantitatively measured the test-taking experience of neurotypical test-takers and individuals with ADHD, dyslexia, or autism on an image- and questionnaire-based assessment, finding that neurotypical test-takers had a more positive experience than neurodivergent on both formats and that the image-based format did not improve the test-taker experience relative to the questionnaire-based format. Moreover, the questionnaire-based format was rated as more compatible with their neurotype than the image-based by test-takers with autism and ADHD and autism, while there was no difference in compatibility for other neurotypes. While it was rated as fairer, the image-based format was rated higher in external attribution and lower in concentration and ease by neurodivergent test-takers. Study Six investigated how accurate algorithms trained on the general population were when applied to neurodivergent test-takers and subgroup differences in scores. Convergent validity with the questionnaire-based measure of personality was similar for neurodivergent and neurotypical test-takers and subgroup differences were acceptable. Results provide initial insights into the effects of personality assessment format on the assessment experiences of different neurotypes. Test publishers and hiring managers may use this information to create more inclusive assessment processes. Limitations and implications are discussed.

Introduction

Applicant reactions to an organisation's selection procedures can have important implications for the talent that companies can attract and onboard given their influence on organisational attractiveness and the likelihood of an applicant accepting a job offer (Chapman et al., 2005; Hausknecht et al., 2004). Considering the current labour shortages (Office for National Statistics, 2021a) and the so-called war for talent (Chambers et al., 1998), organisations must maximise candidate experience during pre-employment testing. To address this challenge, employers are increasingly turning to novel assessment formats for their pre-employment tests, including game- and image-based assessments, many of which use algorithms to make decisions on the suitability of candidates (Albert, 2019; Guenole et al., 2023). Indeed, recent estimates suggest that around 25% of businesses are using intelligent automation in their talent management practices, particularly recruitment and hiring (Maurer, 2022), and 7% of organisations are fully automating their talent sourcing (Laurano, 2022). However, the use of algorithmic tools can raise a number of ethical concerns surrounding bias, consent, explainability, and transparency (Hunkenschroer & Luetge, 2022; Tippins et al., 2021), highlighting the need to research how to combine insights from computer science and industrial-organisational psychology to effectively mitigate bias in technical ways.

There is also a growing field of research exploring perceptions of and reactions to algorithmic recruitment tools from the test-taker perspective, particularly concerning their perceived procedural fairness. As such, we begin by exploring the procedural fairness of and reactions to algorithmic tools before exploring the benefits of image-based assessments. We then examine how neurodivergent test-takers may have differential experiences with pre-employment tests compared to neurotypical, which could impact the procedural fairness of tools, and how image-based formats may be beneficial for neurodivergent applicants due to

their benefits including reduced test-taking anxiety and support for more visual ways of thinking that are associated with being neurodivergent. Subsequently, we describe two quantitative studies.

Study Five explored the experiences of test-takers with ADHD, dyslexia, or autism who completed a questionnaire-based and image-based assessment of personality and aimed to investigate within- and between-group differences in experiences and format preferences, where test-taking experience was measured in terms of test ease, comparative anxiety, external attribution, (lack of) concentration, fairness, and motivation, and format preferences were measured through the neurodivergent compatibility scale. Study Six aimed to investigate whether scoring algorithms developed for the image-based assessment using a general population were equally accurate when applied to neurodivergent test-takers and the presence of subgroup differences. Overall, there was a lack of significant difference in the test-taking experience between the two formats, although qualitative insights suggested that the image-based format was easier to concentrate on, refreshing, and more enjoyable. For neurodivergent test-takers specifically, the questionnaire-based format elicited greater concentration, was perceived as easier, and was lower in external attribution, while the image-based assessment was perceived as fairer by neurodivergent test-takers. Additionally, only those with a diagnosis of ADHD and autism or just autism reported significant differences in format compatibility, favouring the questionnaire-based format, and dyslexic test-takers reported both formats to be more compatible with their neurotypes than these groups. Furthermore, the scoring algorithms generalised well to the neurodivergent test-takers with similar convergent validity to the training and test data as well as neurotypical test-takers from the current study and result in acceptable subgroup differences. This chapter provides first data on neurodivergent experiences with image-based formats and supports the

assumption that well-developed scoring algorithms can be applied to neurodivergent test-takers.

Fairness perceptions of algorithmic recruitment tools

In the context of selection, fairness is a social construct that represents whether test-takers perceive recruitment tools to provide equitable treatment and comparable access for individuals with differing needs, as well as whether outcomes are free from bias (SIOP, 2018). Research into fairness perceptions of recruitment tools typically centres around Gilliland's (1993) model of organizational justice, which conceptualises fairness perceptions into two broad categories: distributive justice and procedural justice. Here, distributive justice concerns equity, equality, and the fulfilment of needs while completing pre-employment tests, while procedural justice is concerned with the procedure used to conduct a pre-employment test. In the context of algorithmic recruitment tools, investigations of fairness perceptions typically focus on the latter (Hilliard, Guenole, et al., 2022), sometimes examining procedural justice in terms of specific views such as social presence, interpersonal treatment, perceived behavioural control, and consistency (Langer et al., 2019). When examining procedural justice at the overall level, findings are often mixed depending on the tool being studied and where it comes in the recruitment funnel. Indeed, algorithmic recruitment tools used at the beginning of the funnel to filter applicants are often seen as fairer than tools used at later stages in the funnel, such as for conducting interviews (Köchling et al., 2022). Moreover, although a game-based situational judgement test was judged to be fairer than a traditional equivalent (Georgiou & Nikolaou, 2020) , there is a lack of preference for algorithmically scored versus human-scored asynchronous video interviews (H. Y. Suen et al., 2019). Instead, it is likely the synchronicity that drives perceptions (Griswold et al., 2022; H. Y. Suen et al., 2019).

When examining perceptions at a more granular level, an interesting finding emerges – although algorithmic tools are perceived as more objective, they are seen as less fair due to

less perceived behavioural control (Kaibel et al., 2019). In other words, test-takers believe they are less able to influence the outcomes of algorithmic tools compared to assessments judged by human raters (Langer et al., 2019) because algorithmic tools are not able to make exceptions for candidates like humans might be able to (Hilliard, Guenole, et al., 2022). Likewise, algorithmic formats are perceived as lower in human presence, or have less of an opportunity to form an interpersonal connection (Langer et al., 2019), than assessments rated by humans even when the human-rated assessments did not involve any direct interaction with humans (Kaibel et al., 2019; Mirowska & Mesnet, 2021). However, the lack of human motive when algorithmic tools are used also gives rise to another interesting phenomenon - gender discrimination in hiring is proposed to result in less moral outrage when discrimination occurs due to an algorithm versus human biases (Bigman et al., 2022). This is because human decision-making can be driven by stereotypes and prejudices whereas algorithmic decisions cannot be prejudicially motivated (Bigman et al., 2022). Overall, there is a lack of consensus on exactly how fair algorithmic tools are perceived to be.

Reactions to novel assessment formats

The same can be said for emerging research into applicant reactions to novel and algorithmic assessment formats compared to traditional formats, where applicant reactions refer to how selection procedures are perceived and responded to by applicants (McCarthy et al., 2017). For example, Leutner et al. (2021) found that game-based measures of emotional intelligence and a video interview designed to measure conscientiousness found the novel assessment formats were judged to be significantly more immersive and better designed than questionnaire-based measures of the same traits. Similar findings have also been reported for a gamified situational judgement test; compared to a standard (non-gamified) situational judgement test, the game-based test was reported to be perceived as more satisfying and resulted in greater organisational attractiveness (Georgiou & Nikolaou, 2020). Moreover, game-based assessments elicit less test-taking anxiety compared to traditional formats

(Georgiou & Nikolaou, 2020; Mavridis & Tsiatsos, 2017) and can offer shorter test-taking times compared to traditional formats (Atkins et al., 2014; Leutner et al., 2021), therefore subjecting test-takers to any remaining test-taking anxiety for a shorter period of time and limiting the duration of high cognitive demand. These findings concur with results from sentiment analysis of online reviews of mobile apps for game-based assessments, where the overall sentiment was positive and the most commonly expressed emotion in reviews was joy (al-Qallawi & Raghavan, 2022).

However, a more recent study found conflicting findings, where reactions to a game-based assessment of cognitive ability were worse than for a pencil-based agility test in terms of face validity, predictive validity, procedural justice, opportunity to perform, procedural justice, and organisational attractiveness. Moreover, males and participants with more video game experience had more positive perceptions than females and those with less video game experience (Ohlms et al., 2023). However, these findings could be confounded by the use of a computer in the delivery of the game-based assessment, where a general preference for paper-based formats might influence perceptions of the computer-based game-based assessment. On the other hand, the previously examined studies compared alternative formats with traditional, computer-based formats such as online questionnaires. Further, there has been a move towards computer-based and proctored internet testing within the last two to three decades, even before the rise of algorithmic based assessments (Tippins et al., 2006), meaning that paper-based tests may not be an accurate representation of the current pre-employment test landscape.

Image-based formats and neurodiversity

Although there has been less research into image-based formats compared to other novel formats such as game-based and gamified assessments, like other formats, they can offer shorter testing times than questionnaire-based formats (Hilliard, Kazim, et al., 2022b) since images elicit stronger reactions and preferences compared to text (Meissner &

Rothermund, 2015), meaning it can be easier and therefore quicker to make a decision with an image-based format. This could have benefits for individuals who are neurodiverse, particularly job seekers with ADHD who may struggle to concentrate on unengaging, lengthy tests compared to neurotypical test takers (A. Mueller et al., 2017). Furthermore, the fact that image-based assessments are language agnostic could have benefits for individuals with dyslexia who can have difficulty processing written information and have a preference for a visual way of thinking (De Beer et al., 2014) by reducing cognitive demands. This may also be beneficial to individuals with ADHD who can be more reliant on visual processing due to deficits in verbal processing (Fassbender & Schweitzer, 2006). The more game-like nature of image-based assessments may also help to alleviate or reduce test-taking anxiety, which can be especially beneficial for neurodivergent applicants, who are more prone to test-taking anxiety than neurotypical populations (Lewandowski et al., 2015; Nelson et al., 2014, 2015). This is important to address since test-taking anxiety can impact performance (Hembree, 1988; McCarthy & Goffin, 2005). However, given the reliance on interpreting social cues depictured when using an image-based format and autistic individuals have an atypical approach to interpreting and processing social cues (Ashwin et al., 2015), autistic test-takers may have less positive perceptions of the image-based formats than other neurotypes. Informed by this research and the findings of Study Four, we hypothesise:

H3: There will be differences in the experiences of neurotypical and neurodivergent test-takers. Specifically:

H3.1: Neurotypical test-takers will have a more positive overall test-taking experience than neurodivergent test-takers for the questionnaire-based format.

H3.2: Neurotypical test-takers will have a more positive overall test-taking experience than neurodivergent test-takers for the image-based format.

H3.3: Neurotypical test-takers will have a more positive experience compared to neurodivergent test-takers on the image-based format across the six dimensions of test-taking experience (comparative anxiety, test-ease, concentration, fairness, external attribution, and motivation).

H4: There will be differences in the test-taking experience for the image-based and questionnaire-based formats. Specifically:

H4.1: Test-takers will rate image-based formats as having a better test-taking experience than the questionnaire-based format.

H4.2: The image-based assessment will be rated higher across the six subscales of test-taking experience (comparative anxiety, test-ease, concentration, fairness, external attribution, and motivation) compared to questionnaire-based measure.

H5: There will be differences in format compatibility for neurodivergent test-takers. Specifically:

H5.1: Neurodivergent test-takers will rate the image-based assessment more positively across the six subscales of test-taking experience (comparative anxiety, test-ease, concentration, fairness, external attribution, and motivation) compared to the questionnaire-based measure.

H5.2: Neurodivergent test-takers rate the image-based formats as more compatible with their neurotype than the questionnaire-based format.

H5.3: Test-taker neurotype will influence whether the image or questionnaire-based format is rated as more compatible with neurotype.

Moreover, machine learning based scoring algorithms for image-based assessments generalise well to unseen samples (Hilliard et al., 2022b,2022a) and recent research has indicated that algorithms trained in a low-stakes context are generalisable to high-stakes

contexts (Stevenor et al., 2024). With this in mind, and given that around 20% of the population is neurodivergent (Doyle, 2020) and therefore likely present in the data used for algorithm training, we hypothesise that:

H6: Scoring algorithms developed using the general population will generalise well to neurodivergent test-takers, having similar levels of accuracy.

Study rationale

The potential of novel assessment formats to make pre-employment tests more accessible is currently underexplored, with the only study investigating the impacts of alternative selection formats on neurodiverse candidates focusing solely on autistic individuals and the outcomes of the assessment instead of how it was experienced. Indeed, Willis et al. (2021) compared the performance of autistic and non-autistic job seekers on two packages of game-based assessments of cognitive ability and found mixed results, where there was no difference for the package containing the maths and memory game, but non-autistic test takers scored higher on the package containing the sequencing game compared to autistic participants.

The unique strengths of neurodiverse candidates, such as creativity, intense focus on interesting work, and out-of-the-box ways of thinking that can support problem-solving (Beetham et al., 2017; Cope & Remington, 2022; De Beer et al., 2014; Hoogman et al., 2020; Kannangara et al., 2018; McDowall et al., 2023; Sarkis, 2014; Sauter & McPeek, 1993; Sedgwick et al., 2019; Steele et al., 2021; Weinberg & Doyle, 2017), can be desirable to employers. As such, given the fact that experiences of pre-employment tests can influence job acceptance and organisational attractiveness (Chapman et al., 2005; Hausknecht et al., 2004), it is in the interest of both candidates and employers to determine how alternative assessment formats are perceived and experienced by neurodivergent individuals and whether these potential advantages that they might offer can be realised.

Therefore, this study sought to understand reactions to an image-based measure of personality that is scored using machine learning and designed for use in selection compared to a questionnaire-based measure of the same traits. Test-taking experiences were investigated for neurodivergent adults with a diagnosis of ADHD, dyslexia, or autism, as well as neurotypical adults. How well the scoring algorithms for the image-based format applied to neurodivergent test-takers was also investigated. Specifically, Study Five sought to understand how test-taking experience varied for neurotypical and neurodivergent test-takers, whether an image-based format improved experience relative to a questionnaire-based format, and how the two formats were experienced by neurodivergent test-takers specifically, including their compatibility with different ways of thinking. Study Six examined the performance of the scoring algorithms for the image-based assessment, that were developed using a general population, for the neurodivergent test-takers compared to neurotypical test-takers and the test set performance during algorithm training and explored the presence of subgroup differences. As such, Study Five tested H3-H5 and Study Six tested H6.

Although we realise that self-diagnosis is valid, and that official diagnosis can be difficult to obtain, a dichotomous approach was taken for this exploratory study to create more delineation and pave the way for future studies that may take a more nuanced approach (Romualdez, Walker, et al., 2021).

Study Five

Study Five sought to compare the experience of neurodivergent and neurotypical test-takers on an image-based and questionnaire-based personality assessment, investigate whether an image-based format resulted in a better test-taking experience, and how compatible the image-based format was for neurodivergent ways of thinking for adults with a diagnosis of ADHD, autism, and/or dyslexia.

Method

Participants

Participants were recruited through the panel provider Prolific Academic, with eligibility contingent on English fluency and being employed either full-time or part-time. For the panels seeking to recruit neurodivergent respondents, eligibility was also contingent on having a diagnosis of ADHD, dyslexia, or autism spectrum disorder depending on what diagnosis population the panel was looking to recruit. On the other hand, for the neurotypical panel, eligibility was contingent on not having a diagnosis of any of these conditions nor anxiety disorder, depression, other mental health diagnosis, or mild cognitive impairment to control for other neurotypes.

566 individuals successfully completed the study across the neurotypical and neurodivergent panels. Of those that were neurodivergent, 102 had a diagnosis of ADHD, 88 had autism spectrum disorder, 86 had dyslexia, 72 had ADHD and autism, 43 had dyslexia and ADHD, 15 had dyslexia and autism, and 14 had dyslexia, autism, and ADHD. Therefore, a dyslexia neurotype was present in 158 participants, ADHD in 231, and autism in 189. Additionally, there were 146 neurotypical participants.

There was a relatively even split between male (n = 270) and female (n = 275) respondents, with 20 identifying as another gender. Participant ages ranged from 18 to 73 (M = 33.23, SD = 11.49). The majority (n = 425) of participants were White, followed by Black (n = 76), Hispanic/Latino (n =24), and Asian (n = 22). Additionally, nine represented two or more ethnic groups and eight represented another ethnic group. 64 had less than one year of work experience, 78 had 1-2 years, 111 had between two and 5 years, 98 had five to ten years, and 215 had over 10 years of work experience. Most (n = 170) participants resided in the UK, with other well-represented countries including South Africa (n = 79), Poland (n = 49), the United States (n = 47), and Portugal (n = 40).

Procedure

Upon enrolling in the study through their Prolific dashboard, participants were redirected to Qualtrics where they were informed of the purpose of the research – to investigate perceptions of different formats of selection assessments – and that it was being carried out as part of a doctoral research project, where findings of the study may also be published in an academic journal. They were also provided with contact details for the research team.

After consenting, respondents then provided demographic information before completing the questionnaire- and image-based assessments and perception scales for each. At the start of each of the two personality assessments, participants were asked to imagine that they were taking the assessment as part of a job application process where success would be dependent on the outcomes. The order of completing the image- and questionnaire-based assessment was randomised to account for order effects.

To complete the image-based assessments, participants followed a hyperlink in the survey to an external webpage before returning to the previous tab to resume the survey. Participants signed up for the image-based assessment using their Prolific ID, which was embedded in the Qualtrics survey for participants to copy and paste so that their responses could be tied back to them.

Three attention checks were included in the survey, which asked participants to select a particular answer to demonstrate that they were paying attention, where those that passed two or more attention check questions were approved for the panel. The survey typically took 30 minutes to complete, with each assessment format estimated to take around six to eight minutes to complete.

Measures

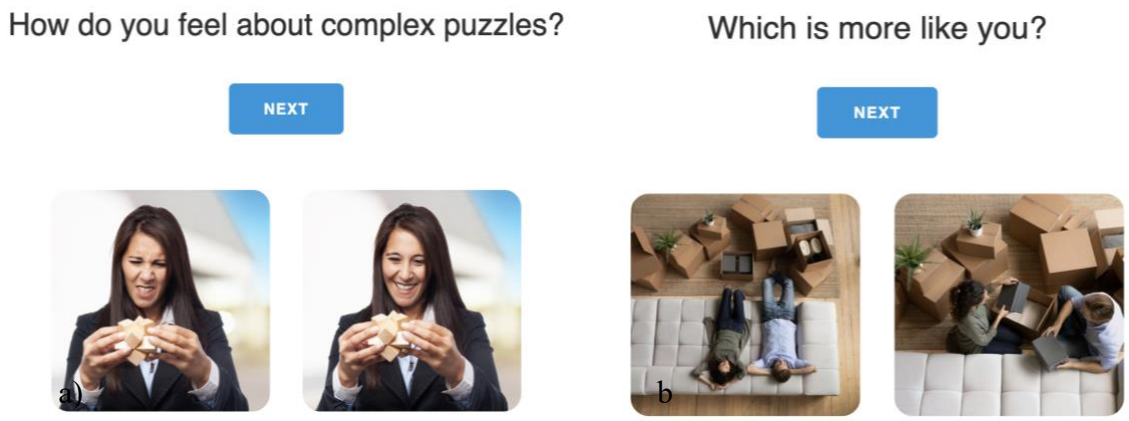
Image-based personality assessment. Based on the conceptual model of career success by Hogan et al. (2013), the image-based assessment was developed for use in

recruitment as part of an unpublished commercial project. The assessment measures a number of traits across personality and creativity, namely cognitive flexibility (Martin & Rubin, 1995), curiosity and exploration (Kashdan et al., 2009), benevolence (Thornton & Kline, 1982), core self-evaluations (Judge et al., 2003), openness to experience, agreeableness, extraversion, emotional stability (Goldberg, 1992), self-discipline, dutifulness, and achievement-striving (Costa & McCrae, 1992).

The image-based assessment presents test-takers with a series of questions that have between two and five image response options, where test-takers are asked to pick which image they identify with most. Questions either present a scenario and ask respondents how they would react/feel or simply ask which image is most like them. An example of each can be seen in Figure 7. A small number of questions were accompanied by an adjective or short phrase to reduce ambiguity. The assessment is scored using machine learning based predictive algorithms and has good convergent validity with questionnaire-based measures of the same traits: .60 for openness, .65 for achievement, .73 for extraversion, .62 for agreeableness, and .71 for emotional stability. Scores also correlate well with self-reported performance on the Individual Work Performance Questionnaire (Koopmans et al., 2014; openness: r = .35, achievement-striving: r = .53, extraversion r = .27, agreeableness r = .30, emotional stability r = .43) and the Work Effort Scale (De Cooman et al., 2009; openness: r = .34, achievement-striving: r = .57, extraversion r = .25, agreeableness r = .31, emotional stability r = .37).

Figure 7

Examples of the image-based assessment items. a) is mapped to the "Like to solve complex problems." statement from openness to experience. b) is mapped to the "Have difficulty starting tasks" statement from the self-discipline facet of conscientiousness.



Questionnaire-based assessment – The 50-item International Personality Item Pool (IPIP) scales (Goldberg, 1992) – a widely used, validated personality assessment that measures the Big Five personality traits using a five-point Likert scale – was used as the questionnaire-based assessment. The 10 conscientiousness items were swapped for 10 achievement-striving items from the IPIP representation of Costa and McCrae's (1992) NEO-PI-R Facets scale to converge with the facet-level measurement of conscientiousness in the assessment.

Test-taking experience. Test-taking experience for both formats was measured using scales and subscales selected based on themes identified from the interviews described in Study Four. As elaborated below, the majority were measured using existing subscales from the Test Attitude Survey (Arvey et al., 1990) and the Selection Procedural Justice Scale (Bauer et al., 2001). However, there was a lack of existing measures that addressed compatibility with the different ways of thinking that are associated with neurodiversity, so a custom scale was developed. Perceptions of each format were rated using the following measures on a five-point Likert scale:

- Subscales from the Test Attitude Survey (Arvey et al., 1990) that was specifically developed to measure perceptions of pre-employment tests:
    - **Motivation** – five items from the motivation subscale that measure whether respondents tried their best on the test.
    - **Concentration** – the four-item lack of concentration subscale that measures whether respondents were distracted during the test and were reversed such that higher scores were indicative of greater concentration.
    - **Comparative anxiety** – five items from the comparative anxiety subscale measuring anxiety and nervousness during the test.
    - **Test ease** – the four-item test ease subscale that measures the extent to which the assessment was too easy or too difficult.
    - **External attribution** – the five-item external attribution subscale that measures how factors such as preoccupation with time, pressure, and well-being influence performance.
- **Fairness** – the four-item chance to perform subscale from the Selection Procedural Justice Scale (Bauer et al., 2001), which was derived from Gilliland's (1993) procedural justice rules.
- **Neurodivergent Compatibility Questionnaire (**Completed by neurodivergent participants only) – a five-item scale developed using quotes from the interviews conducted in Study Four with neurodivergent adults that measures whether the format was compatible with test-takers' neurotypes. Scores were reversed such that a higher score indicated greater compatibility.
- **Open-ended responses** – participants, regardless of whether they were neurotypical or neurodivergent, were able to provide open-ended responses to four questions for each format that asked if they i) had anything to share about how the test format

affected their motivation, ii) the difficulty of the test format, iii) whether anything about the test format caused anxiety, or iv) how the test format might affect their chances of getting a job.

Results

Study Five sought to examine whether:

- There are differences in the experiences of neurotypical and neurodivergent test-takers (H3).
- There are differences in the test-taking experience for the image-based and questionnaire-based formats (H4).
- There are differences in format compatibility for neurodivergent test-takers (H5).

As a first step, the internal validity and factor structure for the neurodivergent compatibility scale were examined before each hypothesis was tested in turn.

Neurodivergent compatibility scale

The neurodivergent compatibility scale was created to measure the compatibility of the two assessments with neurodivergent ways of thinking based on the interviews described in Study Four due to a lack of an existing scale. Each statement in the scale was directly informed by quotes from the interviews, as can be seen in Table 16.

Table 16

Interview quotes informing each statement in the Neurodivergent Compatibility Scale.

Item	Interview quotes
1. The amount of text used in the assessment affected my performance due to my neurodivergent way of thinking	"There is a lot of like maybe writing is OK because it's active like lots of reading, for instance, that's hard" It does take maybe some time to understand written stuff on the computer. I'm very good at reading on paper and pick up a lot quicker, but when's the computer I don't know if it is the brightness or what it is so and it's like I struggled a bit. The first question they asked and they actually give you it in writing as well and then you're just given that time to process the question, to understand it, but in writing. the question stayed up as a written question on the screen that I could look to Having it be on a screen rather than verbal reduced my anxiety
2. The graphics used in this measure affected my performance due to my neurodivergent way of thinking	That's not to say the graphics of the program shouldn't be considered. High contrast is really hard with dyslexia so an interface that is easy for me to physically read makes the world of the difference. I'd rather use a dark background or dark mode when I'm reading the screens, especially when I need to do deep thinking.
3. I felt overstimulated due to sensory issues exacerbated by the assessment format	Just the fact that I can turn off the brightness allows me to focus on the problem itself, rather than having to use a bit of energy to ignore the brightness from the screen High contrast is really hard with dyslexia so an interface that is easy for me to physically read makes of the difference I'd rather not see you or not having to take your input at the same time because then that distracts me from my thought process trying to elaborate my answer.
4. This assessment was not designed for people with my neurodivergent way of thinking	I think pre-employment tests are generally not suitable for everyone most assessments are not designed to be inclusive.AI I think it's much less diverse
5. This assessment format is inclusive of different ways of thinking	Personality assessments are very, again they are biased towards 80% of the population right AI, I think it's much less diverse I think probably an area that is well-adjusted for bias is more reliable than a human being algorithms in general amplify human biases, and this is a very strong bias that you have just even among humans. I feel they might be biased Is there anyone/ a particular group that you feel they would be more

	biased against? Yeah Neurodivergent people like me
6. I feel that the format of the assessment would affect my chances of getting a job due to my neurodivergent way of thinking	I applied for a job a couple of times and I was very open about it and I think that they didn't hire me because I said I was autistic. In the ideal world, ADHD and dyslexia would somehow be accounted for in the pre-employment tests, but then again we don't always want to declare our SPDs up front so it's a difficult situation either way.

To examine the factor structure of the neurodivergent compatibility scale, principal component analysis was conducted for responses to the scale for the questionnaire-based measure and image-based assessment. First, correlation matrices were created for responses to each assessment format. As can be seen in Table 17, with the exception of question five, the questions have moderate intercorrelations, suggesting they measure the same underlying factor.

Table 17

Correlation matrix for questions in the custom neurodivergent compatibility scale.

	1	**2**	**3**	**4**	**5**	**6**
Question 1	-	.469**	.468**	.369**	.093*	.293**
Question 2	.536**	-	.504**	.549**	.230**	.537**
Question 3	.476**	.514**	-	.537**	.255**	.379**
Question 4	.404**	.302**	.294**	-	.491**	.596**
Question 5	.123**	.037	.043	.412**	-	.198**
Question 6	.422**	.354**	.321**	.555**	.170**	-

Note. Correlations for the image-based and questionnaire-based formats are above and below the diagonal, respectively.

The intercorrelations between the questions on the scale were further confirmed by Bartlett's test of sphericity, which had a p-value < .001 for both the questionnaire- and image-based measures, indicating that the item correlations do not form an identity matrix (M. S. Bartlett, 1950). Moreover, the Kaiser-Meyer-Olkin values for the questionnaire-based and image-based assessments were .750 and .780, respectively, above the desirable .70 threshold (Kaiser, 1974). These statistics indicated that the variables were suitable for factor analysis (M. S. Bartlett, 1950; Kaiser, 1974).

As such, principal component analysis with varimax rotation was carried out for the scale for ratings on both the image- and questionnaire-based format. As can be seen in Table 18 and Table 19, the initial analysis resulted in the extraction of one component for the image-based measure and two components for the questionnaire-based measure, where questions one to three formed one factor and questions four to six formed another.

Table 18

Component matrix for the image-based measure.

Item	Loading
Question 4	.844
Question 2	.795
Question 3	.751
Question 6	.725
Question 1	.627
Question 5	.481

Table 19

Component matrix for the questionnaire-based measure.

Unrotated solution			Varimax rotated solution		
Item	**Component 1**	**Component 2**	**Item**	**Component 1**	**Component 2**
Question 1	.774	-.210	Question 2	.821	
Question 4	.727	.463	Question 3	.788	
Question 6	.722	.170	Question 1	.768	.233
Question 2	.719	-.400	Question 5	-.136	.836
Question 3	.687	-.388	Question 4	.370	.778
Question 5	.328	.781	Question 6	.522	.527

Given the discrepancy between the factor structures for the scale for the two assessment formats and the fact that question five had low correlations with the other questions in the scale, the principal component analysis was repeated without question five. This resulted in a KMO value of .772 and .803 for the questionnaire- and image-based assessments, respectively, increasing from the initial values. The p-value for Bartlett's tests remained < .001. As seen in Table 20, principal component analysis extracted a single factor for both the questionnaire- and image-based measure, with one component explaining 53.52% and 57.86% of the variance, respectively.

Table 20

PCA factor loadings for the five-item neurodivergence scale for both assessment formats.

Item	Factor loading	
	Questionnaire-based	**Image-based**
Question 1	.788	.663
Question 2	.746	.814
Question 3	.712	.760
Question 4	.688	.814
Question 6	.720	.742

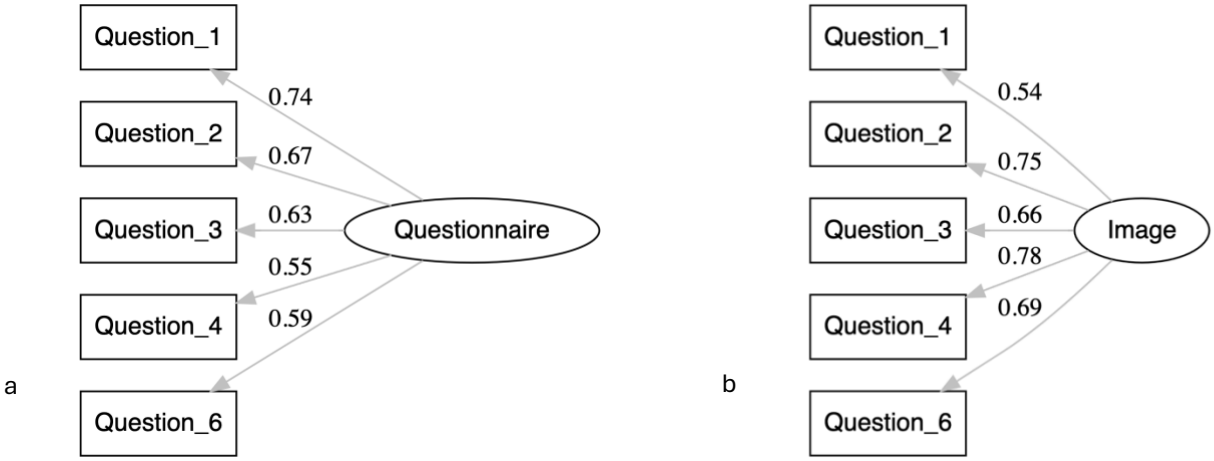
Confirmatory factor analysis was subsequently conducted for both the questionnaire- and image-based measures to further confirm the factor structure of the neurodivergent compatibility scale. As can be seen in Table 21, the model had a better fit for ratings of the image-based assessment than the questionnaire-based assessment, with the CFI value exceeding the recommended .90 threshold and the TLI value just shy of the same threshold, indicating a good model fit. On the other hand, although the SRMR values for both formats were below the recommended .08 threshold, the RMSEA value exceeded the recommended .05 threshold. Overall, the model had an acceptable fit for the two formats. Measurement Models can be seen in Figure 8.

Table 21

Confirmatory factor analysis results for the questionnaire- and image-based measure.

Format	$\chi 2$	df	RMSEA	SRMR	CFI	TLI
Questionnaire	72.15**	5	.18	.07	.88	.76
Image-based	44.15**	5	.14	.05	.94	.89

Figure 8
Measurement Model for the Confirmatory Factor Analysis (CFA) for the questionnaire-based assessment (a) and image-based assessment (b).



Furthermore, internal reliability analysis of the five-item scale resulted in an alpha value of .836 (.782 for the questionnaire-based measure and .817 for the image-based measure), with no items indicated to improve the value if removed (see Table 22).

Table 22
Internal reliability analysis for the questionnaire and image-based assessment.

Item	Questionnaire-based			Image-based		
	Corrected Item-Total Correlation	Squared Multiple Correlation	Cronbach's Alpha if Item Deleted	Corrected Item-Total Correlation	Squared Multiple Correlation	Cronbach's Alpha if Item Deleted
Question 1	.621	.404	.719	.503	.293	.811
Question 2	.565	.383	.738	.676	.460	.760
Question 3	.528	.330	.751	.609	.398	.781
Question 4	.518	.346	.754	.673	.492	.762
Question 6	.554	.369	.742	.581	.418	.789

Differences in experiences of neurotypical and neurodivergent test-takers

H3 predicted that there would be differences in the overall test-taking experience of neurotypical and neurodivergent test-takers for both the questionnaire-based and image-based assessment. Principal component analysis with Varimax rotation indicated that the six rating scales (motivation, concentration, anxiety, ease, fairness, and attribution) could be reduced to a single component that explained 50.33% and 42.91% of the variance for the image-based and questionnaire-based format, respectively. As such, an overall experience score was

calculated for each format. To do so, comparative anxiety and external attribution were reversed such that all scales measured a positive experience. Descriptive statistics for the overall experience of each format can be seen in Table 23.

Table 23

Descriptive statistics for overall experience for the image-based and questionnaire-based assessment.

Group	N	Mean	SD	Min	Max	Skewness	Kurtosis
Questionnaire-based format							
All	566	3.46	.38	2.23	4.50	-.06	-.11
Neurotypical	146	3.62	.36	2.71	4.33	-.16	-.29
Neurodivergent	420	3.41	.37	2.23	4.50	-.02	.02
Image-based format							
All	566	3.45	.42	1.93	4.38	-.31	.09
Neurotypical	146	3.62	.36	2.78	4.38	-.11	-.27
Neurodivergent	420	3.39	.42	1.93	4.38	-.28	.06

To examine whether neurotypical test-takers have a more positive test-taking experience than neurodivergent test-takers on the questionnaire-based assessment (H3.1) and image-based assessment (H3.2), two independent t-tests were carried out using overall experience scores. This indicated that neurotypical test-takers (M = 3.62, SD = .36) had a significantly more positive experience than neurodivergent test-takers (M = 3.41, SD = .37) on the questionnaire-based assessment with a medium effect size, $t(564) = 6.06$, $p < .001$, d = .582, 95% CI [.391, .773]. The same pattern was seen for the image-based assessment, where neurotypical test-takers (M = 3.62, SD = .36) had a significantly more positive test-taking experience than neurodivergent (M = 3.39, SD = .42) with a medium effect size $t(294.061) = 6.39$, $p < .001$, d = .568, 95% CI [.377, .759].

To further investigate the factors driving the differential experience for the image-based format and examine H3.3, a MANOVA analysis was conducted to investigate whether there were differences in ratings on the six experience scales for the image-based assessment – namely motivation, concentration, anxiety, ease, attribution, and fairness – for

neurodivergent and neurotypical test-takers. Multivariate tests indicated a statistically significant difference in scores between the two groups, Pillai's trace = .109, $F(6, 559) = 11.43$, $p < .001$, $\eta^2 = .109$**.** Pairwise comparisons with Bonferroni corrections for each of the scales resulted in the same pattern of findings. As can be seen in Table 24, neurotypical test-takers experienced greater motivation and concentration, less anxiety and external attribution, and perceived the test as fairer and easier in comparison to neurodivergent test-takers. Given that neurotypical and neurodivergent test-takers had differential experiences on both formats, H3.1, H3.2, and H3.3 were supported.

Table 24

Univariate tests for neurodivergent and neurotypical test-takers on the six experience scales for the image-based assessment (N = 566).

Experience scale	Neurotypical		Neurodivergent		Sum of Squares	Mean Square	F	η²
	Mean	SD	Mean	SD				
Motivation	4.61	.49	4.48	.60	1.93	1.93	5.81*	.01
Concentration	4.24	.94	3.69	1.18	32.95	32.95	26.12**	.04
Comparative anxiety	1.8	.63	2.39	.82	36.59	36.59	60.45**	.10
Ease	3.69	.79	3.42	.92	7.68	7.68	9.70**	.02
External attribution	1.66	.68	2.13	.87	24.32	24.32	36.01**	.06
Fairness	3.1	1.3	2.62	1.25	25.16	25.16	15.81**	.03

** Significant at the .05 level
** Significant at the .01 level

Differences in experience by format

To investigate H4.1, whether the test-taking experience was improved for the image-based format compared to the questionnaire-based format, a paired samples t-test was carried out for the test experience scores for the image-based and questionnaire-based format, for all participants. This indicated that there was not a statistically significant difference in the ratings $t(565) = 1.29$, $p = .197$. As such, H4.1 was not supported, indicating that the image-based format did not significantly improve the test-taking experience. Consequently, H4.2, which builds on H4.1, was not tested.

Differences in format compatibility for neurodivergent test-takers

To investigate H5.1, whether the image-based assessment is rated more positively across the experience sub-scales than the questionnaire-based assessment for neurodivergent test-takers, a repeated measures MANOVA was carried out with the six experience subscales. Multivariate tests indicated a main effect of format, Pillai's Trace = .199, $F(6, 414) = 17.13$, $p < .001$, $\eta^2 = .199$. Furthermore, concentration and ease were significantly higher for the questionnaire-based assessment, while the image-based assessment was rated as fairer but higher in external attribution, as can be seen in Table 25. As such, H5.1 was only supported for fairness.

Table 25

Univariate comparisons for ratings of the questionnaire- and image-based assessment by neurodivergent test-takers (n = 420).

Measure	Questionnaire-based		Image-based		Type III Sum of Squares	Mean Square	F	η²
	Mean	SD	Mean	SD				
Motivation	4.43	.63	4.48	.60	.44	.44	3.03	.01
Concentration	3.9	1.03	3.69	1.18	8.96	8.96**	15.5	.04
Ease	3.65	.78	3.42	.92	10.30	10.30**	26.28	.06
External attribution	1.85	.78	2.13	.87	16.80	16.80**	58.84	.12
Fairness	2.46	1.2	2.62	1.25	5.42	5.42**	12.19	.03
Comparative Anxiety	2.35	.82	2.39	.82	.32	.32	1.19	.00

** Significant at the .01 level

H5.2 predicted that the image-based assessment would be more compatible with neurodivergent ways of thinking than the questionnaire-based assessment. To examine this, a paired samples t-test was carried out comparing neurodivergent compatibility scores on the two formats. This indicated that the questionnaire-based assessment ($M = 3.64$) was rated as significantly more compatible with neurotype than the image-based assessment ($M = 3.46$), $t(419) = 3.851$, $p < .001$, $d = .188$, 95% CI [.091, .284]. Consequently, H5.2 was not supported.

Finally, H5.3 predicted that there would be differences in the compatibility of formats by neurotype. To test this, a repeated measures ANOVA with diagnosis as a between-subjects variable was carried out. In line with the findings above, multivariate tests indicated that the questionnaire-based format was rated as significantly more compatible with neurotype than the image-based format, $F(1, 413) = 4.48$, $p = .035$, $\eta^2 = .011$. There was also a significant effect of diagnosis on compatibility ratings, $F(6, 413) = 5.23$, $p < .001$, $\eta^2 = .071$, which was further explored using pairwise comparisons. This resulted in two significant results; test-takers with a diagnosis of dyslexia ($M = 3.89$) had significantly higher compatibility ratings than test-takers with a diagnosis of ADHD and autism ($M = 3.23$) and test-takers with only autism ($M = 3.51$), $MD = .657$, $SE = .123$, $p < .001$, 95% CI [.282, 1.033] and $MD = .376$, $SE = .117$, $p = .029$, 95% CI [.019, .733], respectively.

Moreover, there was a significant interaction effect between diagnosis and format on compatibility ratings, $F(6, 413) = 2.63$, $p = .016$, $\eta^2 = .037$. Pairwise comparisons indicated that the questionnaire-based format ($M = 3.60$) was rated as more compatible with neurotype than the image-based format ($M = 3.47$), $MD = .131$, $SE = .062$, $p = .035$, 95% CI [.009, .253]. Pairwise comparisons for format by diagnosis indicated that test-takers with a diagnosis of both ADHD and autism rated the questionnaire-based format ($M = 3.37$) as significantly more compatible with their neurotype than the image-based format ($M = 3.10$), $MD = .272$, $SE = .112$, p = .015, 95% CI [.052, .492]. The same pattern was seen for test-takers with a diagnosis of just autism, with the questionnaire-based measure (3.72) rated as more compatible than the image-based measure (3.31), $MD = .402$, $SE = .101$, $p < .001$, 95% CI [.203, .601]. For the remaining neurotypes, there was no significant difference in the compatibility ratings of the two formats, as can be seen in Table 26. Accordingly, H5.3 was supported.

Table 26
Pairwise comparisons for ratings of neurodivergent compatibility for each format by diagnosis group.

Diagnosis	Mean		Mean Difference (questionnaire-image)	Std. Error	95% CI
	Questionnaire-based	Image-based			
ADHD	3.641	3.49	.151	.094	[-.034, .336]
ADHD and autism	3.369	3.097	.272*	.112	[.052, .492]
Autism	3.716	3.314	.402**	.101	[.203, .601]
Dyslexia	3.923	3.858	.065	.102	[-.136, .266]
Dyslexia and ADHD	3.484	3.488	-.005	.145	[-.289, .28]
Dyslexia, ADHD, and autism	3.557	3.1	.457	.254	[-.042, .956]
Dyslexia and autism	3.493	3.92	-.427	.245	[-.908, .055]

* Significant at the .05 level
** Significant at the .01 level

Qualitative insights

Open-ended responses to questions about how the format impacted anxiety or motivation and likelihood of getting a job based on format were largely consistent across participants, regardless of their neurotype or whether they were neurotypical. Based on these responses, four key themes emerged:

- **Ease of faking or cheating** – given the instruction to imagine they were taking the assessments as part of a job application, participants shared that they gave responses that they "thought an employer would want to hear", but this can cause anxiety about "knowing if all my lies are consistent". On the other hand, while some participants thought that it would be easy to change their answers to make them more desirable, they were "too honest" but that their honest responses "may not be appealing to most employers". While this honesty struggle was present across both formats, there were 30 references to this struggle for the questionnaire-based format and only nine for the image-based format, indicating that the image-based format was less susceptible to faking.

- **Ambiguity of questions** – across both formats, participants raised concerns about ambiguity, where questionnaire-based formats had statements that were "too vague and up for interpretation" due to issues such as sentence structure and not being context-specific. The images were "confusing at times", which resulted in anxiety for some participants as they "didn't fully understand what the pictures were trying to say". This was particularly true for those with an autism neurotype due to the reliance on the interpretation of facial expressions and non-verbal cues, providing some additional support for H5. To minimise the ambiguity of some of the images, test-takers of multiple neurotypes suggested combining more of the images with adjectives: "Both pictures and text would have been the perfect combo"; "the ones with text were better".
- **Images support concentration –** dyslexic test-takers in particular shared that the questionnaire-based format was harder to focus on due to long lines of text and the black-and-white, monochromatic format. On the other hand, the relative lack of text in the image-based assessment "helps with dyslexia". Moreover, test-takers with dyslexic, ADHD, and neurotypical neurotypes shared that the "use of visuals kept attention", contradicting the finding that the questionnaire-based format was easier to concentrate on and providing some qualitative support for H4.
- **Refreshing image-based format** – the image-based format was "refreshing", "well-presented", and "fun". Accordingly, the image-based assessment was "more engaging" and held attention "better and longer" relative to the questionnaire-based format, thus being more motivating and providing a source of amusement. Contrastingly, the questionnaire-based format was "little dry", "very stale", "monotonous" and "tedious" across test-takers of all neurotypes. As such, this provides additional support for H4.

Participants shared that both formats were too long, particularly the image-based format, highlighting the need to balance assessment validity and depth with conciseness. However, the image-based assessment that participants completed was undergoing validation at the time of the study, meaning the image bank was purposely designed to be larger than required to allow for refinement.

Study Six

Study Six sought to investigate whether the image-based assessment scoring algorithms trained on a general population have similar accuracy for neurotypical and neurodivergent test takers, therefore investigating H6. It further established whether the assessments showed adverse impact against neurodivergent test takers.

Scoring algorithms

The assessment is scored using a machine learning based Lasso regression approach, where algorithms were trained to predict scores on the questionnaire-based measure for each outcome (i.e., IPIP scores) using image choices. To do so, each image was binarised to a dummy variable such that one indicated that the image was selected by the respondent, and zero represented not being selected. As with Study Two, this approach was selected due to Lasso's regularisation parameter that removes some predictors from the model and makes it more generalisable by reducing overfitting (McNeish, 2015; Tibshirani, 1996). The scoring algorithms were trained using a general population with a train-test split, where they were first trained on 70% of the data, with the remaining 30% acting as an unseen sample, allowing the generalisability of the models beyond the training dataset to be examined (Jacobucci et al., 2016). They were then applied to the sample from Study Five. The optimal value of lambda for each algorithm was determined using 10-fold cross-validation. The final images to include in the models from the item bank were identified using an iterative process that used Lasso regression to examine how many models the images were retained across,

where those retained by few images were removed as it was indicated that they did not perform well. The final subset contained 215 images across 94 questions, where all images were used to predict each outcome irrespective of the trait they were designed to measure since personality dimensions intercorrelate (Chang et al., 2012), so an image mapped to one area may be useful in predicting another trait (Speer & Delacruz, 2021).

Training data

The sample of the general population used to initially train the scoring algorithms was recruited through Prolific Academic, where several parallel panels were run to ensure there was good demographic representation based on ethnicity, disability, and veteran status. Eligibility for the panel was contingent on English fluency, high Prolific approval rate, and full or part-time employment, with participation limited to the countries served most by the commercial assessment (UK, Europe, US, Canada, Australia).

1833 participants who completed the new bank of images and associated outcome scales successfully passed at least two of the three attention checks included in the survey. 876 were White, 260 were Asian, 261 were Black, 280 were Hispanic, 5 were Native American or Alaskan Native, 1 was Native Hawaiian or Other Pacific Islander, and 111 were Mixed ethnicity. 811 were female and 992 were male. Age ranged from 18 to 77 (M= 33.96, SD = 11.14). 46 were veterans and 456 reported a health condition, with 195 having a condition that affects daily activity. Moreover, 111 reported a learning difference: 42 participants indicated that they had a diagnosis of dyslexia and a further 12 indicated they had dyslexia in combination with another condition. 15 had ADHD and 2 had ADHD with other conditions, and autism was reported by 3 participants.

Results

To explore the implications of using algorithms trained on the general population to evaluate neurodivergent candidates, the convergent validity and model accuracy of the

algorithms was compared for neurodivergent and neurotypical respondents and adverse impact analysis was carried out.

Algorithm performance

Descriptive statistics for the personality scores for neurotypical and neurodivergent test-takers for each format can be seen in Table 27, where neurotypical test-takers generally score higher than neurodivergent across both formats. Scores on the image-based assessment were transformed to range from one to one hundred for ease of interpretation. Scores on the questionnaire-based measures range from one to five.

Table 27

Personality scores on the two assessment formats for neurodivergent and neurotypical test-takers.

Group	Trait	Mean	SD	Min	Max	Skew	Kurtosis
Questionnaire-based							
All (*n* = 566)	Openness	3.85	.62	1.4	5	-.53	.27
	Achievement	3.67	.56	1.5	5	-.41	.16
	Extraversion	2.79	.93	1	5	.06	-.79
	Agreeableness	3.85	.72	1.1	5	-.75	.49
	Emotional stability	3.01	.97	1	5	-.04	-.78
Neurotypical (*n* = 146)	Openness	3.8	.54	2.2	4.8	-.32	.09
	Achievement	3.78	.50	1.9	5	-.38	.62
	Extraversion	3.06	.81	1.2	4.7	-.06	-.67
	Agreeableness	4	.56	2.2	4.9	-.51	.05
	Emotional stability	3.65	.68	1.8	5	-.21	-.14
Neurodivergent (*n* = 420)	Openness	3.87	.65	1.4	5	-.60	.29
	Achievement	3.63	.58	1.5	5	-.38	.01
	Extraversion	2.7	.95	1	5	.16	-.79
	Agreeableness	3.79	.76	1.1	5	-.69	.26
	Emotional stability	2.79	.95	1	5	.22	-.68
Image-based							
All (*n* = 566)	Openness	57.87	19.68	1	100	-.40	-.17
	Achievement	58.46	19.9	1	100	-.51	-.49
	Extraversion	50.71	22.54	1	100	.00	-.95
	Agreeableness	51.05	22.66	1	100	.03	-1.03
	Emotional stability	59.61	19.26	1	100	-.37	-.51
Neurotypical (*n* = 146)	Openness	58.94	16.75	1	92	-.51	.56
	Achievement	64.48	16.5	19	100	-.69	.03
	Extraversion	64.21	18.4	17	100	-.40	-.49
	Agreeableness	57.76	20.57	5	96	-.25	-.59
	Emotional stability	66.84	14.92	27	100	-.40	-.27
Neurodivergent (*n* = 420)	Openness	57.5	20.61	2	100	-.36	-.36
	Achievement	56.37	20.56	1	96	-.40	-.66
	Extraversion	46.02	21.96	1	94	.21	-.87
	Agreeableness	48.72	22.91	1	100	.16	-1.06
	Emotional stability	57.09	19.97	1	97	-.24	-.68

The performance of the algorithm for each outcome was evaluated in terms of accuracy/convergent validity by correlating the actual and predicted scores for the trait and R^2 to examine the variance in the data explained by the models. Mean absolute error, mean squared error, and root mean squared error were also calculated. Mean absolute error sums

the magnitude of errors and divides it by the number of observations, while mean squared error sums the individual squared errors and again divides this by the number of observations, and the root mean squared error takes the square root of this (Karunasingha, 2022). The larger the value of these metrics, the more error there is, or the less well the model fits the data. There is debate in the literature about the best metric to use, with some arguing that the mean absolute error should be favoured over the root mean squared error since the latter does not represent the average error well (Chai & Draxler, 2014; Karunasingha, 2022; Willmott & Matsuura, 2005), but the purpose of the metrics in the current study is to examine the magnitude of the disparity of the values between the different datasets to evaluate whether the models perform well when applied to neurodivergent test takers, particularly since the training data did not have a substantial representation of individuals with dyslexia, ADHD, or autism.

As can be seen in Table 28, for the most part, the models perform similarly for the current study compared to the test set, which can be used as an indicator of how the algorithm might perform on other unseen examples. Indeed, there are several instances where the convergent validity or accuracy for subgroups in the current study exceeds that of the test set – such as for those with dyslexia and autism for openness, autism, and autism, ADHD and dyslexia for emotional stability, and autism and dyslexia for extraversion. There are, however, also instances where the scoring algorithms did not perform as well for particular subgroups. For example, there is a negative r-squared value for the dyslexia, ADHD and autism group for openness and neurotypical and dyslexia and autism groups for achievement, indicating that the model has a worse fit compared to a horizontal line (Chicco et al., 2021). Overall, H6 is generally supported, with the algorithms generalising well to neurodivergent test-takers.

Table 28

Performance metrics for the scoring algorithm for each trait for different subsets of data.

Data Subset	N	*r*	R^2	MAE	MSE	RMSE
Openness						
Training	1283	.66**	.43	8.67	116.97	10.82
Test	550	.60**	.35	9.27	131.89	11.48
Current study	556	.53**	.26	10.41	170.18	13.05
Neurotypical	146	.55**	.23	9.39	134.28	11.59
Neurodivergent	420	.52**	.27	10.76	182.66	13.52
ADHD	102	.48**	.21	10.73	175.23	13.24
ADHD, autism	72	.46**	.18	11.08	201.24	14.19
Autism	88	.55**	.30	10.97	192.07	13.86
Dyslexia	86	.59**	.31	11.27	206.37	14.37
Dyslexia, ADHD	43	.59**	.35	9.03	120.09	10.96
Dyslexia, ADHD, autism	14	.49**	-.26	11.09	153.35	12.38
Dyslexia, autism	15	.71**	.34	9.89	159.49	12.63
Achievement						
Training	1283	.72**	.52	9.03	127.34	11.28
Test	550	.65**	.42	10.07	160.35	12.66
Current study	556	.59**	.18	9.92	154.69	12.44
Neurotypical	146	.51**	-.05	9.79	156.06	12.49
Neurodivergent	420	.61**	.22	9.96	154.22	12.42
ADHD	102	.55**	.17	9.90	157.10	12.53
ADHD, autism	72	.60**	.17	10.77	173.75	13.18
Autism	88	.71**	.45	8.09	103.82	10.19
Dyslexia	86	.59**	.04	11.07	169.28	13.01
Dyslexia, ADHD	43	.56**	.16	9.89	160.54	12.67
Dyslexia, ADHD, autism	14	.72**	.47	9.88	177.03	13.31
Dyslexia, autism	15	.34**	-.27	11.46	210.77	14.52
Extraversion						
Training	1283	.78**	.60	11.43	207.92	14.42
Test	550	.73**	.53	12.90	257.83	16.06
Current study	556	.69**	.46	13.52	278.65	16.69
Neurotypical	146	.61**	.30	13.52	274.59	16.57
Neurodivergent	420	.70**	.48	13.52	280.06	16.73
ADHD	102	.73**	.52	12.75	242.78	15.58
ADHD, autism	72	.66**	.43	13.40	286.57	16.93
Autism	88	.66**	.43	13.63	285.12	16.89
Dyslexia	86	.68**	.41	13.48	279.31	16.71
Dyslexia, ADHD	43	.64**	.37	15.87	333.81	18.27
Dyslexia, ADHD, autism	14	.57**	.28	15.60	449.38	21.20
Dyslexia, autism	15	.87**	.73	10.25	164.79	12.84

Agreeableness						
Training	1283	.70**	.48	9.97	155.67	12.48
Test	550	.62**	.38	10.59	179.35	13.39
Current study	556	.60**	.30	11.39	216.03	14.70
Neurotypical	146	.48**	.08	10.61	170.72	13.07
Neurodivergent	420	.62**	.32	11.67	231.78	15.22
ADHD	102	.55**	.27	12.14	217.44	14.75
ADHD, autism	72	.60**	.23	13.75	351.70	18.75
Autism	88	.68**	.39	11.24	214.53	14.65
Dyslexia	86	.56**	.25	10.60	199.62	14.13
Dyslexia, ADHD	43	.60**	.23	11.60	217.48	14.75
Dyslexia, ADHD, autism	14	.78**	.61	7.83	101.18	10.06
Dyslexia, autism	15	.82**	.49	10.84	202.25	14.22
Emotional stability						
Training	1283	.76**	.58	11.78	225.13	15.00
Test	550	.71**	.50	13.15	272.83	16.52
Current study	556	.71**	.49	13.55	285.83	16.91
Neurotypical	146	.52**	.01	13.38	273.26	16.53
Neurodivergent	420	.69**	.46	13.60	290.19	17.04
ADHD	102	.59**	.32	15.22	372.85	19.31
ADHD, autism	72	.72**	.46	14.16	300.21	17.33
Autism	88	.82**	.67	1.49	178.60	13.36
Dyslexia	86	.60**	.33	15.12	338.86	18.41
Dyslexia, ADHD	43	.73**	.48	12.38	226.05	15.04
Dyslexia, ADHD, autism	14	.77**	.56	13.73	270.49	16.45
Dyslexia, autism	15	.72**	.24	12.88	257.90	16.06

Note. *r* = Pearson correlation coefficient for actual and predicted scores; R^2 = proportion of variance explained; MAE = mean absolute error; MSE = mean squared error; RMSE = root mean squared error.

Subgroup differences

As an exploratory analysis to provide some first data, the presence of subgroup differences between neurotypes was investigated. To do so, scores on the questionnaire-based measure and predicted scores from the algorithms were binarised based on whether they were above or below the median score for that trait in line with the approach of New York City Local Law 144 (The New York City Council, 2021). Group differences based on neurodiversity, dyslexic neurotype, ADHD neurotype, and autistic neurotype were then examined using the following metrics:

- **Four-fifths rule** – compares the pass rates of subgroups to the group with the highest rate to calculate an impact ratio, where ratios below .80 can indicate adverse impact (Equal Employment Opportunity Commission, 1978).
- **Two standard deviations rule** (also known as the z-test) – compares the expected and observed pass rates of each group based on the proportion of data that each subgroup represents, where values > 2 indicate that there is a statistically significant discrepancy in expected and observed pass rates (D. Morgan, 2010; S. B. Morris & Lobsenz, 2000).
- **Cohen's *d*** – a measure of effect size of the difference between means, where values above .20, .50, and .80 indicate small, medium, and large effect sizes, respectively (Cohen, 1992). The current study used a threshold of +/- .20 as indicative of group differences.

Given that the three metrics can result in discrepant results (Hilliard, Kazim, et al., 2022a; S. B. Morris & Lobsenz, 2000), an agreement of two or more metrics was used as an indication of the presence of group differences.

As can be seen in **Error! Reference source not found.**, for predicted scores, subgroup differences were indicated for non-dyslexic individuals for openness, for neurodivergent and ADHD neurotype for achievement, neurodivergent, and autistic neurotype for extraversion, neurodivergent, and autistic neurotype for agreeableness, and neurodivergent, ADHD neurotype, and autistic neurotype for emotional stability. However, these group differences generally follow a similar pattern to group differences in actual scores on the questionnaire-based measure. This indicates that there could be genuine subgroup differences in personality, rather than the algorithm or image-based assessment format not working well for specific groups. A full adverse impact analysis against gender, age, and ethnicity can be seen in Appendix H.

Table 29

Adverse impact analysis for scores on the image-based assessment/questionnaire-based assessment. Exceptions to the metrics (four-fifths <.80, SD > ±2, Cohen's d > ±.20) are in bold.

Group	Group Size	*n* passing	Pass rate	Impact ratio	2SD	Cohen's d
Openness						
Neurodivergent						
Neurotypical	146	73/55	.50/.38	1.00/**.78**	.00/-**2.28**	.00/**.22**
Neurodivergent	420	210/204	.50/.49	1.00/1.00	.00/**2.28**	.00/.00
Dyslexic neurotype						
Non-dyslexic	408	190/183	.47/.45	.79/.93	-2.62/-.70	.25/.07
Dyslexic	158	93/76	.59/.48	1.00/1.00	2.62/.70	.00/.00
ADHD neurotype						
Non-ADHD	335	170/145	.51/.43	1.00/.88	.43/-1.42	.00/.12
ADHD	231	113/114	.49/.49	.96/1.00	-.43/1.42	-.04/.00
Autism neurotype						
Non-autistic	377	193/163	.51/.43	1.00/.85	.8/-1.7	.00/.15
Autistic	189	90/96	.48/.51	.93/1.00	-.8/1.7	-.07/.00
Achievement						
Neurodivergent						
Neurotypical	146	97/77	.66/.53	1.00/1.00	4.61/2.17	.00/.00
Neurodivergent	420	186/178	.44/.42	.67/.80	-4.61/-2.17	-.46/-.21
Dyslexic neurotype						
Non-dyslexic	408	203/183	.50/.45	.98/.98	-.19/-.15	.02/.01
Dyslexic	158	80/72	.51/.46	1.00/1.00	.19/.15	.00/.00
ADHD neurotype						
Non-ADHD	335	194/169	.58/.50	1.00/1.00	4.53/3.11	.00/.00
ADHD	231	89/86	.39/.37	.67/.74	-4.53/-3.11	-.39/-.27
Autism neurotype						
Non-autistic	377	198/169	.53/.45	1.00/99	1.69/-.15	.00/.01
Autistic	189	85/86	.45/.46	.86/1.00	-1.69/.15	-.15/.00
Extraversion						
Neurodivergent						
Neurotypical	146	91/85	.62/.58	1.00/1.00	3.46/3.20	.00/.00
Neurodivergent	420	192/180	.46/.43	.73/.74	-3.46/-3.20	-.34/-.31
Dyslexic neurotype						
Non-dyslexic	408	194/183	.48/.45	.84/.86	-1.87/-1.51	.18/.14
Dyslexic	158	89/82	.56/.52	1.00/1.00	1.87/1.51	.00/.00
ADHD neurotype						
Non-ADHD	335	180/166	.54/.5	1.00/1.00	2.14/1.57	.00/.00
ADHD	231	103/99	.45/.43	.83/.86	-2.14/-1.57	-.18/-.13
Autism neurotype						
Non-autistic	377	211/201	.56/.53	1.00/1.00	4.01/4.37	.00/.00

Autistic	189	72/64	.38/.34	.68/.64	-4.01/-4.37	-.36/-.4
Agreeableness						
Neurodivergent						
Neurotypical	146	93/81	.64/.55	1.00/1.00	3.84/1.93	.00/.00
Neurodivergent	420	190/194	.45/.46	.71/.83	-3.84/-1.93	-.38/-.19
Dyslexic neurotype						
Non-dyslexic	408	197/196	.48/.48	.89/.96	-1.31/-.42	.12/.04
Dyslexic	158	86/79	.54/.50	1.00/1.00	1.31/.42	.00/.00
ADHD neurotype						
Non-ADHD	335	177/158	.53/.47	1.00/93	1.62/-.82	0/.07
ADHD	231	106/117	.46/.51	.87/1.00	-1.62/.82	-.14/.00
Autism neurotype						
Non-autistic	377	211/200	.56/.53	1.00/1.00	4.01/3.00	.00/.00
Autistic	189	72/75	.38/.40	.68/.75	-4.01/-3.00	-.36/-.27
Emotional stability						
Neurodivergent						
Neurotypical	146	112/118	.77/.81	1.00/1.00	7.49/9.2	.00/.00
Neurodivergent	420	171/154	.41/.37	.53/.45	-7.49/-9.2	-.78/-1.00
Dyslexic neurotype						
Non-dyslexic	408	203/205	.50/.5	.98/1.00	-.19/1.67	.02/.00
Dyslexic	158	80/67	.51/.42	1.00/84	.19/-1.67	.00/-.16
ADHD neurotype						
Non-ADHD	335	194/196	.58/.59	1.00/1.00	4.53/5.99	.00/.00
ADHD	231	89/76	.39/.33	.67/.56	-4.53/-5.99	-.39/-.53
Autism neurotype						
Non-autistic	377	213/207	.56/.55	1.00/1.00	4.37/4.61	.00/.00
Autistic	189	70/65	.37/.34	.66/.63	-4.37/-4.61	-.04/-.42

Discussion

This study sought to investigate the effect of personality assessment format – questionnaire-based and image-based – on test-taker reactions using a sample of neurodivergent and neurotypical applications and examine how well scoring algorithms for the image-based assessment work for neurodivergent test-takers.

Study Five

Study Five investigated the effect of assessment format (questionnaire-based or image-based) on test-taking experience, both overall and on six test-taking scales: motivation,

concentration, test ease, comparative anxiety, external attribution, and fairness. It also investigated the compatibility of the assessments with neurodivergent ways of thinking through a scale developed through interviews with neurodivergent adults. Specifically, it sought to determine whether neurotypical and neurodivergent test-takers had different test-taking experiences (H3), whether the questionnaire-based and image-based assessment resulted in different test-taking experiences (H4), and whether there were differences in the compatibility of assessment format between neurotypes (H5). For both the image-based and questionnaire-based assessment, neurotypical test-takers had a more positive overall experience and had significantly higher ratings for motivation, concentration, ease, and fairness and lower ratings of external attribution and comparative anxiety on the image-based assessment compared to neurodivergent test-takers, supporting H3.1, H3.2, and H3.3. This also concurred with the perspectives of interviewees from Study Four, who suggested that pre-employment tests can present actual and psychological barriers for neurodivergent applicants. Moreover, there was a lack of significant difference in the overall test-taking experience between the two formats, meaning H4 was not supported. However, this does indicate that while the image-based format did not improve the test-taking experience relative to the questionnaire-based format, it also did not exacerbate the barriers associated with pre-employment tests.

Focusing on neurodivergent test-takers, the questionnaire-based format was rated as significantly higher in concentration and ease than the image-based format, while the image-based format was rated as significantly higher in fairness despite also being rated higher in external attribution. Additionally, the questionnaire-based format was rated as more compatible with neurodivergent ways of thinking overall. However, comparisons by group indicated that there was only a significant difference in the compatibility of the two formats for test-takers with a diagnosis of autism, as well as ADHD and autism, suggesting that other

neurotypes did not find either format to improve or worsen their experience relative to the other. This is despite the fact that dyslexic individuals rely more on visual and semantic processing abilities to compensate for phonological and verbal deficits caused by their neurotype (Bacon & Handley, 2010), meaning that preference for the image-based format among dyslexic test-takers would be expected. Likewise, individuals with ADHD can have a greater reliance on visual processing (Fassbender & Schweitzer, 2006), suggesting that other factors, may have influenced compatibility. Indeed, individuals with ADHD have greater difficulty concentrating compared to those without ADHD (A. Mueller et al., 2017). Moreover, autism is associated with greater perceptual inference (Skewes et al., 2015) or the ability to notice details, which may help with the interpretation of the visual items, but given that many of the images rely on the interpretation of social cues and autistic individuals have an atypical approach to interpreting and processing them (Ashwin et al., 2015), this could explain why the image-based format was less compatible with autistic participants. Overall, the results indicated that different neurotypes experience the formats differently, supporting H5.

Furthermore, qualitative insights from open-ended responses indicated that both formats were associated with a moral dilemma about whether to respond genuinely or in the way that participants predicted employers to desire, with being faking a significant concern for the validity of selection assessments (Melchers et al., 2020). Participants also shared concerns about questions being ambiguous and lacking context across both formats, where items measured test-taker personality in general, rather than specifically in the workplace context. However, the image-based assessment's visual and text-based format for some questions helped to provide clarity and reduce some of the ambiguity, with the format also being perceived as refreshing and fun, which could have important implications for the talent pool employers have access to (Chapman et al., 2005; Hausknecht et al., 2004). Overall,

Study Five did not provide evidence to support image-based assessments reducing differential experiences of neurodivergent test-takers compared to neurotypical test-takers nor improving the test-taking experience for neurodivergent test-takers in general but equally did not find that the image-based format presented additional barriers. As such, Study Five provides some first research that can be further built upon.

Study Six

Study Six sought to examine how well scoring algorithms for the image-based assessment that were trained on a general population perform for neurodivergent applicants in terms of convergent validity between predicted scores and actual personality scores, as measured by the questionnaire-based assessments (H6). It also sought to provide first data on subgroup differences in scores based on neurotype.

While there were some instances where the scoring algorithms had a negative r-squared value and non-significant convergent correlations when applied to neurodivergent respondents, for the most part, the performance of the models was similar to the test data and neurotypical sample. Moreover, in some cases, the performance of the models when applied to neurodivergent test-takers exceeded that of the test data or neurotypical sample. This suggests that algorithms trained on general samples can generally be applied to neurodivergent test takers, supporting H6, although efforts should be made to ensure the representativeness of training data to ensure algorithms are optimised for all groups that they may be used to score (Tay et al., 2022). Adverse impact analysis based on the four-fifths rule, Cohen's *d*, and the two standard deviations rule indicated that group differences in predicted scores for multiple personality traits. However, analysis of questionnaire-based scores showed a similar pattern of group differences, indicating that there could be genuine differences in personality between groups. Indeed, meta-analytic findings have indicated that there can be moderate differences in the personality traits of different races (Foldes et al.,

2008) and across nationalities (D. P. Schmitt et al., 2007). There can also be subgroup differences in personality traits based on neurotype (Lodi-Smith et al., 2019; Nigg et al., 2002; Tops et al., 2013), although this is something that is less well explored. Overall, this indicates that differences in personality could be genuine and not an artefact of the scoring algorithms or assessment format since both the questionnaire and image-based format had similar patterns of group differences, although this should be further investigated. Moreover, since the total score for the assessment combines all personality outcomes and cognitive ability, group differences can be balanced out at the procedure level.

Limitations and future directions

While several previous studies investigating reactions to and perceptions of various assessment formats only subjected participants to hypothetical scenarios (Kaibel et al., 2019; Köchling et al., 2022; Langer et al., 2019; Lee, 2018; Mirowska & Mesnet, 2021), in the current study, participants experienced both formats first-hand. However, participants only imagined that they had completed the assessment as part of a hiring scenario. Although qualitative insights suggested that participants took this seriously in that they were considering the responses an employer would find desirable, simply imagining that the assessment was being taken in a high-stakes context might result in stronger relationships between variables than if the assessments were taken as part of an authentic selection process (Hausknecht et al., 2004). As such, a direction for future research could be to explore applicant reactions when the assessment is applied in practice (e.g., Tilston et al., 2024).

This research could also be extended by investigating whether perceptions vary based on other characteristics such as age or gender, and whether this interacts with diagnosis. Indeed, technology self-efficacy can vary with age (Ellison et al., 2020) and older applicants find technology less useful (Hauk et al., 2018), which could drive disparities in ratings for older versus younger applicants. Moreover, there are gender differences in technology

attitudes, with males viewing technology more favourably than females (Cai et al., 2017). This may lead to gender differences in perceptions of novel assessment formats, where females have been found to react more positively to increased preparation time for video interviews (Tilston et al., 2024).

Finally, the characteristics of the image-based assessment may have impacted reported perceptions in the current study. While this study made use of an image-based assessment that was in development, allowing insights to be applied to improve the test-taking experience, the fact that the item bank had not yet been refined meant that the assessment was longer than it would have been if fully developed, which may have impacted test-taking experience, with qualitative insights flagging the assessment length. As such, once the measure is finalised and implemented, the study could be conducted again to examine whether assessment length influences perceptions. Furthermore, the lower perceived ease of the image-based format over the questionnaire-based format may have been due to the differences in the scales of the two assessments, with the questionnaire-based format using a Likert scale and many of the image pairs only having two scale points and hence being forced-choice. Given that forced-choice assessments are designed to elicit deeper processing of response options (Smyth et al., 2006), they are often seen as more difficult than single statement measures (B. Zhang et al., 2020), meaning that the perceived greater ease of the questionnaire-based assessment may be an artefact of differences in question style between the two formats. As such, future research could compare perceptions of text-based and image-based forced-choice assessments to control for differences in question style and could also be extended to investigate different assessment formats, such as game-based assessments and video interviews, to examine how perceptions differ across formats.

Conclusion

This study provides insights into the key perceptions of selection assessments of neurodivergent adults, who can experience greater anxiety about their performance on assessments compared to neurotypical test-takers and be more prone to their performance being influenced by external factors. While image-based formats are perceived as fairer in terms of opportunity to perform and are associated with greater enjoyment, amusement, and focus, there is a lack of evidence that they can completely alleviate barriers associated with pre-employment tests for neurodivergent applicants, although further research is needed with more concise assessments that balance images and text to investigate this potential. Furthermore, this study supports the use of machine learning to score novel assessment formats, where algorithms developed on a general population performed well for neurodivergent test-takers, and sometimes had improved performance compared to training or test data used during algorithm development. Nevertheless, there should be an impetus to ensure that training data represents a range of demographics and thinking styles to ensure that algorithms are optimised for various groups that they will be used to score. Future research should continue to investigate how alternative assessment formats might benefit neurodivergent job applicants to further diversity, equity, and inclusion.

Chapter 7. General discussion

General discussion

This chapter discusses the main findings of the six studies described in this thesis, as well as their contributions to the interdisciplinary field of algorithmic recruitment tools. This research aimed to explore the implications of algorithmic assessments on bias and fairness in recruitment through the perspectives of I-O psychology and computer science. First, it explored the compatibility of the two fields' definitions of bias and approaches to mitigating it before investigating how the two fields can come together in practice to create valid and unbiased selection assessments. It then explored how such tools impact the fairness of test-taking experiences of neurodivergent and neurotypical test-takers and whether these tools have the potential to close the gap and make assessments fairer. Two commercially-developed image-based assessments of personality designed to be used in selection that were undergoing validation were used as a vehicle to investigate these phenomena due to the access available to the researcher and the lack of research on image-based formats in general. Using a real-life tool over a hypothetical scenario also helped to make the described studies more realistic and increased their ecological validity.

This chapter first discusses the main findings of each of the studies and their implications. It then discusses the limitations of these studies and potential areas of future research before ending with a general conclusion.

Main findings

The six studies described in this research have two key themes:

i) The compatibility of machine learning and psychology in the context of algorithmic selection assessments in terms of assessment validity and measuring, mitigating, and sources of bias.

ii) Whether algorithmic selection assessments increase the fairness of pre-employment tests, including whether they provide neurodivergent applicants

with an overall better test-taking experience and reduce the gap between the experiences of neurotypical and neurodivergent applicants.

Studies One, Two, Three, and Six investigated the first, while Studies Four and Five investigated the second. These themes are used to anchor the discussion of the key findings in the following subsections.

Using machine learning to score psychometric assessments

Study One. The first study described the creation of an image-based personality assessment in collaboration with an industry partner for use in selection. The assessment presents respondents with image pairs mapped to the Big Five traits and asks them to indicate which trait in the pair is more like them. The creation and validation of other novel assessment formats, such as game-based assessments and asynchronous video interviews, have increasingly been described in the literature (Collmus & Landers, 2019; Ellison et al., 2020; Georgiou et al., 2019; Hickman, Bosch, et al., 2021; Hickman. et al., 2019; Hickman, Saef, et al., 2021; Landers, Armstrong, et al., 2022; Landers & Sanchez, 2022; Montefiori, 2016; Ventura & Shute, 2013). However, despite showing promise (Leutner et al., 2017; H. Zhang et al., 2017) image-based assessments of personality have been explored to a much lesser extent. As such, Study One created an initial item bank of image pairs to form the basis of an image-based assessment of personality, where images were mapped to IPIP statements (Goldberg, 1992) to support measure validity. The initial item bank was then refined such that only the best-performing images, or those that demonstrated clear differences in the personality of individuals selecting each image in the pair, were retained.

Study Two. Study Two sought to create machine learning based predictive scoring algorithms based on image choices for the measure described in Study One. This study aimed to investigate the implications of combining psychological theory with computer science based scoring approaches on the convergent and discriminant validity of the measure, as well

as any resulting bias in the scores. Specifically, it investigated H1, which predicted that through the use of machine learning based scoring algorithms, the image-based assessment would have strong convergent and discriminant validity. Convergent validity with the IPIP-NEO-120 (J. A. Johnson, 2014) was strong, ranging from .60 for agreeableness to .78 for extraversion. This exceeded the validity of previously created non-verbal personality assessments, with Paunonen et al.'s (2001) measure ranging from .45 for emotional stability to .59 for agreeableness.

Moreover, the convergent validity of the image-based assessment described in this study also exceeded the convergence between two questionnaire-based measures of personality, which ranged from .50 for agreeableness to .76 for emotional stability (Lim & Ployhart, 2006). Discriminant validity was also generally strong, with algorithms measuring the trait they were designed to, rather than other traits. As such, H1 was supported. Furthermore, adverse impact analysis indicated that the assessment and scoring algorithm combination resulted in acceptable subgroup differences in accordance with psychology approaches to adverse impact and equal opportunity laws. As such, Study Two provided support for the use of machine learning based scoring of image-based assessments, finding that the intersection of psychology and machine learning did not result in biased outcomes and resulted in a valid, accurate, and fast assessment of personality that could be used in selection.

Study Three. Building on the findings of Study Two, Study Three explored the implications of using different types of predictive algorithms and predictor combinations on the validity of the measure described in Studies One and Two, as well as which approach was more effective at preventing biased outcomes. To do this, four scoring approaches were used: Lasso regression, Ridge regression, ordinary least squares regression, and a manual, summative approach. Lasso regression has an L1 regularisation parameter that causes the

coefficients of the model to be shrunk by an equal amount, resulting in the coefficient of some predictors being reduced to zero and therefore effectively removing the predictor from the model (McNeish, 2015; Tibshirani, 1996). Ridge regression, on the other hand, uses an L2 regularisation parameter that reduces the coefficients in a way that is proportional to their size, meaning no predictors are removed from the model (McNeish, 2015). Regulation parameters are particularly useful for models trained on one set of data that will then be applied to other data as they increase the generalisability of the model by reducing overfitting (McNeish, 2015). In contrast, ordinary least squares regression does not have a regularisation parameter and is, therefore, prone to overfitting the data, thereby reducing the generalisability of the model, particularly when there is a small ratio between the number of predictors and the number of participants (McNeish, 2015; Putka et al., 2018). Finally, the summative approach did not use regression but instead manually added the number of times an image designed to measure a particular trait was selected. Study Three investigated H2, which predicted that the machine learning based approaches would result in stronger validity.

For the regression-based models, three predictor combinations were examined – using *all* images to predict each trait, using the images *intended* to measure each trait, and using the images *mapped* to each trait using a data-driven approach during the item bank refinement in Study One. This led to ten different approaches to scoring being compared: a) Lasso all, b) Lasso intended, c) Lasso mapped, d) Ridge all, e) Ridge intended, f) Ridge mapped, g) OLS all, h) OLS intended, i) OLS mapped, and j) summative. The Lasso all models are those that were used in Study Two.

For each predictor combination, convergent and discriminant validity were examined, and adverse impact analysis was carried out using the four-fifths rule, Cohen's *d*, and the two standard deviations rule. Subgroup differences, or potential adverse impact, were identified when two or more metrics were violated, where acceptable values are $> .80$, $< |.20|$, and $< |2|$,

respectively. The number of differences resulting from each scoring approach for each personality trait was then compared to the training data to examine whether they diminished what could be genuine subgroup differences in personality or whether they unjustifiably resulted in more subgroup differences.

Across predictor combinations, convergent validity was greatest for the machine learning based approaches compared to the OLS and summative approaches, and the summative approaches generally had much lower convergent validity, around .51. Moreover, the machine learning models also demonstrated greater generalisability, with less of a gap in the performance for the training and test sets compared to the remaining approaches, likely due to the regularisation parameter reducing overfitting (McNeish, 2015). This demonstrates that using data-driven machine learning based approaches can help to increase the validity and generalisability of measures, particularly those that use a forced-choice format, as with the image-based assessment used in this study. The machine learning models – the Lasso and Ridge models – performed similarly in terms of convergent validity and generalisability, but Lasso achieved this performance with fewer predictors due to the removal of features from the model (Tibshirani, 1996). As such, the Lasso-based approach could give rise to a shorter but equally valid measure that retained only the best-performing images.

An analysis of subgroup differences indicated that the Lasso and Ridge mapped and intended models had similar exceptions to scores on the IPIP-NEO-120 (the training data), while the summative approach resulted in the most exceptions. As such, the Lasso and Ridge mapped and intended models likely reflected genuine subgroup differences in personality, although this should be investigated further (SIOP, 2018), while the summative approach introduced novel and potentially unjustifiable subgroup differences. On the other hand, the OLS mapped model resulted in the fewest exceptions, but this could be at the expense of genuine subgroup differences in personality. As such, the combination of psychological

theory and machine learning resulted in a more valid assessment than psychology alone without causing bias. However, the study does highlight the need to examine different predictor combinations and machine learning approaches to optimise outcomes.

Study Six. Study Six sought to investigate whether scoring algorithms developed using data from the general population generalised well to neurodivergent test-takers. Convergent validity scores on a questionnaire-based measure of personality and the image-based scores produced by the algorithms based on the test subset of data from when the algorithms were created ranged from .60 for openness to experience to .73 for extraversion. When applied to the neurotypical test-takers in the current study, convergent validity ranged from .48 for agreeableness to .61 for extraversion, while for neurodivergent test-takers, convergent validity ranged from .52 for openness to experience to .70 for extraversion, meaning that interestingly the algorithms had stronger convergent validity for the neurodivergent test-takers than neurotypical. Although convergent validity for all groups decreased from the test set, having only a test set and training set can lead to overestimations of model performance compared to also having a validation sample (Xu & Goodacre, 2018), so this slight decrease in performance was to be expected. This analysis was also carried out for each diagnosis group individually, with convergent validity for multiple subgroups exceeding that of the test set – such as in the case of dyslexia and autism for openness, autism, and autism, ADHD and dyslexia for emotional stability, and autism and dyslexia for extraversion.

In addition, Study Six also sought to investigate subgroup differences in scores on the image-based assessment that have resulted from the format and/or scoring algorithms. Based on the four-fifths rule, Cohen's *d*, and 2SD rule, potential adverse impact was identified for non-dyslexic individuals for openness, for neurodivergent and ADHD neurotype for achievement, neurodivergent, and autistic neurotype for extraversion, neurodivergent, and

autistic neurotype for agreeableness, and neurodivergent, ADHD neurotype, and autistic neurotype for emotional stability. However, these subgroup differences reflected subgroup differences that were present in scores on the questionnaire-based assessment, suggesting that they were not the result of the image-based format or scoring algorithms and could, therefore, be genuine subgroup differences (SIOP, 2018).

Bias worked example. Finally, the bias worked example in Chapter 2 demonstrated that machine learning approaches to bias mitigation can be effectively applied to algorithmic recruitment tools to mitigate subgroup differences in the outputs of the algorithms. Indeed, the Prejudice Remover Regularizer (Kamishima et al., 2011) effectively mitigated bias against males and mixed ethnicity test-takers for the emotional stability and conscientiousness algorithms, respectively, without a considerable negative impact on model performance. Moreover, the two models had different magnitudes of subgroup differences, where the impact ratio of .77 for emotional stability was only slightly below the .80 threshold to be considered unbiased, while the .36 impact ratio for conscientiousness was considerably below the threshold. Because the in-processing Prejudice Remover Regularizer effectively mitigated the subgroup differences in both instances, this demonstrated the effectiveness of the mitigation regardless of the size of subgroup differences. On the other hand, as a pre-processing approach, Learning Fair Representations (Zemel et al., 2013) transformed the input data, changing it from binary image choices to floats. Although this approach did also rectify the subgroup differences, it did so in a way that made the input data no longer representative of image choices. Finally, although Equalized Odds (Hardt et al., 2016) is widely used in computer science, it is a post-processing approach that aims to ensure that the true and false positive rates are equal across groups, not that outcomes are comparable. As such, it did not effectively mitigate bias as measured in psychology and equal opportunity laws. Furthermore, the changing of scores based on subgroup membership is not compatible

with psychology and could violate equal opportunity laws, so is not appropriate for use with recruitment tools. Accordingly, for the current example, the in-progressing approach was most effective in terms of mitigating subgroup differences measured by metrics used by psychologists and being compatible with equal opportunity laws. The pre-processing approach also showed the same potential but was not compatible with the data format in this particular example.

In summary. Overall, studies One, Two, Three, and Six demonstrate that I-O psychology and machine learning can effectively come together to create assessments that are valid and unbiased and combining the two fields can result in more valid and less biased assessments than psychology alone can. This is particularly true in the context of novel assessment formats that use untraditional data and formats that psychology may be less optimised to take advantage of. As such, combining computer science and psychology can give rise to a better candidate experience through more innovative assessments (al-Qallawi & Raghavan, 2022; Georgiou & Nikolaou, 2020; Leutner et al., 2021; Mavridis & Tsiatsos, 2017; Smits & Charlier, 2011), such as image-based formats. Moreover, these studies have demonstrated the importance of considering the legal landscape and social implications when applying bias mitigation approaches from computer science in order to ensure compliance with equal opportunity laws while rectifying subgroup differences. Furthermore, the format of the data and what it represents must be considered to ensure that any transformations made to the data still encode the original insights represented in the data. Nevertheless, they did highlight the potential for computer science and machine learning to effectively come together to prevent and mitigate bias in algorithmic recruitment tools.

Test-taking experience of neurodivergent adults

Study Four. Since much of the research into the experiences of neurodivergent individuals focuses on children and adolescents in an educational setting (Leather & Kirwan, 2012), the experiences of neurodivergent job applicants are not well understood. This is

despite the fact that more diverse workplaces result in more collaborative teams in the workplace (Gomathy, 2023) and can lead to better business outcomes, including higher profit (Herring, 2009). Neurodiverse employees can also bring unique strengths to the workplace, such as creativity, passion, out-of-the-box thinking, and problem-solving (Beetham et al., 2017; Cope & Remington, 2022; De Beer et al., 2014; Hoogman et al., 2020; Kannangara et al., 2018; McDowall et al., 2023; Sarkis, 2014; Sauter & McPeek, 1993; Sedgwick et al., 2019; Steele et al., 2021; Weinberg & Doyle, 2017). As such, understanding the factors that might present a barrier for neurodivergent applicants is essential.

Accordingly, Study Four conducted interviews with neurodivergent adults, with a diagnosis of ADHD, dyslexia, or autism who were or had recently been employed, on their experiences with recruitment tools. Specifically, it sought to understand barriers to employment or optimal performance that pre-employment tests could pose and associated accommodations that could help to minimise these barriers. The interviews also endeavoured to gather insights on how algorithmic formats might alleviate or exacerbate these barriers or provide or remove adjustments in comparison to traditional formats.

During the interviews, interviewees shared rich insights about their experiences with recruitment tools, identifying a number of barriers that affected their performance. This included a narrow focus, both in terms of performance in a narrow time frame and a focus on narrow skills or traits that did not let them showcase all of their strengths, ambiguous language and a lack of feedback, the pressure of completing an assessment in a short period of time, and an internal battle about whether or not to disclose their condition to access adjustments due to concerns about being met with bias and stigma. Many of these barriers were present for both traditional and algorithmic formats, but a key distinction between the two was human presence, where the lack of human involvement in algorithmic assessments relative to traditional assessments was seen as a way to reduce pressure and judgement by

some and as taking away opportunities to ask questions for others. Moreover, multiple interviewees expressed that algorithmic tools did not work well and were not complex enough compared to what could be achieved given advancements in technology, including using AI to provide instant, personalised feedback. Furthermore, although issues such as stress that resulted from being primed with a test scenario and time pressures to complete the task within the allotted window can apply to all test-takers, the remaining barriers were spoken about specifically in relation to being neurodivergent, suggesting that interviewees perceived a gap in their experiences compared to neurotypical test-takers. Unsurprisingly, these barriers and concerns about having a differential experience compared to neurotypical test-takers led to a stressful experience when completing pre-employment tests.

In terms of accommodations, many of the suggestions served as ways to reduce stress. For example, interviewees suggested that giving feedback and allowing for preparation time could optimise performance, respectively, and written communication could help to reduce ambiguity. Allowing neurodivergent test-takers to have extra time if desired was also suggested as a mechanism to reduce disparities between neurodivergent and neurotypical test-takers, and widening the focus of assessments or conducting them over a longer period of time could allow neurodivergent applicants to better show their unique skills. Furthermore, interviewees advised that the format and display of assessments should be taken into consideration so as to not create sensory issues, distractions, or reduce readability, such as providing a toggle for dark mode. Finally, gamification was suggested by multiple interviewees as a way to reduce test-taking anxiety and make the assessment more pleasant. As such, while the majority of these adjustments could be implemented for both traditional and algorithmic formats, gamification is more conducive to algorithmic formats.

These findings indicated that there was potential for algorithmic formats to make the test-taking experience more pleasant for neurodivergent candidates due to the customisability

of their interface and inclusion of elements of game. They also highlighted the importance of ensuring that the validation of these tools considers neurodivergent applicants and other disabled applicants to ensure that their needs can be met. While each individual is likely to have their own needs and preferences, accessibility adjustments are likely to benefit a range of applicants, whether or not they have a disability or are neurodivergent (Evett & Brown, 2005; Leather & Kirwan, 2012). This was also supported by the interviews, where although interviews represented a variety of neurotypes that included ADHD, autism, dyslexia, anxiety disorder, dyspraxia, and dyscalculia, all of the barriers and adjustments were identified by multiple individuals with various neurotypes. As such, although candidate experiences are unique to individuals and adjustments should be evaluated on a case-by-case basis (Moody, 2015), the adjustments recommended by interviewees could serve as a first step to improve accessibility and test-taking experience for all. These insights also informed the creation of the neurodivergent compatibility scale that was used in Study Five.

Study Five. Study Five quantitatively measured i) differences in test-taking experience between neurodivergent and neurotypical test-takers, ii) differences in test-taking experience between an image-based and questionnaire-based assessment of personality, and iii) differences in the compatibility of formats with different ways of thinking. Test-taking experience was measured in terms of motivation, concentration, ease, comparative anxiety, external attribution, and perceived fairness. These subscales were also aggregated to calculate overall test-taking experience where external attribution and comparative anxiety were reversed such that all scales measured a positive experience. Compatibility of the formats with neurotypes was measured through the neurodivergent compatibility scale that was informed by the interviews in Study Four and aims to measure how the format and display of the assessment interact with different ways of thinking.

Before any statistical analysis was carried out, the neurodivergent compatibility scale was examined for internal reliability and factor structure. Originally composed of six statements that were directly informed by quotes from the interviews described in Study Four, internal reliability analysis and principal component analysis indicated that the fifth statement did not fit well with the other statements so was removed from the scale. The scale was then re-tested, resulting in strong internal reliability and an alpha value of .836. Principle component analysis subsequently indicated that the scale represented a single factor with the fifth statement removed, and confirmatory factor analysis indicated a relatively good model fit.

To investigate H3, which predicted that there would be differences in the overall test-taking experience of neurotypical and neurodivergent test-takers for both formats, using ratings of the test-taking experience on both formats from neurodivergent and neurotypical test-takers, an independent t-test was carried out for each format. These tests indicated that neurotypical test-takers reported a significantly more positive test-taking experience compared to neurodivergent test-takers for both the image- and questionnaire-based format. This supported H3.1 and H3.2, which predicted neurotypical applicants would have a more test-taking experience on the questionnaire-based and image-based assessment, respectively. To investigate H3.3, whether neurotypical test-takers had more favourable ratings on all of the experience subscales for the image-based format, a MANOVA analysis was carried out. Here, neurotypical test-takers experienced significantly greater motivation and concentration, less anxiety and external attribution, and perceived the test as fairer and easier in comparison to neurodivergent test-takers. As such, H3 was supported. Accordingly, the views shared by interviewees in Study Four that neurodivergent and neurotypical test-takers have differential experiences were supported quantitatively. This also concurred with previous research that

indicated that selection assessments can be perceived as lacking inclusivity for neurodiverse applicants and only being optimised for neurotypical applicants (Vincent & Fabri, 2022).

H4 proposed that there would be differences in the test-taking experience between the two formats. To test this, a paired samples t-test was carried out using overall experience scores for the two formats, finding that there were no significant differences. As such, H4.1 was not supported and H4.2, which proposed that there would be differences between the formats on the experience subscales, was not tested. This did not concur with findings from research into other, similar formats such as game-based assessments, which are perceived as more satisfying, immersive, better designed, and less anxiety-inducing compared to traditional measurements of the same traits (Georgiou & Nikolaou, 2020; Leutner et al., 2021; Mavridis & Tsiatsos, 2017; Smits & Charlier, 2011). However, this study provided first data on test-taking experiences of image-based assessments and was conducted with an image-based assessment that was undergoing validation and was hence longer than it would be once deployed in practice.

H5.1 proposed that the image-based format would lead to a better test-taking experience on the subscales for neurodivergent test-takers specifically compared to the questionnaire format. This is because image-based assessments largely remove the language element from the assessment and represent choices visually, which could benefit individuals with ADHD and dyslexia in particular, who can have a preference for visual processing (De Beer et al., 2014; Fassbender & Schweitzer, 2006). Moreover, given that alternative assessment formats can help to reduce anxiety (Mavridis & Tsiatsos, 2017; Smits & Charlier, 2011), this could make the test more compatible with neurodivergent test-takers' ways of thinking as they can experience greater test-taking anxiety (Lewandowski et al., 2015; Nelson et al., 2014, 2015), which can impact performance (Hembree, 1988; McCarthy & Goffin, 2005).

To test this, a repeated measures MANOVA was carried out. This indicated that concentration and ease were significantly higher for the questionnaire-based assessment, while the image-based assessment was rated as fairer but higher in external attribution. H5.2 predicted that the image-based assessment would be more compatible with neurodivergent ways of thinking than the questionnaire-based assessment. To test this, ratings on the neurodivergent compatibility scale were compared using a paired samples t-test, which indicated that the questionnaire-based assessment ($M = 3.64$) was rated as significantly more compatible with neurotype than the image-based assessment ($M = 3.46$), meaning H5.2 was not supported.

Finally, given that different neurotypes can experience different barriers and have unique needs (Moody, 2015), H5.3 predicted that there would be differences in the compatibility of formats by neurotype. To test for this, a repeated measures ANOVA with diagnosis as a between-subjects variable was carried out. Multivariate tests indicated that both format and diagnosis had significant effects, and there was also an interaction effect for the two variables. Pairwise comparisons for diagnosis indicated that test-takers with a diagnosis of dyslexia ($M = 3.89$) had significantly higher compatibility ratings than test-takers with a diagnosis of autism ($M = 3.51$) and ADHD and autism ($M = 3.23$) in general. Furthermore, pairwise comparisons for the interaction between format and diagnosis found that test-takers with a diagnosis of ADHD and autism rated the questionnaire-based format ($M = 3.37$) as significantly more compatible with their neurotype than the image-based format ($M = 3.10$). The same pattern was seen for test-takers with a diagnosis of just autism, with the questionnaire-based measure ($M = 3.72$) rated as more compatible than the image-based measure ($M = 3.31$). Other diagnosis groups did not have a significant difference in their compatibility ratings of the two formats. As such, H5.3 was supported. The fact that autistic test-takers in particular found the questionnaire-based format more compatible with

their neurotype also concurs with the atypical approach to interpreting and processing social cues that autistic individuals can have (Ashwin et al., 2015). Indeed, since the interpretation of the images relied on social cues, this could have lacked compatibility with autistic ways of thinking and interpretation.

In summary. Overall, Studies Four and Five indicated that there are a number of factors that can impact the test-taking experience. While selection assessments are not typically a particularly pleasant experience for any applicant, they are often perceived as more unpleasant by neurodivergent applicants and thus could be considered unfair due to the disparity in the experience. This is particularly true as some barriers to optimal performance may be more readily or even exclusively associated with being neurodivergent or otherwise disabled. This, therefore, highlights an important distinction between fairness and bias in psychology, where despite an assessment not being biased as defined by differential outcomes for different groups, it can still be unfair if it results in differential test-taking experiences since fairness is driven by social perceptions (SIOP, 2018). Moreover, Study Five highlights the complexity of the factors that can influence test-taking experience, where although the image-based format was seen as more difficult to complete, harder to concentrate on, and higher in external attribution, it was still rated as fairer in terms of having a chance to perform.

Further, although the findings of these studies indicate that there is *potential* for algorithmic formats to improve the test-taking experience for neurodivergent candidates, and applicants in general, this potential was not realised in the current study. However, the findings of these studies also did not indicate that image-based formats negatively impact the test-taking experience or introduce image-based assessments, so still support the potential of these assessments.

Limitations and future research

Although largely exploratory and building on the small amount of research that does exist in relation to the fairness and bias of algorithmic recruitment tools, this research is not without limitations. Furthermore, given that this research provided some first data, it opens up the potential for future research that builds on these findings and addresses the limitations of the present research.

Extrapolating from a low-stakes context

The data in the current study was collected in a low-stakes context, where test-takers were recruited through panels and compensated for their time and responses. As such, despite the fact that responses can be rejected or returned on Prolific due to poor quality, which would provide some incentive to perform, this was a relatively low-stakes context. Consequently, responses to the measure may not reflect responses that would have been given in a high-stakes context since applicants can try to inflate their scores to appear more favourably when completing measures as part of an application process (Arthur et al., 2010; Le et al., 2011). This could explain some of the subgroup differences, which are discussed below, if certain groups tried to inflate their scores more than others. Notwithstanding this, recent research has indicated that algorithms trained in a low-stakes context are generalisable to high-stakes contexts (Stevenor et al., 2024). Nevertheless, future research should endeavour to use real-life data. For example, Study Six could be replicated with real-life recruitment data from (consenting) applicants self-reporting that they are neurodivergent.

Recruiting respondents through Prolific

All of the quantitative data that supported this research were sourced from Prolific Academic. Prolific has been demonstrated to be associated with lower levels of dishonest behaviour (Peer et al., 2017) and better-quality data (Douglas et al., 2023) than other crowd sourcing platforms such as Mturk. Moreover, Prolific Academic has a number of filters and pre-screening criterion that can be used to limit participation to allow targeted data collection

and only permit responses from individuals that have previously provided high-quality responses, as judged by prior participation, where poor response quality can result in panellists being penalised (Palan & Schitter, 2018). However, despite the advantages of Prolific over other panel providers, the use of a panel provider could limit the generalisability of the findings of this research. Indeed, individual differences can play a part in how likely an individual is to sign up to an online panel provider and participate in surveys; research indicates that individuals who are more intrinsically motivated are more likely to participate in online panels (Brüggen et al., 2011). Moreover, paid survey participation has been associated with lower levels of openness to experience (Buchanan, 2018; Valentino et al., 2021) and self-selected participants have been found to have higher extraversion and lower conscientiousness than non-participants (Ljepava, 2023). As such, this may have implications for the performance of the algorithms when applied to non-panel participants and generalisability of findings regarding test-taking experience given that the participants in these studies are likely to perceive questionnaires differently to those who do not regularly participate in questionnaire research.

Subgroup differences in scores

In computer science, subgroup differences in outcomes can be seen as inherently bad and as something that should be corrected in order to have an unbiased algorithm or assessment, such as in the case of the notion of independence (Barocas & Hardt, 2017; Hardt et al., 2016). However, in psychology, subgroup differences are not necessarily a cause for concern if they represent genuine differences in abilities or traits and are not a result of the measurement tool (measurement bias) or regression line (predictive bias; SIOP, 2018). Studies Two and Six found that subgroup differences in the scores predicted by the algorithms mirrored subgroup differences in the questionnaire-based scores that were used as training data. This highlights how biases in the training data can be reflected in algorithm outputs (Tay et al., 2022). While this finding did indicate that these differences were likely

not a result of the algorithm or format, there was a lack of investigation into whether these subgroup differences in the training data were a result of genuine differences in ability or measurement bias in the questionnaire-based measure since performance data was not collected. As such, in the event that training data demonstrates subgroup differences in scores, future studies should make an effort to investigate the source of these subgroup differences, such as by examining whether there are differences in performance (SIOP. This also highlights the importance of collecting performance data when validating pre-employment tests in order to evaluate their predictive validity and provide additional support for their use in line with the Uniform Guidelines (EEOC, 1978).

Bias mitigation

The bias mitigation worked examples served as a demonstration of how computer science approaches to bias mitigation work in practice and a vehicle to examine how compatible the transformations they make are with psychology and equal opportunity laws. As such, while the in-processing approach was the most compatible for the present study, this claim cannot be generalised beyond the specific mitigation approaches used in the worked example, namely Learning Fair Representations (Zemel et al., 2013), Prejudice Remover Regularizer (Kamishima et al., 2011), and Equalised Odds (Hardt et al., 2016). Accordingly, future research could build on these findings to compare a variety of pre-processing, in-processing and post-processing bias mitigation techniques to determine which specific techniques are compatible with psychology. This could be further extended by comparing different types of data that may be collected from algorithmic tools, such as binary data, continuous data, and even language. This research would provide psychologists with a toolkit of techniques they can apply from computer science with the knowledge that they are compatible with the social context of algorithmic recruitment tools. It could also give rise to greater collaboration between the two fields in order to develop further techniques that draw

on the varying approaches to bias in computer science (Verma & Rubin, 2018), building on emerging efforts in this area (Rottman et al., 2023).

The link between bias and fairness

Although bias and fairness are distinct concepts (SIOP, 2018), biased outcomes can influence fairness perceptions (Wang et al., 2020). However, in the current research, bias and fairness were examined in isolation; Studies Two, Three, and Six examined bias while Studies Four and Five focused on fairness. Moreover, the factors driving fairness perceptions in Study Six were not explored. This could have provided insight into why the image-based format was seen as fairer than the questionnaire-based format despite being rated more difficult, harder to concentrate on, and more prone to external attribution. As such, the findings of Study Six could be built upon by examining whether the remaining experience scales predicted fairness perceptions. Follow-up interviews could also have been conducted with respondents to gain richer insights into their perceptions of the two formats and their justifications for the ratings they provided on the questionnaire (Bowen et al., 2017). This also may have helped to provide clarifications in cases where quantitative ratings and qualitative responses to the open-ended questions were not aligned.

Moreover, the link between biased outcomes and fairness could be examined in relation to bias audits. New York City Local Law 144 (The New York City Council, 2021) is the first in the world to require independent, impartial bias audits of automated employment decision tools. The law requires impact ratios to be calculated for the outputs of tools used to make employment decisions – hiring or promotions – in New York City, where the impact ratios must be calculated in a similar way to the four-fifths rule (DCWP, 2023). It also imposes transparency requirements, where a summary of the results of the audit must be made publicly available on the website of the employer or employment agency that is subject to the audit. This has already influenced the proposal of several similar laws in the US, including in New York State, New Jersey, and Pennsylvania (Hilliard et al., 2024), meaning

that bias audits and such transparency are likely to become the norm in the coming years. However, while only a small number of employers are currently required to comply with the New York City law (Hilliard et al., 2024), this transition period represents a unique opportunity to examine the real-life impact that bias audits have on fairness perceptions of algorithmic tools. Specifically, research could examine perceptions of a tool before and after it has been subject to a bias audit and the summary of results has been published. Assuming that the outcome of the audit is positive, this would provide an opportunity to investigate whether evidence of unbiased outcomes improves fairness perceptions.

This could also be examined further in a laboratory setting to examine whether the amount of information shared about the audit and the audit outcome itself influence fairness decisions. For example, fairness perceptions of an algorithmic tool could be compared across four conditions: i) no information provided about the audit, ii) participants are informed that the tool has been audited by a third-party with no outcome specified, iii) participants are informed that the tool has been audited by a third-party with no evidence of bias found, and iv) participants are informed that the tool has been audited by a third-party and evidence of bias was identified. This would provide insights into the direct influence of bias on fairness perceptions and the optimal amount of information that should be shared about adverse impact testing to avoid additional information unintentionally resulting in more concerns (Langer et al., 2021).

Moreover, although auditing is well established in the financial services sector, bias audits of algorithmic tools are a new practice and could lead to concerns about audits being conducted due to suspicions of wrong-doing or tools being biased (Landers & Behrend, 2022), rather than them being conducted as a standard practice. As such, perceptions of bias audits and the influence of bias audits on fairness perceptions could be studied through longitudinal research as bias audits become the norm.

Focus groups to identify specific adjustments

Sticking with the theme of fairness, the results of Studies Four and Five indicated that there was potential for image-based assessments to increase the fairness of selection assessments for neurodivergent candidates, but this potential was not realised. While these studies provided first data, they did not provide particularly actionable insights into how an image-based format may be elevated by certain features in order to increase accessibility. As such, future research could conduct focus groups with neurodivergent individuals to gather rich insights into how specific tools could be adjusted to support different needs, including features that may be toggled on and off. These interviews could focus on addressing the principles of universal design, for example, to ensure that assessments are optimised for equitable use, flexibility in use, a simple and intuitive design, perceptible information, tolerance for error, low physical effort, and approach and use (Doyle, 2023; The Center for Universal Design, 1997). By having these features directly informed by insights from neurodivergent test-takers, they would likely be more effective than accommodations that non-disabled test designers may assume that disabled applicants need.

Neurodivergent compatibility scale

The neurodivergent compatibility scale was specifically created for the purposes of Study Five because there was a lack of an existing scale that considered how assessment formats and displays might interact with different neurotypes. As such, the scale was validated and then used to quantify perceptions with a single dataset. However, given that the scale has been partially validated in terms of internal reliability and factor structure and that there is a clear gap for such a scale, this validation could be built upon through a dedicated validation study. Specifically, split-half reliability and test-retest reliability should be evaluated (Fenn et al., 2020) through additional panels. These panels could also include different neurotypes beyond those examined (ADHD, dyslexia, and autism) to determine whether the scale can be applied to neurodiversity research in general. Subject matter experts

who specialise in neurodiversity could also be consulted to ensure the scale's content validity and construct validity (Fenn et al., 2020). As a consequence of this more robust validation, the scale could be made available to researchers and pave the way for more research into the compatibility of different assessment formats for neurodiverse applicants. Given the lack of research in this area, a scale like this could prove valuable and help to increase diversity in applicant pools and ultimately in the workplace.

Investigating a wider range of neurotypes

Finally, diagnoses of ADHD, dyslexia, and autism were specifically chosen for the current research due to the benefits that image-based formats could specifically have for individuals with ADHD and dyslexia and the potential issues with interpreting social cues that could come for test-takers with autism. However, a limitation of this deliberate limiting of the scope of the research also means that the findings cannot be generalised beyond these diagnosis groups as each of the conditions under the umbrella term of neurodiversity has unique symptoms and associated barriers (British Psychological Society, 2021). As such, future research could take a similar approach as Study Five, taking into consideration the recommended avenues for future research discussed above, but with other neurotypes. This would provide an opportunity to ensure that recruitment tools can accommodate as many different needs as possible.

Implications

This research provided first data on a number of applied considerations in relation to the use of algorithms and alternative assessment formats in recruitment, particularly with respect to image-based assessments, which have rarely been reported on in the literature. This is likely because image-based formats are not used at the same scale as other formats such as game-based assessments and asynchronous video interviews, as well as the fact that information pertaining to assessment performance is often commercial intellectual property and, therefore, not publicly shared.

In particular, this research supported the use of machine learning scored image-based assessments in recruitment in terms of their validity and lack of bias, where a machine learning based approach outperformed typical psychology approaches and could allow candidate experience to be further enhanced through short but highly accurate assessments, for example. It also found that some of the existing machine learning approaches to bias mitigation are compatible with algorithmic recruitment tools and can effectively reduce subgroup differences in scores, providing a starting point for the development of additional compatible tools and signposting the approaches that psychologists can use in the interim, providing that the subgroup differences do not reflect genuine group differences in performance.

This research also highlighted the need to ensure that selection procedures are as inclusive to as many different needs as possible and provided a starting point for practitioners to do so through the principles of universal design, for example. Moreover, first data was provided on the potential of image-based formats to level the playing field, providing practitioners a starting point to make informed decisions about how different assessment formats might impact the fairness of test-taking experiences and affect the accessibility of the procedure for different groups.

General conclusion

The overall aim of this research was to investigate the fairness and bias associated with algorithmic recruitment tools through the lens of image-based assessments, specifically the implications of combining psychology and machine learning on test bias and experiences of fairness. This thesis provided first data comparing how psychology and machine learning approaches to scoring impact test validity and subgroup differences and examined the compatibility of machine learning based mitigation approaches with psychology and equal opportunity laws, paving the way for the development of metrics that fit these specifications

by interdisciplinary teams. Moreover, this research found that algorithms trained on the general population can generalise to neurodivergent test-takers, providing some reassurance that well-trained algorithmic tools are not only optimised to evaluate neurotypical test-takers.

In regard to fairness, this research found that there are disparities in the test-taking experience of neurotypical and neurodivergent applicants, where neurotypical have a more positive experience. This research also demonstrated that there is potential for image-based assessments to help close this gap, although this impact was not realised, and additional research is required to further investigate how pre-employment tests can cater for a variety of different needs. Furthermore, this research created a partially validated measure of the compatibility of assessment formats with neurodivergent ways of thinking that could be validated in a dedicated study, providing a mechanism to better explore the suitability of different assessment formats for different groups. Overall, this thesis has laid the foundations to further explore how psychology and machine learning can come together to improve the test-taking experience and create assessments that are unbiased and fair for all.

Chapter 8. References

Abreu, S. (2018). Navigating choppy waters: Reasonable accommodations in standardized testing and the workplace for individuals with ADHD. *Quinnipiac Health Law Journal*, *22*. https://heinonline.org/HOL/Page?handle=hein.journals/qhlj22&id=5

Adamou, M., Arif, M., Asherson, P., Aw, T. C., Bolea, B., Coghill, D., Gudjónsson, G., Halmøy, A., Hodgkins, P., Müller, U., Pitts, M., Trakoli, A., Williams, N., & Young, S. (2013). Occupational issues of adults with ADHD. In *BMC Psychiatry* (Vol. 13, Issue 1, pp. 1–7). BioMed Central. https://doi.org/10.1186/1471-244X-13-59

Aguinis, H., Mazurkiewicz, M. D., & Heggestad, E. D. (2009). Using web-based frame-of-reference training to decrease biases in personality-based job analysis: An experimental field study. *Personnel Psychology*, *62*(2), 405–438. https://doi.org/10.1111/j.1744-6570.2009.01144.x

Aiken, J. R., & Hanges, P. J. (2015). Teach an I-O io fish: Integrating data science into I-O graduate education. *Industrial and Organizational Psychology*, *8*(4), 539–544. https://doi.org/10.1017/IOP.2015.80

Aizenberg, E., Dennis, M. J., & van den Hoven, J. (2023). Examining the assumptions of AI hiring assessments and their impact on job seekers' autonomy over self-representation. *AI and Society*, *1*, 3. https://doi.org/10.1007/s00146-023-01783-1

Albert, E. T. (2019). AI in talent acquisition: A review of AI-applications used in recruitment and selection. *Strategic HR Review*, *18*(5), 215–221. https://doi.org/10.1108/SHR-04-2019-0024

Alexander-Passe, N. (2015). The dyslexia experience: Difference, disclosure, labelling, discrimination and stigma. *Asia Pacific Journal of Developmental Differences*, *2*(2), 202–233. https://doi.org/10.3850/S2345734115000290

Allison, C., Auyeung, B., & Baron-Cohen, S. (2012). Toward brief "red flags" for autism screening: The short Autism Spectrum Quotient and the short Quantitative Checklist in 1,000 cases and 3,000 controls. *Journal of the American Academy of Child and Adolescent Psychiatry*, *51*(2), 202-212.e7. https://doi.org/10.1016/j.jaac.2011.11.003

al-Qallawi, S., & Raghavan, M. (2022). A review of online reactions to game-based assessment mobile applications. *International Journal of Selection and Assessment*, *30*(1), 14–26. https://doi.org/10.1111/ijsa.12346

Ameri, M., Schur, L., Adya, M., Bentley, F. S., McKay, P., & Kruse, D. (2018). The disability employment puzzle: A field experiment on employer hiring behavior. *ILR Review*, *71*(2), 329–364. https://doi.org/10.1177/0019793917717474

American Educational Research Association American Psychological Association National Council on Measurement in Education. (2014). *Standards for educational and psychological testing*. American Psychological Association (APA).

American Psychiatric Association. (2013). *Diagnostic and Statistical Manual of Mental Disorders*. American Psychiatric Association. https://doi.org/10.1176/appi.books.9780890425596

American Psychological Association. (2017). *Ethical principles of psychologists and code of conduct*.

Americans with Disabilities Act (1990). https://www.govinfo.gov/content/pkg/STATUTE-104/pdf/STATUTE-104-Pg327.pdf

Armstrong, M. B., Ferrell, J. Z., Collmus, A. B., & Landers, R. N. (2016). Correcting misconceptions about gamification of assessment: More than SJTs and badges. *Industrial and Organizational Psychology*, *9*(3), 671–677. https://doi.org/10.1017/iop.2016.69

Armstrong, M. B., Landers, R. N., & Collmus, A. B. (2016). Gamifying Recruitment, Selection, Training, and Performance Management. In H. Gangadharbatla & D. Z. Davis (Eds.), *Emerging Research and Trends in Gamification* (pp. 140–165). Information Science Reference. https://doi.org/10.4018/978-1-4666-8651-9.ch007

Arnold, B., Easteal, P., Easteal, S., & Rice, S. (2010). It just doesn't add up: ADHD/ADD, the workplace and discrimination. *Melbourne University Law Review*, *34*(2), 359–391. https://heinonline.org/HOL/Page?handle=hein.journals/mulr34&div=18&g_sent=1&casa_token=SygOez8CgJEAAAAA:T8StgqTpdnHlvj-7boH-

2qLVHmePp2F0sIQ0PChrhVgmEePR_yC03ZpPuI3RpsKrE3bQHWVLrTo&collection=journals

Arthur, W., Glaze, R. M., Villado, A. J., & Taylor, J. E. (2010). The magnitude and extent of cheating and response distortion effects on unproctored internet-based tests of cognitive ability and personality. *International Journal of Selection and Assessment*, *18*(1), 1–16. https://doi.org/10.1111/j.1468-2389.2010.00476.x

Ashwin, C., Hietanen, J. K., & Baron-Cohen, S. (2015). Atypical integration of social cues for orienting to gaze direction in adults with autism. *Molecular Autism*, *6*(1), 5. https://doi.org/10.1186/2040-2392-6-5

Atkins, S. M., Sprenger, A. M., Colflesh, G. J. H., Briner, T. L., Buchanan, J. B., Chavis, S. E., Chen, S., Iannuzzi, G. L., Kashtelyan, V., Dowling, E., Harbison, J. I., Bolger, D. J., Bunting, M. F., & Dougherty, M. R. (2014). Measuring working memory is all fun and games. *Experimental Psychology*, *61*(6), 417–438. https://doi.org/10.1027/1618-3169/a000262

Auer, E. M., Mersy, G., Marin, S., Blaik, J., & Landers, R. N. (2022). Using machine learning to model trace behavioral data from a game-based assessment. *International Journal of Selection and Assessment*, *30*(1), 82–102. https://doi.org/10.1111/IJSA.12363

Bachrach, Y., Kosinski, M., Graepel, T., Kohli, P., & Stillwell, D. (2012). Personality and patterns of Facebook usage. *Proceedings of the 4th Annual ACM Web Science Conference*, 24–32. https://doi.org/10.1145/2380718.2380722

Bacon, A. M., & Handley, S. J. (2010). Dyslexia and reasoning: The importance of visual processes. *British Journal of Psychology*, *101*(3), 433–452. https://doi.org/10.1348/000712609X467314

Bajorek, J. P. (2019). *Voice recognition still has significant race and gender biases*. https://hbr.org/2019/05/voice-recognition-still-has-significant-race-and-gender-biases

Baldwin, S., Costley, D., & Warren, A. (2014). Employment activities and experiences of adults with high-functioning autism and asperger's disorder. *Journal of Autism and Developmental Disorders*, *44*(10), 2440–2449. https://doi.org/10.1007/s10803-014-2112-z

Baltrunas, D., Elmokashfi, A., & Kvalbein, A. (2014). Measuring the reliability of mobile broadband networks. *Proceedings of the 2014 Conference on Internet Measurement Conference*, 45–58. https://doi.org/10.1145/2663716.2663725

Barocas, S., & Hardt, M. (2017). *Fairness in machine learning - NIPS 2017 Tutorial*. Neural Information Processing Systems. https://fairmlbook.org/tutorial1.html

Barocas, S., Hardt, M., & Narayanan, A. (2023). *Fairness and machine learning: Limitations and opportunities*. MIT Press. https://fairmlbook.org/pdf/fairmlbook.pdf

Barrick, M. R., & Mount, M. K. (1991). The Big Five personality dimensions and job performance: A meta-analysis. *Personnel Psychology*, *44*(1), 1–26. https://doi.org/10.1111/j.1744-6570.1991.tb00688.x

Bartlett, M. S. (1950). Tests of significance in factor analysis. *British Journal of Statistical Psychology*, *3*(2), 77–85. https://doi.org/10.1111/j.2044-8317.1950.tb00285.x

Basch, J., & Melchers, K. (2019). Fair and flexible?! Explanations can improve applicant reactions toward asynchronous video interviews. *Personnel Assessment and Decisions*, *5*(3). https://doi.org/10.25035/pad.2019.03.002

Bauer, T. N., Truxillo, D. M., Sanchez, R. J., Craig, J. M., Ferrara, P., & Campion, M. A. (2001). Applicant reactions to selection: Development of the Selection Procedural Justice Scale (SPJS). *Personnel Psychology*, *54*(2), 387–419. https://doi.org/10.1111/j.1744-6570.2001.tb00097.x

Baxter, A. J., Brugha, T. S., Erskine, H. E., Scheurer, R. W., Vos, T., & Scott, J. G. (2015). The epidemiology and global burden of autism spectrum disorders. *Psychological Medicine*, *45*(3), 601–613. https://doi.org/10.1017/S003329171400172X

Bayerl, M., Dielentheis, T. F., Vucurevic, G., Gesierich, T., Vogel, F., Fehr, C., Stoeter, P., Huss, M., & Konrad, A. (2010). Disturbed brain activation during a working memory task in drug-naive adult patients with ADHD. *NeuroReport*, *21*(6), 442–446. https://doi.org/10.1097/WNR.0B013E328338B9BE

Beatrice, A. (2021). *LinkedIn Coughed Out AI Bias! Is AI in Recruitment Reliable?* https://industrywired.com/linkedin-coughed-out-ai-bias-is-ai-in-recruitment-reliable/

Beetham, J., Okhai, L., Beetham, J., & Okhai, L. (2017). Workplace Dyslexia & Specific Learning Difficulties—Productivity, Engagement and Well-Being. *Open Journal of Social Sciences*, *5*(6), 56–78. https://doi.org/10.4236/JSS.2017.56007

Bernick, M. (2021). *The state of autism employment in 2021*. https://www.forbes.com/sites/michaelbernick/2021/01/12/the-state-of-autism-employment-in-2021/

Berry, C. M., & Zhao, P. (2015). Addressing criticisms of existing predictive bias research: Cognitive ability test scores still overpredict African Americans' job performance. *Journal of Applied Psychology*, *100*(1), 162–179. https://doi.org/10.1037/a0037615

Bertrand, M., & Mullainathan, S. (2004). Are Emily and Greg More Employable Than Lakisha and Jamal? A Field Experiment on Labor Market Discrimination. *American Economic Review*, *94*(4), 991–1013. https://doi.org/10.1257/0002828042002561

Bhatta, A., Albiero, V., Bowyer, K. W., & King, M. C. (2023). The Gender Gap in Face Recognition Accuracy Is a Hairy Problem. *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV) Workshops*, 303–312. https://openaccess.thecvf.com/content/WACV2023W/DVPBA/html/Bhatta_The_Gender_Gap_in_Face_Recognition_Accuracy_Is_a_Hairy_WACVW_2023_paper.html

Biel, J.-I., & Gatica-Perez, D. (2013). The YouTube lens: Crowdsourced personality impressions and audiovisual analysis of vlogs. *IEEE Transactions on Multimedia*, *15*(1), 41–55. https://doi.org/10.1109/TMM.2012.2225032

Biel, J.-I., Teijeiro-Mosquera, L., & Gatica-Perez, D. (2012). FaceTube: Predicting personality from facial expressions of emotion in online conversational video. *Proceedings of the ACM International Conference on Multimodal Interaction*, 53–56. https://doi.org/10.1145/2388676.2388689

Bigman, Y. E., Wilson, D., Arnestad, M. N., Waytz, A., & Gray, K. (2022). Algorithmic discrimination causes less moral outrage than human discrimination. *Journal of Experimental Psychology: General*. https://doi.org/10.1037/xge0001250

Black, R. D., Weinberg, L. A., & Brodwin, M. G. (2015). Universal design for learning and instruction: Perspectives of students with disabilities in higher education. *Exceptionality Education International*, *25*(2), 1–26. https://doi.org/10.5206/eei.v25i2.7723

Blodgett, S. L., & O'Connor, B. (2017). *Racial disparity in natural language processing: A case study of social media African-American English*.

Bolukbasi, T., Chang, K. W., Zou, J., Saligrama, V., & Kalai, A. (2016). Man is to computer programmer as woman is to homemaker? Debiasing word embeddings. *Advances in Neural Information Processing Systems*, 4356–4364. https://proceedings.neurips.cc/paper/2016/file/a486cd07e4ac3d270571622f4f316ec5-Paper.pdf

Bonaccio, S., Connelly, C. E., Gellatly, I. R., Jetha, A., & Martin Ginis, K. A. (2020). The Participation of People with Disabilities in the Workplace Across the Employment Cycle: Employer Concerns and Research Evidence. *Journal of Business and Psychology*, *35*(2), 135–158. https://doi.org/10.1007/s10869-018-9602-5

Bowen, P. W., Rose, R., & Pilkington, A. (2017). Mixed methods-theory and practice. Sequential, explanatory approach. *International Journal of Quantitative and Qualitative Research Methods*, *5*(2), 10–27. http://nectar.northampton.ac.uk/9608/

Bozgeyikli, L., Bozgeyikli, E., Raij, A., Alqasemi, R., Katkoori, S., & Dubey, R. (2017). Vocational rehabilitation of individuals with autism spectrum disorder with virtual reality. *ACM Transactions on Accessible Computing*, *10*(2). https://doi.org/10.1145/3046786

Braun, V., & Clarke, V. (2006). Using thematic analysis in psychology. *Qualitative Research in Psychology*, *3*(2), 77–101. https://doi.org/10.1191/1478088706qp063oa

Braun, V., & Clarke, V. (2019). *To saturate or not to saturate? Questioning data saturation as a useful concept for thematic analysis and sample-size rationales*. https://doi.org/10.1080/2159676X.2019.1704846

Braun, V., & Clarke, V. (2022). Conceptual and design thinking for thematic analysis. *Qualitative Psychology*, *9*(1), 3–26. https://doi.org/10.1037/qup0000196

British Psychological Society. (2006). *Using online assessment tools for recruitment*. https://ptc.bps.org.uk/sites/ptc.bps.org.uk/files/guidance_documents/using_online_assessment_tools_for_recruitment.pdf

British Psychological Society. (2021). *Working with autism Best practice guidelines for psychologists*. https://explore.bps.org.uk/binary/bpsworks/9fabc49fd3a5d277/8fa6f84a95698866f37795234835fc3ccb672854ae62f1ca1e1d71bd861a5154/bpsrep_rep156.pdf

Brown, A., & Maydeu-Olivares, A. (2011). Item response modeling of forced-choice questionnaires. *Educational and Psychological Measurement*, *71*(3), 460–502. https://doi.org/10.1177/0013164410375112

Brown, A., & Maydeu-Olivares, A. (2013). How IRT can solve problems of ipsative data in forced-choice questionnaires. *Psychological Methods*, *18*(1), 36–52. https://doi.org/10.1037/a0030641

Brüggen, E., Wetzels, M., De Ruyter, K., & Schillewaert, N. (2011). Individual Differences in Motivation to Participate in Online Panels. *International Journal of Market Research*, *53*(3), 369–390. https://doi.org/10.2501/IJMR-53-3-369-390

Buolamwini, J., & Gebru, T. (2018). Gender shades: Intersectional accuracy disparities in commercial gender classification. *Proceedings of Machine Learning Research*, *81*, 1–15. http://proceedings.mlr.press/v81/buolamwini18a/buolamwini18a.pdf

Burke, S. L., Bresnahan, T., Li, T., Epnere, K., Rizzo, A., Partin, M., Ahlness, R. M., & Trimmer, M. (2018). Using Virtual Interactive Training Agents (ViTA) with adults with autism and other developmental disabilities. *Journal of Autism and Developmental Disorders*, *48*(3), 905–912. https://doi.org/10.1007/s10803-017-3374-z

Byrge, L., Dubois, J., Tyszka, J. M., Adolphs, R., & Kennedy, D. P. (2015). Idiosyncratic brain activation patterns are associated with poor social comprehension in autism. *Journal of Neuroscience*, *35*(14), 5837–5850. https://doi.org/10.1523/JNEUROSCI.5182-14.2015

Byrne, B. M., & van de Vijver, F. J. R. (2010). Testing for measurement and structural equivalence in large-scale cross-cultural studies: Addressing the issue of

nonequivalence. *International Journal of Testing*, *10*(2), 107–132. https://doi.org/10.1080/15305051003637306

Cai, Z., Fan, X., & Du, J. (2017). Gender and attitudes toward technology use: A meta-analysis. *Computers & Education*, *105*, 1–13. https://doi.org/10.1016/J.COMPEDU.2016.11.003

Calders, T., Kamiran, F., & Pechenizkiy, M. (2009). Building classifiers with independency constraints. *ICDM Workshops 2009 - IEEE International Conference on Data Mining*, 13–18. https://doi.org/10.1109/ICDMW.2009.83

Camden, W. L., Allen, K. S., Specht, M. L., Bennett, M. W., Badr, K. H., Mottram, C. L., & Gutierrez, S. L. (2024). Cognitive ability: A promising option for assessing neurodiverse talent. *Consulting Psychology Journal*, *76*(1), 20–41. https://doi.org/10.1037/cpb0000271

Campbell, D. T., & Fiske, D. W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin*, *56*(2), 81–105. https://doi.org/10.1037/H0046016

Campion, M. A., Outtz, J. L., Zedeck, S., Schmidt, F. L., Kehoe, J. F., Murphy, K. R., & Guion, R. M. (2001). The controversy over score banding in personnel selection: Answers to 10 key questions. *Personnel Psychology*, *54*(1), 149–185. https://doi.org/10.1111/j.1744-6570.2001.tb00090.x

Canagasuriam, D., & Lukacik, E. R. (2024). ChatGPT, can you take my job interview? Examining artificial intelligence cheating in the asynchronous video interview. *International Journal of Selection and Assessment*. https://doi.org/10.1111/IJSA.12491

Cascade, E., Kalali, A. H., & Wigal, S. B. (2010). Real-world data on attention deficit hyperactivity disorder medication side effects. *Psychiatry (Edgemont)*, *7*(4), 13–15. http://www.ncbi.nlm.nih.gov/pubmed/20508803

Cascio, W. F., & Aguinis, H. (2001). The federal uniform guidelines on employee selection procedures (1978): An update on selected issues. *Review of Public Personnel Administration*, *21*(3), 200–218. https://doi.org/10.1177/0734371X0102100303

Cascio, W. F., Alexander, R. A., & Barrett, G. V. (1988). Setting cutoff scores: Legal, psychometric, and professional issues and guidelines. *Personnel Psychology*, *41*(1), 1–24. https://doi.org/10.1111/j.1744-6570.1988.tb00629.x

Cascio, W. F., Outtz, J., Zedeck, S., & Goldstein, I. L. (1995). Statistical implications of six methods of test score use in personnel selection. *Human Performance*, *8*(3), 133–164. https://doi.org/10.1207/s15327043hup0803_2

Central Digital & Data Office. (2017). *Simone: dyslexic user*. https://www.gov.uk/government/publications/understanding-disabilities-and-impairments-user-profiles/simone-dyslexic-user

Chai, T., & Draxler, R. R. (2014). Root mean square error (RMSE) or mean absolute error (MAE)? *Geosci. Model Dev. Discuss*, *7*, 1525–1534. https://doi.org/10.5194/gmdd-7-1525-2014

Chambers, E. G., Foulon, M., Handfield-Jones, H., Hankin, S. M., & Michaels, E. G. (1998). The war for talent. *The McKinsey Quarterly*, *3*, 44–57. https://doi.org/10.4018/jskd.2010070103

Chamorro-Premuzic, T. (2017). *The talent delusion: Why data, not intuition, is the key to unlocking human potential*. Piatkus.

Chamorro-Premuzic, T., Akhtar, R., Winsborough, D., & Sherman, R. A. (2017). The datafication of talent: How technology is advancing the science of human potential at work. *Current Opinion in Behavioral Sciences*, *18*, 13–16. https://doi.org/10.1016/j.cobeha.2017.04.007

Chang, L., Connelly, B. S., & Geeza, A. A. (2012). Separating method factors and higher order traits of the Big Five: A meta-analytic multitrait–multimethod approach. *Journal of Personality and Social Psychology*, *102*(2), 408–426. https://doi.org/10.1037/a0025559

Chapman, D. S., Uggerslev, K. L., Carroll, S. A., Piasentin, K. A., & Jones, D. A. (2005). Applicant attraction to organizations and job choice: A meta-analytic review of the correlates of recruiting outcomes. *Journal of Applied Psychology*, *90*(5), 928–944. https://doi.org/10.1037/0021-9010.90.5.928

Chen, E. E., & Wojcik, S. P. (2016). A practical guide to Big Data research in psychology. *Psychological Methods*, *21*(4), 458–474. https://doi.org/10.1037/met0000111

Cheng, M. M., & Hackett, R. D. (2021). A critical review of algorithms in HRM: Definition, theory, and practice. *Human Resource Management Review*, *31*(1). https://doi.org/10.1016/j.hrmr.2019.100698

Chicco, D., Warrens, M. J., & Jurman, G. (2021). The coefficient of determination R-squared is more informative than SMAPE, MAE, MAPE, MSE and RMSE in regression analysis evaluation. *PeerJ Computer Science*, *7*, 1–24. https://doi.org/10.7717/PEERJ-CS.623/SUPP-1

Civil Rights Act of 1991 (1991).

Cohen, J. (1992). A power primer. *Psychological Bulletin*, *112*(1), 155–159. https://doi.org/10.1037/0033-2909.112.1.155

Collins, M. W., & Morris, S. B. (2008). Testing for adverse impact when sample size is small. *Journal of Applied Psychology*, *93*(2), 463–471. https://doi.org/10.1037/0021-9010.93.2.463

Collmus, A. B., & Landers, R. N. (2019). Game-framing to improve applicant perceptions of cognitive assessments. *Journal of Personnel Psychology*, *18*(3), 157–162. https://doi.org/10.1027/1866-5888/a000227

Cope, R., & Remington, A. (2022). The Strengths and Abilities of Autistic People in the Workplace. *Autism in Adulthood*, *4*(1), 22–31. https://doi.org/10.1089/aut.2021.0037

Cornell, T. (2023). *How to detect shared scripts when interviewing*. https://www.hirevue.com/blog/hiring/how-to-detect-shared-scripts-when-interviewing-candidates

Costa, P. T., & McCrae, R. R. (1992). *Revised NEO Personality Inventory and NEO Five Factor Inventory Professional Manual*. Psychological Assessment Resources.

Costa, P. T., & McCrae, R. R. (2008). The Revised NEO Personality Inventory (NEO-PI-R). In G. J. Boyle, G. Matthews, & D. H. Saklofske (Eds.), *The SAGE Handbook of*

Personality Theory and Assessment: Volume 2 — Personality Measurement and Testing (pp. 179–198). SAGE Publications Inc. https://doi.org/10.4135/9781849200479.n9

Courey, S. J., Tappe, P., Siker, J., & LePage, P. (2013). Improved lesson planning with Universal Design for Learning (UDL). *Teacher Education and Special Education: The Journal of the Teacher Education Division of the Council for Exceptional Children*, *36*(1), 7–27. https://doi.org/10.1177/0888406412446178

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika*, *16*(3), 297–334. https://doi.org/10.1007/BF02310555

Cui, Z., & Gong, G. (2018). The effect of machine learning regression algorithms and sample size on individualized behavioral prediction with functional connectivity features. *NeuroImage*, *178*, 622–637. https://doi.org/10.1016/j.neuroimage.2018.06.001

Cummins, C., Pellicano, E., & Crane, L. (2020). Autistic adults' views of their communication skills and needs. *International Journal of Language & Communication Disorders*, *55*(5), 678–689. https://doi.org/10.1111/1460-6984.12552

Dastin, J. (2018). *Amazon scraps secret AI recruiting tool that showed bias against women*. https://www.reuters.com/article/us-amazon-com-jobs-automation-insight-idUSKCN1MK08G

DCWP. (2023). *Notice of Adoption of Final Rules*. https://rules.cityofnewyork.us/wp-content/uploads/2023/04/DCWP-NOA-for-Use-of-Automated-Employment-Decisionmaking-Tools-2.pdf

De Beer, J., Engels, J., Heerkens, Y., & Van Der Klink, J. (2014). Factors influencing work participation of adults with developmental dyslexia: A systematic review. *BMC Public Health*, *14*(1), 1–22. https://doi.org/10.1186/1471-2458-14-77

De Cooman, R., De Gieter, S., Pepermans, R., Jegers, M., & Van Acker, F. (2009). Development and validation of the work effort scale. *European Journal of Psychological Assessment*, *25*(4), 266–273. https://doi.org/10.1027/1015-5759.25.4.266

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

De Corte, W., Sackett, P. R., & Lievens, F. (2011). Designing pareto-optimal selection systems: Formalizing the decisions required for selection system development. *Journal of Applied Psychology*, *96*(5), 907–926. https://doi.org/10.1037/a0023298

de Montjoye, Y.-A., Quoidbach, J., Robic, F., & Pentland, A. (2013). Predicting personality using novel mobile phone-based metrics. *International Conference on Social Computing, Behavioral-Cultural Modeling, and Prediction*, 48–55. https://doi.org/10.1007/978-3-642-37210-0_6

Delgado, P., Vargas, C., Ackerman, R., & Salmerón, L. (2018). Don't throw away your printed books: A meta-analysis on the effects of reading media on reading comprehension. *Educational Research Review*, *25*, 23–38. https://doi.org/10.1016/J.EDUREV.2018.09.003

Di Sarno, M., Zimmermann, J., Madeddu, F., Casini, E., & Di Pierro, R. (2020). Shame behind the corner? A daily diary investigation of pathological narcissism. *Journal of Research in Personality*, *85*, 103924. https://doi.org/10.1016/J.JRP.2020.103924

Dibbets, P., Evers, E. A. T., Hurks, P. P. M., & Jolles. (2010). *Differential brain activation patterns in adult attention-deficit hyperactivity disorder (ADHD) associated with task switching Citation for published version (APA)*. https://doi.org/10.1037/a0018997

Digman, J. M. (1990). Personality structure: emergence of the five-factor model. *Annual Review of Psychology*, *41*(1), 417–440. https://doi.org/10.1146/annurev.ps.41.020190.002221

Douglas, B. D., Ewell, P. J., & Brauer, M. (2023). Data quality in online human-subjects research: Comparisons between MTurk, Prolific, CloudResearch, Qualtrics, and SONA. *PLOS ONE*, *18*(3), e0279720. https://doi.org/10.1371/JOURNAL.PONE.0279720

Downes-Le Guin, T., Baker, R., Mechling, J., & Ruyle, E. (2012). Myths and realities of respondent engagement in online surveys. *International Journal of Market Research*, *54*(5), 613–633. https://doi.org/10.2501/ijmr-54-5-613-633

Doyle, N. (2020). Neurodiversity at work: a biopsychosocial model and the impact on working adults. *British Medical Bulletin*, *135*, 108–125. https://doi.org/10.1093/bmb/ldaa021

Doyle, N. (2023). Universal design in psychometric testing. *Assessment and Development Matters*, *15*(2), 4–9. https://doi.org/10.53841/BPSADM.2023.15.2.4

Doyle, N., & McDowall, A. (2019). Context matters: A review to formulate a conceptual framework for coaching as a disability accommodation. *PLOS ONE*, *14*(8), e0199408. https://doi.org/10.1371/journal.pone.0199408

Doyle, N., & McDowall, A. (2022). Diamond in the rough? An "empty review" of research into "neurodiversity" and a road map for developing the inclusion agenda. *Equality, Diversity and Inclusion*, *41*(3), 352–382. https://doi.org/10.1108/EDI-06-2020-0172

Doyle, N., McDowall, A., & Waseem, U. (2022). Intersectional stigma for autistic people at work: A compound adverse impact effect on labor force participation and experiences of belonging. *Autism in Adulthood*, *4*(4), 340–356. https://doi.org/10.1089/aut.2021.0082

Drehmer, D., & LaVan, H. (1999). Diagnosing and making reasonable accommodation under ADA for attention-deficit hyperactivity disorder in adults. *SAM Advanced Management Journal*, *64*(3), 28–34. https://go.gale.com/ps/i.do?id=GALE%7CA55804651&sid=googleScholar&v=2.1&it=r&linkaccess=abs&issn=07497075&p=AONE&sw=w&userGroupName=anon%7Eb92ab137

Dubois, D. J., Holliday, N., Waddell, K., & Choffnes, D. (2024). Fair or Fare? Understanding automated transcription error bias in social media and videoconferencing platforms. *Proceedings of the International AAAI Conference on Web and Social Media*, *18*, 367–380. https://doi.org/10.1609/icwsm.v18i1.31320

Dunlop, P. D., Holtrop, D., & Wee, | Serena. (2022). *How asynchronous video interviews are used in practice: A study of an Australian-based AVI vendor*. https://doi.org/10.1111/ijsa.12372

Dwork, C., Hardt, M., Pitassi, T., Reingold, O., & Zemel, R. (2012). Fairness through awareness. *ITCS 2012 - Innovations in Theoretical Computer Science Conference*, 214–226. https://doi.org/10.1145/2090236.2090255

Ekman, P., & Friesen, W. V. (1978). *The facial action coding system (FACS)*. Consulting psychologist press.

Electronic Privacy Information Center. (2019). *Complaint and Request for Investigation, Injunction, and Other Relief*. https://epic.org/wp-content/uploads/privacy/ftc/hirevue/EPIC_FTC_HireVue_Complaint.pdf

Elia, J., & Devoto, M. (2007). ADHD genetics: 2007 update. *Current Psychiatry Reports*, *9*(5), 434–439. https://doi.org/10.1007/s11920-007-0057-z

Ellison, L. J., McClure Johnson, T., Tomczak, D., Siemsen, A., & Gonzalez, M. F. (2020). Game on! Exploring reactions to game-based selection assessments. *Journal of Managerial Psychology*, *35*(4), 241–254. https://doi.org/10.1108/JMP-09-2018-0414

Equal Employment Opportunity Commission. (1978). Uniform guidelines on Employee Selection Procedures. *Federal Register*, *43*(166), 38290–38315.

Equal Employment Opportunity Commission. (1991). *Civil Rights Act of 1991*. https://www.eeoc.gov/civil-rights-act-1991-original-text

Equal Employment Opportunity Commission. (2021). *Artificial Intelligence and Algorithmic Fairness Initiative*. https://www.eeoc.gov/ai

Equal Employment Opportunity Commission. (2022). *The Americans with Disabilities Act and the use of software, algorithms, and artificial intelligence to assess job applicants and employees*. https://www.eeoc.gov/laws/guidance/americans-disabilities-act-and-use-software-algorithms-and-artificial-intelligence

Equality Act (2010). https://www.legislation.gov.uk/ukpga/2010/15/contents

Erbeli, F., Rice, M., & Paracchini, S. (2022). Insights into dyslexia genetics research from the last two decades. *Brain Sciences*, *12*(1), 27. https://doi.org/10.3390/brainsci12010027

European Union. (2000). *Charter of Fundamental Rights of the European Union*.

Evett, L., & Brown, D. (2005). Text formats and web design for visually impaired and dyslexic readers - Clear Text for All. *Interacting with Computers*, *17*(4), 453–472. https://doi.org/10.1016/j.intcom.2005.04.001

Fairlie, R. W. (2017). Have we finally bridged the digital divide? Smart phone and Internet use patterns race and ethnicity. *EScholarship*.

Fassbender, C., & Schweitzer, J. B. (2006). Is there evidence for neural compensation in attention deficit hyperactivity disorder? A review of the functional neuroimaging literature. *Clinical Psychology Review*, *26*(4), 445–465. https://doi.org/10.1016/J.CPR.2006.01.003

Federal Register. (2011). *Regulations To Implement the Equal Employment Provisions of the Americans With Disabilities Act, as Amended*. https://www.federalregister.gov/documents/2011/03/25/2011-6056/regulations-to-implement-the-equal-employment-provisions-of-the-americans-with-disabilities-act-as

Feldman, M., Friedler, S. A., Moeller, J., Scheidegger, C., & Venkatasubramanian, S. (2015). Certifying and removing disparate impact. *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, *2015-Augus*, 259–268. https://doi.org/10.1145/2783258.2783311

Fenn, J., Tan, C.-S., & George, S. (2020). *Development, validation and translation of psychological tests*. https://doi.org/10.1192/bja.2020.33

Fergus, J. (2021). *A bookshelf in your job screening video makes you more hirable to AI*. https://www.inputmag.com/culture/a-bookshelf-in-your-job-screening-video-makes-you-more-hirable-to-ai

Filippi, G., Zannone, S., Hilliard, A., & Koshiyama, A. (2023). *Local Law 144: A Critical Analysis of Regression Metrics*. http://arxiv.org/abs/2302.04119

Fisher, G. G., Truxillo, D. M., Finkelstein, L. M., & Wallace, L. E. (2017). Age discrimination: Potential for adverse impact and differential prediction related to age. *Human Resource Management Review*, *27*(2), 316–327. https://doi.org/10.1016/J.HRMR.2016.06.001

Floridi, L., & Cowls, J. (2019). A unified framework of five principles for AI in society. *Harvard Data Science Review*, *1*(1). https://doi.org/10.1162/99608f92.8cd550d1

Floridi, L., Cowls, J., Beltrametti, M., Chatila, R., Chazerand, P., Dignum, V., Luetge, C., Madelin, R., Pagallo, U., Rossi, F., Schafer, B., Valcke, P., & Vayena, E. (2018). AI4People—An ethical framework for a good AI society: Opportunities, risks, principles, and recommendations. *Minds and Machines*, *28*(4), 689–707. https://doi.org/10.1007/s11023-018-9482-5

Flower, R. L., Hedley, D., Spoor, J. R., & Dissanayake, C. (2019). *An alternative pathway to employment for autistic job-seekers: a case study of a training and assessment program targeted to autistic job candidates*. https://doi.org/10.1080/13636820.2019.1636846

Foldes, H. J., Duehr, E. E., & Ones, D. S. (2008). Group differences in personality: Meta-analyses comparing five U.S. racial groups. *Personnel Psychology*, *61*(3), 579–616. https://doi.org/10.1111/j.1744-6570.2008.00123.x

Fontechia, S. A., Miltenberger, R. G., Smith, T. J., & Berkman, K. (2019). Evaluating video modeling for teaching professional E-mailing skills in transition-age job seekers with autism. *Journal of Applied Rehabilitation Counseling*, *50*(1), 73–90. https://doi.org/10.1891/0047-2220.50.1.73

Frauendorfer, D., & Mast, M. S. (2014). The impact of nonverbal behavior in the job interview. In *The Social Psychology of Nonverbal Communication* (pp. 220–247). Palgrave Macmillan.

Fritts, M., & Cabrera, F. (2021). AI recruitment algorithms and the dehumanization problem. *Ethics and Information Technology*, *23*(4), 791–801. https://doi.org/10.1007/s10676-021-09615-w

Fuermaier, A. B. M., Tucha, L., Butzbach, M., Weisbrod, M., Aschenbrenner, S., & Tucha, O. (2021). ADHD at the workplace: ADHD symptoms, diagnostic status, and work-related functioning. *Journal of Neural Transmission*, *128*(7), 1021–1031. https://doi.org/10.1007/s00702-021-02309-z

Fugita, S. S., Wexley, K. N., & Hillery, J. M. (1974). Black-white differences in nonverbal behavior in an interview setting. *Journal of Applied Social Psychology*, *4*(4), 343–350. https://doi.org/10.1111/j.1559-1816.1974.tb02606.x

Gable, S. L., & Haidt, J. (2005). What (and why) is positive psychology? *Review of General Psychology*, *9*(2), 103–110. https://doi.org/10.1037/1089-2680.9.2.103

Gajane, P., & Pechenizkiy, M. (2017). *On formalizing fairness in prediction with machine learning*. http://arxiv.org/abs/1710.03184

Gaugler, T., Klei, L., Sanders, S. J., Bodea, C. A., Goldberg, A. P., Lee, A. B., Mahajan, M., Manaa, D., Pawitan, Y., Reichert, J., Ripke, S., Sandin, S., Sklar, P., Svantesson, O., Reichenberg, A., Hultman, C. M., Devlin, B., Roeder, K., & Buxbaum, J. D. (2014). Most genetic risk for autism resides with common variation. *Nature Genetics*, *46*(8), 881. https://doi.org/10.1038/NG.3039

Georgiou, K., Gouras, A., & Nikolaou, I. (2019). Gamification in employee selection: The development of a gamified assessment. *International Journal of Selection and Assessment*, *27*(2), 91–103. https://doi.org/10.1111/ijsa.12240

Georgiou, K., & Nikolaou, I. (2020). Are applicants in favor of traditional or gamified assessment methods? Exploring applicant reactions towards a gamified selection method. *Computers in Human Behavior*, *109*, 106356. https://doi.org/10.1016/j.chb.2020.106356

Gerber, P. J., Ginsberg, R., & Reiff, H. B. (1992). Identifying alterable patterns in employment success for highly successful adults with learning disabilities. *Journal of Learning Disabilities*, *25*(8), 475–487. https://doi.org/10.1177/002221949202500802

Gerber, P. J., & Price, L. A. (2008). Self-Disclosure and Adults with Learning Disabilities: Practical Ideas about a Complex Process. *Learning Disabilities: A Multidisciplinary Journal*, *15*(1), 21–24. https://js.sagamorepub.com/index.php/ldmj/article/view/5399

Gershon, J. (2002). A meta-analytic review of gender differences in ADHD. *Journal of Attention Disorders*, *5*(3), 143–154. https://doi.org/10.1177/108705470200500302

Geschwind, D. H. (2011). Genetics of autism spectrum disorders. *Trends in Cognitive Sciences*, *15*(9), 409–416. https://doi.org/10.1016/J.TICS.2011.07.003

Gillespie-Lynch, K., Brooks, P. J., Someki, F., Obeid, R., Shane-Simpson, C., Kapp, S. K., Daou, N., & Smith, D. S. (2015). Changing college students' conceptions of autism: An online training to increase knowledge and decrease stigma. *Journal of Autism and Developmental Disorders*, *45*(8), 2553–2566. https://doi.org/10.1007/s10803-015-2422-9

Gilliland, S. W. (1993). The perceived fairness of selection systems: An organizational justice perspective. *Academy of Management Review*, *18*(4), 694–734. https://doi.org/10.5465/amr.1993.9402210155

Goel, N., Yaghini, M., & Faltings, B. (2018). Non-discriminatory machine learning through convex fairness criteria. *32nd AAAI Conference on Artificial Intelligence, AAAI 2018*, 3029–3036. https://doi.org/10.1145/3278721.3278722

Goldberg, L. R. (1992). The development of markers for the Big-Five factor structure. *Psychological Assessment*, *4*(1), 26–42. https://doi.org/10.1037/1040-3590.4.1.26

Gomathy, D. C. K. (2023). Workplace Diversity and its effects on team dynamics and productivity. *International Journal of Scientific Research in Engineering and Management*, *07*(06). https://doi.org/10.55041/ijsrem21469

Griswold, K. R., Phillips, J. M., Kim, M. S., Mondragon, N., Liff, J., & Gully, S. M. (2022). Global differences in applicant reactions to virtual interview synchronicity. *The International Journal of Human Resource Management*, *33*(15), 2991–3018. https://doi.org/10.1080/09585192.2021.1917641

Groves, L., Metcalf, J., Kennedy, A., Vecchione, B., & Strait, A. (2024). *Auditing work: Exploring the New York City algorithmic bias audit regime*. http://arxiv.org/abs/2402.08101

Guardiola, E. (2016). The gameplay loop: A player activity model for game design and analysis. *ACM International Conference Proceeding Series*. https://doi.org/10.1145/3001773.3001791

Guenole, N., Svensson, C., Wille, B., & Aloyan, K. (2023). A European perspective on psychometric measurement technology. In L. Tay, S. E. Woo, & T. Behrend (Eds.), *Technology and measurement around the globe* (pp. 271–307). Cambridge University Press.

Guidotti, R., Monreale, A., Ruggieri, S., Turini, F., Giannotti, F., & Pedreschi, D. (2018). A survey of methods for explaining black box models. *ACM Computing Surveys*, *51*(5). https://doi.org/10.1145/3236009

Guzzo, R. A., Fink, A. A., King, E., Tonidandel, S., & Landis, R. S. (2015). Big data recommendations for industrial-organizational psychology. *Industrial and Organizational Psychology*, *8*(4), 491–508. https://doi.org/10.1017/iop.2015.40

Hagner, D., & Cooney, B. F. (2005). "I do that for everybody": Supervising employees With autism. *Focus on Autism and Other Developmental Disabilities*, *20*(2), 91–97. https://doi.org/10.1177/10883576050200020501

Hand, C. J. (2023). Neurodiverse undergraduate psychology students' experiences of presentations in education and employment. *Journal of Applied Research in Higher Education*, *15*(5), 1600–1617. https://doi.org/10.1108/JARHE-03-2022-0106

Hardt, M., Price, E., & Srebro, N. (2016). Equality of opportunity in supervised learning. *Advances in Neural Information Processing Systems*, 3323–3331.

Hardy, J. H., Tey, K. S., Cyrus-Lai, W., Martell, R. F., Olstad, A., & Uhlmann, E. L. (2021). Bias in context: Small biases in hiring evaluations have big consequences. *Journal of Management*. https://doi.org/10.1177/0149206320982654

Harris, G. J., Chabris, C. F., Clark, J., Urban, T., Aharon, I., Steele, S., McGrath, L., Condouris, K., & Tager-Flusberg, H. (2006). Brain activation during semantic processing in autism spectrum disorders via functional magnetic resonance imaging. *Brain and Cognition*, *61*(1), 54–68. https://doi.org/10.1016/J.BANDC.2005.12.015

Hartnett, H. P., Stuart, H., Thurman, H., Loy, B., & Batiste, L. C. (2011). Employers' perceptions of the benefits of workplace accommodations: Reasons to hire, retain and promote people with disabilities. *Journal of Vocational Rehabilitation*, *34*(1), 17–23. https://doi.org/10.3233/JVR-2010-0530

Harvey, R. J., & Hammer, A. L. (1999). Item Response Theory. *The Counseling Psychologist*, *27*(3), 353–383. https://doi.org/10.1177/0011000099273004

Hauk, N., Hüffmeier, J., & Krumm, S. (2018). Ready to be a silver surfer? A meta-analysis on the relationship between chronological age and technology acceptance. *Computers in Human Behavior*, *84*, 304–319. https://doi.org/10.1016/j.chb.2018.01.020

Hausdorf, P. A., Leblanc, M. M., & Chawla, A. (2003). Cognitive ability testing and employment selection: Does test content relate to adverse impact? *Applied HRM Research*, *7*(2), 41–48. http://applyhrm.asp.radford.edu/2002/ms 7_2_ hausdorf.pdf

Hausknecht, J. P., Day, D. V., & Thomas, S. C. (2004). Applicant reactions to selection procedures: An updated model and meta-analysis. *Personnel Psychology*, *57*(3), 639–683. https://doi.org/10.1111/j.1744-6570.2004.00003.x

Hembree, R. (1988). Correlates, causes, effects, and treatment of test anxiety. *Review of Educational Research*, *58*(1), 47–77. https://doi.org/10.3102/00346543058001047

Hensel, W. F. (2017). People with autism spectrum disorder in the workplace: An expanding legal frontier. *Harvard Civil Rights-Civil Liberties Law Review*, *52*, 73–102. https://heinonline.org/HOL/Page?handle=hein.journals/hcrcl52&id=6

Herring, C. (2009). Does diversity pay?: Race, gender, and the business case for diversity. *American Sociological Review*, *74*(2), 208–224. https://doi.org/10.1177/000312240907400203

Hickman, L., Bosch, N., Ng, V., Saef, R., Tay, L., & Woo, S. E. (2021). Automated video interview personality assessments: Reliability, validity, and generalizability investigations. *Journal of Applied Psychology*. https://doi.org/10.1037/apl0000695

Hickman, L., Saef, R., Ng, V., Woo, S. E., Tay, L., & Bosch, N. (2021). Developing and evaluating language-based machine learning algorithms for inferring applicant personality in video interviews. *Human Resource Management Journal*. https://doi.org/10.1111/1748-8583.12356

Hickman., L., Tay, L., & Woo., S. E. (2019). Validity Investigation of Off-the-SHelf Language-Based Personality Assessment using Video Interviews: Convergent and

Discriminant Relationships with Self and Observer Ratings. *Personnel Assessment and Decisions*. https://www.researchgate.net/publication/332544721_Validity_Investigation_of_Off-the-Shelf_Language-Based_Personality_Assessment_using_Video_Interviews_Convergent_and_Discriminant_Relationships_with_Self_and_Observer_Ratings

Higgins, D. M., Peterson, J. B., Pihl, R. O., & Lee, A. G. M. (2007). Prefrontal cognitive ability, intelligence, Big Five personality, and the prediction of advanced academic and workplace performance. *Journal of Personality and Social Psychology*, *93*(2), 298–319. https://doi.org/10.1037/0022-3514.93.2.298

Hilliard, A., Guenole, N., Davies, J., & Kazim, E. (n.d.). Employment testing in the UK. In W. Arthur, D. Doverspike, & B. D. Schulte (Eds.), *Global perspectives on the definition, assessment, and reduction of bias and unfairness in employment testing*. Cambridge University Press and Assessment.

Hilliard, A., Guenole, N., & Leutner, F. (2022). Robots are judging me: Perceived fairness of algorithmic recruitment tools. *Frontiers in Psychology*, *13*. https://doi.org/10.3389/fpsyg.2022.940456

Hilliard, A., Gulley, A., Koshiyama, A., & Kazim, E. (2024). Bias audit laws: how effective are they at preventing bias in automated employment decision tools? *International Review of Law, Computers & Technology*, 1–17. https://doi.org/10.1080/13600869.2024.2403053

Hilliard, A., Kazim, E., Bitsakis, T., & Leutner, F. (2022a). Measuring personality through images: Validating a forced-choice image-based assessment of the Big Five personality traits. *Journal of Intelligence*, *10*(1), 12. https://doi.org/10.3390/jintelligence10010012

Hilliard, A., Kazim, E., Bitsakis, T., & Leutner, F. (2022b). Scoring a forced-choice image-based assessment of personality: A comparison of machine learning, regression, and summative approaches. *Acta Psychologica*, *228*, 103659. https://doi.org/10.1016/J.ACTPSY.2022.103659

HireVue. (2020). *Bias, AI ethics and the HireVue approach*. https://www.hirevue.com/why-hirevue/ethical-ai

HireVue. (2021). *HireVue customers conduct over 1 million video interviews in just 30 days*. https://www.hirevue.com/press-release/hirevue-customers-conduct-over-1-million-video-interviews-in-just-30-days

Hogan, R., Chamorro-Premuzic, T., & Kaiser, R. B. (2013). Employability and career success: Bridging the gap between theory and reality. *Industrial and Organizational Psychology*, *6*(1), 3–16. https://doi.org/10.1111/iops.12001

Hogan, R., Hogan, J., & Roberts, B. W. (1996). Personality measurement and employment decisions: Questions and answers. *American Psychologist*, *51*(5), 469–477. https://doi.org/10.1037/0003-066X.51.5.469

Hontangas, P. M., de la Torre, J., Ponsoda, V., Leenen, I., Morillo, D., & Abad, F. J. (2015). Comparing traditional and IRT scoring of forced-choice tests. *Applied Psychological Measurement*, *39*(8), 598–612. https://doi.org/10.1177/0146621615585851

Hoogman, M., Stolte, M., Baas, M., & Kroesbergen, E. (2020). Creativity and ADHD: A review of behavioral studies, the effect of psychostimulants and neural underpinnings. *Neuroscience & Biobehavioral Reviews*, *119*, 66–85. https://doi.org/10.1016/J.NEUBIOREV.2020.09.029

Hoppe, S., Loetscher, T., Morey, S. A., & Bulling, A. (2018). Eye movements during everyday behavior predict personality traits. *Frontiers in Human Neuroscience*, *12*, 105. https://doi.org/10.3389/fnhum.2018.00105

Hough, L. M., Oswald, F. L., & Ployhart, R. E. (2001). Determinants, detection and amelioration of adverse impact in personnel selection procedures: Issues, evidence and lessons learned. *International Journal of Selection and Assessment*, *9*(1–2), 152–194. https://doi.org/10.1111/1468-2389.00171

Huang, Y., Hwang, Y. I., Arnold, S. R. C., Lawson, L. P., Richdale, A. L., & Trollor, J. N. (2022). Autistic adults' experiences of diagnosis disclosure. *Journal of Autism and Developmental Disorders*, *52*(12), 5301–5307. https://doi.org/10.1007/s10803-021-05384-z

Hughes, D. J. (2017). Psychometric validity: Establishing the accuracy and appropriateness of psychometric measures. In *The Wiley Handbook of Psychometric Testing: A Multidisciplinary Reference on Survey, Scale and Test Development* (Vols. 2–2, pp. 751–779). John Wiley & Sons, Ltd. https://doi.org/10.1002/9781118489772.ch24

Hughes, J. A. (2021). Does the heterogeneity of autism undermine the neurodiversity paradigm? *Bioethics*, *35*(1), 47–60. https://doi.org/10.1111/BIOE.12780

Hunkenschroer, A. L., & Luetge, C. (2022). Ethics of AI-Enabled Recruiting and Selection: A Review and Research Agenda. *Journal of Business Ethics 2022*, *1*(4), 1–31. https://doi.org/10.1007/S10551-022-05049-6

Hutchinson, B., & Mitchell, M. (2019). 50 years of test (un)fairness. *Proceedings of the Conference on Fairness, Accountability, and Transparency*, 49–58. https://doi.org/10.1145/3287560.3287600

Hutchinson, V. D., Rehfeldt, R. A., Hertel, I., & Root, W. B. (2019). Exploring the efficacy of acceptance and commitment therapy and behavioral skills training to teach interview skills to adults with autism spectrum disorders. *Advances in Neurodevelopmental Disorders*, *3*(4), 450–456. https://doi.org/10.1007/s41252-019-00136-8

Illinois General Assembly. (2020). *820 ICLS 42 - Artificial Intelligence Video Interview Act*. https://www.ilga.gov/legislation/ilcs/ilcs3.asp?ActID=4015&ChapterID=68

Imana, B., Korolova, A., & Heidemann, J. (2021). Auditing for discrimination in algorithms delivering job ads. *The Web Conference 2021 - Proceedings of the World Wide Web Conference, WWW 2021*, *12*, 3767–3778. https://doi.org/10.1145/3442381.3450077

International Dyslexia Association. (2016). *How widespread is dyslexia?* https://dyslexiaida.org/how-widespread-is-dyslexia/

Jacobucci, R., Grimm, K. J., & McArdle, J. J. (2016). Regularized structural equation modeling. *Structural Equation Modeling: A Multidisciplinary Journal*, *23*(4), 555–566. https://doi.org/10.1080/10705511.2016.1154793

Jastrowski, K. E., Berlin, K. S., Sato, A. F., & Davies, W. H. (2007). Disclosure of attention-deficit/hyperactivity disorder may minimize risk of social rejection. *Psychiatry*, *70*(3), 274–282. https://doi.org/10.1521/psyc.2007.70.3.274

Jobin, A., Ienca, M., & Vayena, E. (2019). The global landscape of AI ethics guidelines. *Nature Machine Intelligence*, *1*(9), 389–399. https://doi.org/10.1038/s42256-019-0088-2

Johnson, J. A. (2014). Measuring thirty facets of the Five Factor Model with a 120-item public domain inventory: Development of the IPIP-NEO-120. *Journal of Research in Personality*, *51*, 78–89. https://doi.org/10.1016/j.jrp.2014.05.003

Johnson, K. R., Ennis-Cole, D., & Bonhamgregory, M. (2020). Workplace success strategies for employees With autism spectrum disorder: A new frontier for human resource development. *Https://Doi.Org/10.1177/1534484320905910*, *19*(2), 122–151. https://doi.org/10.1177/1534484320905910

Jordan, J. A., McGladdery, G., & Dyer, K. (2014). Dyslexia in higher education: Implications for maths anxiety, statistics anxiety and psychological well-being. *Dyslexia*, *20*(3), 225–240. https://doi.org/10.1002/dys.1478

Judge, T. A., Erez, A., Bono, J. E., & Thoresen, C. J. (2003). The core self-evaluations scale: Development of a measure. *Personnel Psychology*, *56*(2), 303–331. https://doi.org/10.1111/j.1744-6570.2003.tb00152.x

Judge, T. A., Higgins, C. A., Thoresen, C. J., & Barrick, M. R. (1999). The big five personality traits, general mental ability, and career success across the life span. *Personnel Psychology*, *52*(3), 621–652. https://doi.org/10.1111/j.1744-6570.1999.tb00174.x

Junsomboon, N., & Phienthrakul, T. (2017). Combining over-sampling and under-sampling techniques for imbalance dataset. *ACM International Conference Proceeding Series*, *Part F1283*, 243–247. https://doi.org/10.1145/3055635.3056643

Kahathuduwa, C. N., Wakefield, S., West, B. D., Blume, J., Dassanayake, T. L., Weerasinghe, V. S., & Mastergeorge, A. (2020). Effects of l-theanine–caffeine combination on sustained attention and inhibitory control among children with ADHD: a proof-of-

concept neuroimaging RCT. *Scientific Reports*, *10*(1), 13072. https://doi.org/10.1038/s41598-020-70037-7

Kahn, J. (2021). *HireVue stops using facial expressions to assess job candidates amid audit of its AI algorithms*. https://fortune.com/2021/01/19/hirevue-drops-facial-monitoring-amid-a-i-algorithm-audit/

Kaibel, C., Koch-Bayram, I., Biemann, T., & Mühlenbock, M. (2019). Applicant perceptions of hiring algorithms - Uniqueness and discrimination experiences as moderators. *79th Annual Meeting of the Academy of Management 2019: Understanding the Inclusive Organization, AoM 2019*, *2019*(1), 18172. https://doi.org/10.5465/AMBPP.2019.210

Kaiser, H. F. (1974). An index of factorial simplicity. *Psychometrika*, *39*(1), 31–36. https://doi.org/10.1007/BF02291575

Kallio, H., Pietilä, A. M., Johnson, M., & Kangasniemi, M. (2016). Systematic methodological review: developing a framework for a qualitative semi-structured interview guide. *Journal of Advanced Nursing*, *72*(12), 2954–2965. https://doi.org/10.1111/JAN.13031

Kamiran, F., & Calders, T. (2012). Data preprocessing techniques for classification without discrimination. *Knowledge and Information Systems*, *33*(1), 1–33. https://doi.org/10.1007/s10115-011-0463-8

Kamishima, T., Akaho, S., & Sakuma, J. (2011). Fairness-aware learning through regularization approach. *Proceedings - IEEE International Conference on Data Mining, ICDM*, 643–650. https://doi.org/10.1109/ICDMW.2011.83

Kannangara, C., Carson, J., Puttaraju, S., & Allen, R. (2018). Not All Those Who Wander are Lost: Examining the Character Strengths of Dyslexia. *Global Journal of Intellectual & Developmental Disabilities*, *4*(5). https://doi.org/10.19080/GJIDD.2018.04.555648

Karunasingha, D. S. K. (2022). Root mean square error or mean absolute error? Use their ratio as well. *Information Sciences*, *585*, 609–629. https://doi.org/10.1016/J.INS.2021.11.036

Kashdan, T. B., Gallagher, M. W., Silvia, P. J., Winterstein, B. P., Breen, W. E., Terhar, D., & Steger, M. F. (2009). The curiosity and exploration inventory-II: Development, factor structure, and psychometrics. *Journal of Research in Personality*, *43*(6), 987–998. https://doi.org/10.1016/J.JRP.2009.04.011

Kassab, K., & Kashevnik, A. (2024). Personality Traits Estimation Based on Job Interview Video Analysis: Importance of Human Nonverbal Cues Detection. *Big Data and Cognitive Computing 2024, Vol. 8, Page 173*, *8*(12), 173. https://doi.org/10.3390/BDCC8120173

Kazim, E., & Koshiyama, A. (2020). A high-level overview of AI ethics. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.3609292

Kazim, E., Koshiyama, A. S., Hilliard, A., & Polle, R. (2021). Systematizing audit in algorithmic recruitment. *Journal of Intelligence*, *9*(3), 46. https://doi.org/10.3390/jintelligence9030046

Kerz, E., Qiao, Y., Zanwar, S., & Wiechmann, D. (2022). ♠ SPADE: A big five-Mturk dataset of argumentative speech enriched with socio-demographics for personality detection. *2022 Language Resources and Evaluation Conference, LREC 2022*, 6405–6419. https://aclanthology.org/2022.lrec-1.688/

Kim, H.-Y. (2013). Statistical notes for clinical researchers: assessing normal distribution (2) using skewness and kurtosis. *Restorative Dentistry & Endodontics*, *38*(1), 52. https://doi.org/10.5395/rde.2013.38.1.52

Klein, R. M., Dilchert, S., Ones, D. S., & Dages, K. D. (2015). *Cognitive predictors and age-based adverse impact Among Business Executives*. https://doi.org/10.1037/a0038991

Kleinberg, J., Mullainathan, S., & Raghavan, M. (2016). *Inherent trade-offs in the fair determination of risk scores*.

Köchling, A., Wehner, M. C., & Warkocz, J. (2022). Can I show my skills? Affective responses to artificial intelligence in the recruitment process. *Review of Managerial Science*, 1–30. https://doi.org/10.1007/S11846-021-00514-4

Koopmans, L., Bernaards, C. M., Hildebrandt, V. H., De Vet, H. C. W., & Van Der Beek, A. J. (2014). Measuring individual work performance: Identifying and selecting indicators. *Work*, *48*(2), 229–238. https://doi.org/10.3233/WOR-131659

Koshiyama, A., Kazim, E., Treleaven, P., Rai, P., Szpruch, L., Pavey, G., Ahamat, G., Leutner, F., Goebel, R., Knight, A., Adams, J., Hitrova, C., Barnett, J., Nachev, P., Barber, D., Chamorro-Premuzic, T., Klemmer, K., Gregorovic, M., Khan, S., & Lomas, E. (2021). Towards algorithm auditing: A survey on managing legal, ethical and technological risks of AI, ML and associated algorithms. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.3778998

Kosinski, M., Stillwell, D., & Graepel, T. (2013). Private traits and attributes are predictable from digital records of human behavior. *Proceedings of the National Academy of Sciences of the United States of America*, *110*(15), 5802–5805. https://doi.org/10.1073/pnas.1218772110

Krainikovsky, S., Melnikov, M. Y., & Samarev, R. (2019). Estimation of psychometric data based on image preferences. *Conference Proceedings for Education and Humanities, WestEastInstitute*, 75–82. https://www.westeastinstitute.com/wp-content/uploads/2019/06/EDU-Vienna-Conference-Proceedings-2019.pdf#page=75

Krosnick, J. A. (1991). Response strategies for coping with the cognitive demands of attitude measures in surveys. *Applied Cognitive Psychology*, *5*(3), 213–236. https://doi.org/10.1002/acp.2350050305

Kumar, U., Reganti, A. N., Maheshwari, T., Chakroborty, T., Gambäck, B., & Das, A. (2018). Inducing personalities and values from language use in social network communities. *Information Systems Frontiers*, *20*(6), 1219–1240. https://doi.org/10.1007/s10796-017-9793-8

Kuncel, N. R., Ones, D. S., & Sackett, P. R. (2010). Individual differences as predictors of work, educational, and broad life outcomes. *Personality and Individual Differences*, *49*(4), 331–336. https://doi.org/10.1016/j.paid.2010.03.042

Kusner, M., Loftus, J., Russell, C., & Silva, R. (2017). Counterfactual fairness. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, & R. Garnett

(Eds.), *Advances in Neural Information Processing Systems* (Vols. 2017-Decem, pp. 4067–4077). Curran Associates, Inc. https://proceedings.neurips.cc/paper_files/paper/2017/file/a486cd07e4ac3d270571622f4f316ec5-Paper.pdf

Ladders Inc. (2018). *Eye-Tracking Study 2018*. https://www.theladders.com/static/images/basicSite/pdfs/TheLadders-EyeTracking-StudyC2.pdf

Lambrecht, A., & Tucker, C. (2019). Algorithmic bias? An empirical study of apparent gender-based discrimination in the display of stem career ads. *Management Science*, *65*(7), 2966–2981. https://doi.org/10.1287/mnsc.2018.3093

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

Landers, R. N., Auer, E. M., Mersy, G., Marin, S., & Blaik, J. (2022). You are what you click: using machine learning to model trace data for psychometric measurement. *International Journal of Testing*, *22*(3–4), 243–263. https://doi.org/10.1080/15305058.2022.2134394

Landers, R. N., & Behrend, T. S. (2022). Auditing the AI auditors: A framework for evaluating fairness and bias in high stakes AI predictive models. *American Psychologist*. https://doi.org/10.1037/AMP0000972

Landers, R. N., Brusso, R. C., Cavanaugh, K. J., & Collmus, A. B. (2016). A primer on theory-driven web scraping: Automatic extraction of big data from the internet for use in psychological research. *Psychological Methods*, *21*(4), 475–492. https://doi.org/10.1037/met0000081

Landers, R. N., & Sanchez, D. R. (2022). Game-based, gamified, and gamefully designed assessments for employee selection: Definitions, distinctions, design, and validation. *International Journal of Selection and Assessment*, *30*(1), 1–13. https://doi.org/10.1111/ijsa.12376

Landers, R. N., Tondello, G. F., Kappen, D. L., Collmus, A. B., Mekler, E. D., & Nacke, L. E. (2019). Defining gameful experience as a psychological state caused by gameplay: Replacing the term 'Gamefulness' with three distinct constructs. *International Journal of Human-Computer Studies*, *127*, 81–94. https://doi.org/10.1016/J.IJHCS.2018.08.003

Langer, M., Baum, K., König, C. J., Hähne, V., Oster, D., & Speith, T. (2021). Spare me the details: How the type of information about automated interviews influences applicant reactions. *International Journal of Selection and Assessment*, *29*(2), 154–169. https://doi.org/10.1111/ijsa.12325

Langer, M., König, C. J., & Fitili, A. (2018). Information as a double-edged sword: The role of computer experience and information on applicant reactions towards novel technologies for personnel selection. *Computers in Human Behavior*, *81*, 19–30. https://doi.org/10.1016/j.chb.2017.11.036

Langer, M., König, C. J., & Papathanasiou, M. (2019). Highly automated job interviews: Acceptance under the influence of stakes. *International Journal of Selection and Assessment*, *27*(3), 217–234. https://doi.org/10.1111/ijsa.12246

Laurano, M. (2022). *The power of AI in talent acquisition*.

Le, H., Oh, I.-S., Robbins, S. B., Ilies, R., Holland, E., & Westrick, P. (2011). Too much of a good thing: Curvilinear relationships between personality traits and job performance. *Journal of Applied Psychology*, *96*(1), 113–133. https://doi.org/10.1037/a0021016

Leather, C., Hogh, H., Seiss, E., & Everatt, J. (2011). Cognitive functioning and work success in adults with dyslexia. *Dyslexia*, *17*(4), 327–338. https://doi.org/10.1002/dys.441

Leather, C., & Kirwan, B. (2012). Achieving success in the workplace. In N. Brunswick (Ed.), *Supporting Dyslexic Adults in Higher Education and the Workplace* (pp. 157–166). Wiley Blackwell. https://doi.org/10.1002/9781119945000.ch16

Lee, M. K. (2018). Understanding perception of algorithmic decisions: Fairness, trust, and emotion in response to algorithmic management. *Big Data and Society*, *5*(1). https://doi.org/10.1177/2053951718756684

Leutner, F., & Chamorro-Premuzic, T. (2018). Stronger together: Personality, intelligence and the assessment of career potential. *Journal of Intelligence*, *6*(4), 1–10. https://doi.org/10.3390/jintelligence6040049

Leutner, F., Codreanu, S.-C., Brink, S., & Bitsakis, T. (2023). Game based assessments of cognitive ability in recruitment: Validity, fairness and test-taking experience. *Frontiers in Psychology*, *13*. https://doi.org/10.3389/fpsyg.2022.942662

Leutner, F., Codreanu, S.-C., Liff, J., & Mondragon, N. (2021). The potential of game- and video-based assessments for social attributes: examples from practice. *Journal of Managerial Psychology*, *36*(7), 533–547. https://doi.org/10.1108/JMP-01-2020-0023

Leutner, F., Yearsley, A., Codreanu, S. C., Borenstein, Y., & Ahmetoglu, G. (2017). From Likert scales to images: Validating a novel creativity measure with image based response scales. *Personality and Individual Differences*, *106*, 36–40. https://doi.org/10.1016/j.paid.2016.10.007

Levashina, J., Hartwell, C. J., Morgeson, F. P., & Campion, M. A. (2014). The structured employment interview: Narrative and quantitative review of the research literature. *Personnel Psychology*, *67*(1), 241–293. https://doi.org/10.1111/peps.12052

Levy, F., Hay, D. A., Bennett, K. S., & McStephen, M. (2005). Gender differences in ADHD subtype comorbidity. *Journal of the American Academy of Child and Adolescent Psychiatry*, *44*(4), 368–376. https://doi.org/10.1097/01.chi.0000153232.64968.c1

Lewandowski, L., Hendricks, K., & Gordon, M. (2015). Test-Taking Performance of High School Students With ADHD. *Journal of Attention Disorders*, *19*(1), 27–34. https://doi.org/10.1177/1087054712449183

Li, L., Lassiter, T., Lee, M. K., & Oh, J. (2021). Algorithmic Hiring in Practice: Recruiter and HR Professional's Perspectives on AI Use in Hiring; Algorithmic Hiring in Practice: Recruiter and HR Professional's Perspectives on AI Use in Hiring. *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, *11*(21). https://doi.org/10.1145/3461702

Lieberoth, A. (2015). Shallow gamification: Testing psychological effects of framing an activity as a game. *Games and Culture*, *10*(3), 229–248. https://doi.org/10.1177/1555412014559978

Lim, B. C., & Ployhart, R. E. (2006). Assessing the convergent and discriminant validity of Goldberg's international personality item pool: A multitrait-multimethod examination. *Organizational Research Methods*, *9*(1), 29–54. https://doi.org/10.1177/1094428105283193

Lindsay, S., Osten, V., Rezai, M., & Bui, S. (2021). Disclosure and workplace accommodations for people with autism: a systematic review. *Disability and Rehabilitation*, *43*(5), 597–610. https://doi.org/10.1080/09638288.2019.1635658

Lipton, Z. C., Chouldechova, A., & McAuley, J. (2018). Does mitigating ML's impact disparity require treatment disparity? *Advances in Neural Information Processing Systems*, *2018-Decem*, 8125–8135.

Ljepava, N. (2023). Examining the role of participants' personality traits on data quality in online panel surveys. *International Journal for Quality Research*, *18*(2), 515–530.

Lloyd, K. (2018). *Bias Amplification in Artificial Intelligence Systems*. https://doi.org/10.48550/arxiv.1809.07842

Locke, R., Alexander, G., Mann, R., Kibble, S., & Scallan, S. (2017). Doctors with dyslexia: strategies and support. *The Clinical Teacher*, *14*(5), 355–359. https://doi.org/10.1111/TCT.12578

Lodi-Smith, J., Rodgers, J. D., Cunningham, S. A., Lopata, C., & Thomeer, M. L. (2019). Meta-analysis of Big Five personality traits in autism spectrum disorder. *Autism*, *23*(3), 556–565. https://doi.org/10.1177/1362361318766571

Lofink, C. R. (2021). Resume analysis: A comparison of two methods. *JBERT*, *3*, 81–87. https://www.nabet.us/j_archives/JBET_2021.pdf#page=81

Logan, J. (2009). Dyslexic entrepreneurs: the incidence; their coping strategies and their business skills. *Dyslexia*, *15*(4), 328–346. https://doi.org/10.1002/DYS.388

Logan, J., Hendry, C., Courtney, N., & Brown, J. (2008). *Unlocking the potential of the UK's Hidden Innovators*. www.cass.city.ac.uk/centive/current/entrepreneurship_innovation.html

Lombardi, A. R., Murray, C., & Gerdes, H. (2011). College faculty and inclusive instruction: Self-reported attitudes and actions pertaining to Universal Design. *Journal of Diversity in Higher Education*, *4*(4), 250–261. https://doi.org/10.1037/a0024961

London, A. S., & Landes, S. D. (2021). Cohort change in the prevalence of ADHD among U.S. adults: Evidence of a gender-specific historical period effect. *Journal of Attention Disorders*, *25*(6), 771–782. https://doi.org/10.1177/1087054719855689

Lorenz, T., Frischling, C., Cuadros, R., & Heinitz, K. (2016). Autism and overcoming job barriers: Comparing job-related barriers and possible solutions in and outside of autism-specific employment. *PLoS ONE*, *11*(1), e0147040. https://doi.org/10.1371/journal.pone.0147040

Louizos, C., Swersky, K., Li, Y., Welling, M., & Zemel, R. (2016). The variational fair autoencoder. *4th International Conference on Learning Representations, ICLR 2016 - Conference Track Proceedings*.

Loyer Carbonneau, M., Demers, M., Bigras, M., & Guay, M.-C. (2021). Meta-analysis of sex differences in ADHD symptoms and associated cognitive deficits. *Journal of Attention Disorders*, *25*(12), 1640–1656. https://doi.org/10.1177/1087054720923736

Lum, K., & Johndrow, J. E. (2016). *A statistical framework for fair predictive algorithms*.

Maisog, J. M., Einbinder, E. R., Flowers, D. L., Turkeltaub, P. E., & Eden, G. F. (2008). A meta-analysis of functional neuroimaging studies of dyslexia. *Annals of the New York Academy of Sciences*, *1145*(1), 237–259. https://doi.org/10.1196/annals.1416.024

Mao, A. R., Brams, M., Babcock, T., & Madhoo, M. (2011). A Physician's Guide to Helping Patients with ADHD Find Success in the Workplace. *Postgraduate Medicine*, *123*(5), 60–70. https://doi.org/10.3810/pgm.2011.09.2460

Maras, K., Norris, J. E., Nicholson, J., Heasman, B., Remington, A., & Crane, L. (2021). Ameliorating the disadvantage for autistic job seekers: An initial evaluation of adapted

employment interview questions. *Autism*, *25*(4), 1060–1075. https://doi.org/10.1177/1362361320981319

Marshall, J. E., Fearon, C., Highwood, M., & Warden, K. (2020). “What should I say to my employer… if anything?”- My disability disclosure dilemma. *International Journal of Educational Management*, *34*(7), 1105–1117. https://doi.org/10.1108/IJEM-01-2020-0028

Martin, M. M., & Rubin, R. B. (1995). A new measure of cognitive flexibility. *Psychological Reports*, *76*(2), 623–626. https://doi.org/10.2466/pr0.1995.76.2.623

Mascheretti, S., De Luca, A., Trezzi, V., Peruzzo, D., Nordio, A., Marino, C., & Arrigoni, F. (2017). Neurogenetics of developmental dyslexia: from genes to behavior through brain neuroimaging and cognitive and sensorial mechanisms. *Translational Psychiatry 2017 7:1*, *7*(1), e987–e987. https://doi.org/10.1038/tp.2016.240

Masuch, T. V., Bea, M., Alm, B., Deibler, P., & Sobanski, E. (2019). Internalized stigma, anticipated discrimination and perceived public stigma in adults with ADHD. *ADHD Attention Deficit and Hyperactivity Disorders*, *11*(2), 211–220. https://doi.org/10.1007/s12402-018-0274-9

Maurer, R. (2022). *SHRM Research: AI Use on the Rise, Ethics Questions Remain*. Society for Human Resource Management. https://www.shrm.org/topics-tools/news/technology/shrm-research-ai-use-rise-ethics-questions-remain

Mavridis, A., & Tsiatsos, T. (2017). Game-based assessment: investigating the impact on test anxiety and exam performance. *Journal of Computer Assisted Learning*, *33*(2), 137–150. https://doi.org/10.1111/jcal.12170

McCarthy, J. M., Bauer, T. N., Truxillo, D. M., Anderson, N. R., Costa, A. C., & Ahmed, S. M. (2017). Applicant perspectives during selection: A review addressing “so what?,” “what’s new?,” and “where to next?” *Journal of Management*, *43*(6), 1693–1725. https://doi.org/10.1177/0149206316681846

McCarthy, J. M., & Goffin, R. D. (2005). Selection test anxiety: Exploring tension and fear of failure across the sexes in simulated selection scenarios. *International Journal of*

Selection and Assessment, *13*(4), 282–295. https://doi.org/10.1111/j.1468-2389.2005.00325.x

McColl, R., & Michelotti, M. (2019). Sorry, could you repeat the question? Exploring video-interview recruitment practice in HRM. *Human Resource Management Journal*, *29*(4), 637–656. https://doi.org/10.1111/1748-8583.12249

McCord, J. L., Harman, J. L., & Purl, J. (2019). Game-like personality testing: An emerging mode of personality assessment. *Personality and Individual Differences*, *143*, 95–102. https://doi.org/10.1016/j.paid.2019.02.017

McCrae, R. R., & Costa, P. T. (1985). Updating Norman's "adequacy taxonomy": Intelligence and personality dimensions in natural language and in questionnaires. *Journal of Personality and Social Psychology*, *49*(3), 710–721. https://doi.org/10.1037/0022-3514.49.3.710

McDowall, A., Doyle, N., & Kiseleva, M. (2023). *Neurodiversity at work: demand, supply and a gap analysis.* Birkbeck, University of London.

McEvoy, S. A. (1993). Discrimination in employment and education because of dyslexia. *North East Journal of Legal Studies*, *1*. https://heinonline.org/HOL/Page?handle=hein.journals/neastjol1&id=54&div=7&collection=journals

McIntosh, C. K., Hyde, S. A., Bell, M. P., & Yeatts, P. E. (2023). Thriving at work with ADHD: antecedents and outcomes of proactive disclosure. *Equality, Diversity and Inclusion: An International Journal*, *42*(2), 228–247. https://doi.org/10.1108/EDI-02-2022-0033

McKinsey & Company. (2017). A future that works: Automation, employment, and productivity. In *McKinsey Global Institute* (Issue January).

McLoughlin, D. (2015). Career development and individuals with dyslexia. *Career Planning and Adult Development Journal*, *31*(4), 151–161. https://www.proquest.com/docview/1845685193?accountid=11149&parentSessionId=OyB9QsjRrGQ19vp%2BO%2Brqu4b6ZFwealIGfFvvL1hUmsk%3D&pq-origsite=primo

McNeish, D. M. (2015). Using Lasso for predictor selection and to assuage overfitting: a method long overlooked in behavioral sciences. *Multivariate Behavioral Research*, *50*(5), 471–484. https://doi.org/10.1080/00273171.2015.1036965

Meissner, F., & Rothermund, K. (2015). A thousand words are worth more than a picture? The effects of stimulus modality on the implicit association test. *Social Psychological and Personality Science*, *6*(7), 740–748. https://doi.org/10.1177/1948550615580381

Melchers, K. G., & Basch, J. M. (2021). Fair play? Sex-, age-, and job-related correlates of performance in a computer-based simulation game. *Int J Sel Assess*, 1–14. https://doi.org/10.1111/ijsa.12337

Melchers, K. G., Roulin, N., & Buehl, A. K. (2020). A review of applicant faking in selection interviews. *International Journal of Selection and Assessment*, *28*(2), 123–142. https://doi.org/10.1111/IJSA.12280

Mirowska, A., & Mesnet, L. (2021). Preferring the devil you know: Potential applicant reactions to artificial intelligence evaluation of interviews. *Human Resource Management Journal*, 1748-8583.12393. https://doi.org/10.1111/1748-8583.12393

Mitchell, S., Potash, E., Barocas, S., D'Amour, A., & Lum, K. (2021). Algorithmic fairness: Choices, assumptions, and definitions. *Annual Review of Statistics and Its Application*, *8*(1), 141–163. https://doi.org/10.1146/annurev-statistics-042720-125902

Montefiori, L. (2016). Game-based assessment: Face validity, fairness perception, and impact on employer's brand image. *Assessment & Development Matters*, *8*(2), 19–22.

Moody, S. (2010). Dyslexia in the workplace. In D. Bartlett, S. Moody, & K. Kindersley (Eds.), *Dyslexia in the Workplace: An Introductory Guide:* (2nd ed., pp. 3–11). Wiley-Blackwell. https://doi.org/10.1002/9780470669341.ch1

Moody, S. (2015). How to do a workplace needs assessment. In *Dyslexia and Employment: A Guide for Assessors, Trainers and Managers* (pp. 47–56). John Wiley & Sons, Ltd. https://doi.org/10.1002/9780470749203.CH5

Morgan, D. (2010). *Statistical significance standards for basic adverse impact analysis*. https://adverse-impact.com/wp-content/uploads/2015/02/Statistical-Significance-Testing-for-Adverse-Impact-Measurement.pdf

Morgan, L., Leatzow, A., Clark, S., & Siller, M. (2014). Interview skills for adults with autism spectrum disorder: A pilot randomized controlled trial. *Journal of Autism and Developmental Disorders*, *44*(9), 2290–2300. https://doi.org/10.1007/s10803-014-2100-3

Morris, D. K., & Turnbull, P. A. (2007). The disclosure of dyslexia in clinical practice: Experiences of student nurses in the United Kingdom. *Nurse Education Today*, *27*(1), 35–42. https://doi.org/10.1016/j.nedt.2006.01.017

Morris, S. B., & Lobsenz, R. B. (2000). Significance tests and confidence intervals for the adverse impact ratio. *Personnel Psychology*, *53*(1), 89–111. https://doi.org/10.1111/j.1744-6570.2000.tb00195.x

Mueller, A., Hong, D. S., Shepard, S., & Moore, T. (2017). Linking ADHD to the neural circuitry of attention. *Trends in Cognitive Sciences*, *21*(6), 474–488. https://doi.org/10.1016/J.TICS.2017.03.009

Mueller, L., Norris, D., & Oppler, S. (2007). Implementation based on alternate validation procedures: Ranking, cut, scores, banding, and compensatory models. In S. M. McPhail (Ed.), *Alternate Validation Strategies* (pp. 349–405).

Muhle, R., Trentacoste, S. V., & Rapin, I. (2004). The Genetics of autism. *Pediatrics*, *113*(5), e472–e486. https://doi.org/10.1542/peds.113.5.e472

Mullainathan, S., & Obermeyer, Z. (2017). Does machine learning automate moral hazard and error? *American Economic Review*, *107*(5), 476–480. https://doi.org/10.1257/aer.p20171084

Müller, E., Schuler, A., Burton, B. A., & Yates, G. B. (2003). Meeting the vocational support needs of individuals with Asperger Syndrome and other autism spectrum disabilities - IOS Press. *Journal of Vocational Rehabilitation, 18*(3), 163–175. https://content.iospress.com/articles/journal-of-vocational-rehabilitation/jvr00193

Murphy, K. R. (2009). How a broader definition of the criterion domain changes our thinking about adverse impact. In *Adverse impact: Implications for organizational staffing and high stakes selection* (pp. 137–160). https://doi.org/10.4324/9780203848418

Murphy, K. R., & Jacobs, R. R. (2012). Using effect size measures to reform the determination of adverse impact in equal employment litigation. *Psychology, Public Policy, and Law*, *18*(3), 477–499. https://doi.org/10.1037/a0026350

Nadeau, K. G. (2005). Career choices and workplace challenges for individuals with ADHD. *Journal of Clinical Psychology*, *61*(5), 549–563. https://doi.org/10.1002/JCLP.20119

Nadeau, K. G. (2013). ADD in the workplace: Career consultation and counseling for the adult with ADD. In *A Comprehensive Guide To Attention Deficit Disorder In Adults: Research, Diagnosis and Treatment* (pp. 308–344). Routledge.

Nagib, W., & Wilton, R. (2020). Gender matters in career exploration and job-seeking among adults with autism spectrum disorder: evidence from an online community. *Disability and Rehabilitation*, *42*(18), 2530–2541. https://doi.org/10.1080/09638288.2019.1573936

Nakao, T., Radua, J., Rubia, K., & Mataix-Cols, D. (2011). Gray matter volume abnormalities in ADHD: Voxel-based meta-analysis exploring the effects of age and stimulant medication. *American Journal of Psychiatry*, *168*(11), 1154–1163. https://doi.org/10.1176/appi.ajp.2011.11020281

Nalavany, B. A., Logan, J. M., & Carawan, L. W. (2018). The relationship between emotional experience with dyslexia and work self-efficacy among adults with dyslexia. *Dyslexia*, *24*(1), 17–32. https://doi.org/10.1002/DYS.1575

Nasteski, V. (2017). An overview of the supervised machine learning methods. *HORIZONS.B*, *4*, 51–62. https://doi.org/10.20544/HORIZONS.B.04.1.17.P05

Nelson, J. M., Lindstrom, W., & Foels, P. A. (2014). Test anxiety and college students with attention deficit hyperactivity disorder. *Journal of Psychoeducational Assessment*, *32*(6), 548–557. https://doi.org/10.1177/0734282914521978

Nelson, J. M., Lindstrom, W., & Foels, P. A. (2015). Test anxiety among college students with specific reading disability (dyslexia). *Journal of Learning Disabilities*, *48*(4), 422–432. https://doi.org/10.1177/0022219413507604

Nguyen, L. S., & Gatica-Perez, D. (2016). Hirability in the wild: Analysis of online conversational video resumes. *IEEE Transactions on Multimedia*, *18*(7), 1422–1437. https://doi.org/10.1109/TMM.2016.2557058

NHS Digital. (2014). *Mental health and wellbeing in England: Adult psychiatric morbidity survey 2014*. https://files.digital.nhs.uk/pdf/q/3/mental_health_and_wellbeing_in_england_full_report.pdf

Nicolaidis, C., Raymaker, D., Kapp, S. K., Baggs, A., Ashkenazy, E., McDonald, K., Weiner, M., Maslak, J., Hunter, M., & Joyce, A. (2019). The AASPIRE practice-based guidelines for the inclusion of autistic adults in research as co-researchers and study participants. *Autism*, *23*(8), 2007–2019. https://doi.org/10.1177/1362361319830523

Nicolaidis, C., Raymaker, D. M., Ashkenazy, E., McDonald, K. E., Dern, S., Baggs, A. E. V., Kapp, S. K., Weiner, M., & Boisclair, W. C. (2015). "Respect the way I need to communicate with you": Healthcare experiences of adults on the autism spectrum. *Autism : The International Journal of Research and Practice*, *19*(7), 824. https://doi.org/10.1177/1362361315576221

Nigg, J. T., Blaskey, L. G., Huang-Pollock, C. L., Hinshaw, S. P., John, O. P., Willcutt, E. G., & Pennington, B. (2002). Big five dimensions and ADHD symptoms: Links between personality traits and clinical symptoms. In *Journal of Personality and Social Psychology* (Vol. 83, Issue 2, pp. 451–469). https://doi.org/10.1037/0022-3514.83.2.451

Nigg, J. T., John, O. P., Blaskey, L. G., Huang-Pollock, C. L., Willcutt, E. G., Hinshaw, S. P., & Pennington, B. (2002). Big Five Dimensions and ADHD Symptoms: Links Between Personality Traits and Clinical Symptoms. *Journal of Personality and Social Psychology*, *83*(2), 451–469. https://doi.org/10.1037/0022-3514.83.2.451

Novita, S. (2016). Secondary symptoms of dyslexia: a comparison of self-esteem and anxiety profiles of children with and without dyslexia. *European Journal of Special Needs Education*, *31*(2), 279–288. https://doi.org/10.1080/08856257.2015.1125694

Office for National Statistics. (2021a). *Changing trends and recent shortages in the labour market, UK - Office for National Statistics*. https://www.ons.gov.uk/employmentandlabourmarket/peopleinwork/employmentandemployeetypes/articles/changingtrendsandrecentshortagesinthelabourmarketuk/2016to2021

Office for National Statistics. (2021b). *Outcomes for disabled people in the UK*. https://www.ons.gov.uk/peoplepopulationandcommunity/healthandsocialcare/disability/articles/outcomesfordisabledpeopleintheuk/2020#employment

Office of Federal Contract Compliance Programs. (2020). *Federal contract compliance manual*. https://www.dol.gov/sites/dolgov/files/OFCCP/FCCM/508_FCCM_05012020.pdf

Ohlms, M. L., Melchers, K. G., Uwe, |, & Kanning, P. (2023). *Can we playfully measure cognitive ability? Construct-related validity and applicant reactions*. https://doi.org/10.1111/ijsa.12450

O'Neil Risk Consulting and Algorithmic Auditing. (2020). *Description of Algorithmic Audit: Pre-built Assessments*.

O'Nions, E., Petersen, I., Buckman, J. E. J., Charlton, R., Cooper, C., Corbett, A., Happé, F., Manthorpe, J., Richards, M., Saunders, R., Zanker, C., Mandy, W., & Stott, J. (2023). Autism in England: assessing underdiagnosis in a population-based cohort study of prospectively collected primary care data. *The Lancet Regional Health - Europe*, *29*, 100626. https://doi.org/10.1016/j.lanepe.2023.100626

Oostrom, J. K., Holtrop, D., Koutsoumpis, A., van Breda, W., Ghassemi, S., & de Vries, R. E. (2024). Applicant reactions to algorithm- versus recruiter-based evaluations of an asynchronous video interview and a personality inventory. *Journal of Occupational and Organizational Psychology*, *97*(1), 160–189. https://doi.org/10.1111/JOOP.12465

Oswald, F. L., Behrend, T. S., Putka, D. J., & Sinar, E. (2020). Big data in industrial-organizational psychology and human resource management: Forward progress for

organizational research and practice. *Annual Review of Organizational Psychology and Organizational Behavior*, *7*(1), 505–533. https://doi.org/10.1146/annurev-orgpsych-032117-104553

Oswald, F. L., & Putka, D. J. (2020). Statistical methods for big data: A scenic tour. In S. Tonidandel, E. B. King, & J. M. Cortina (Eds.), *Big Data at Work* (pp. 57–77). Routledge. https://doi.org/10.4324/9781315780504-9

Palan, S., & Schitter, C. (2018). Prolific.ac—A subject pool for online experiments. *Journal of Behavioral and Experimental Finance*, *17*, 22–27. https://doi.org/10.1016/J.JBEF.2017.12.004

Palanica, A., Thommandram, A., Lee, A., Li, M., & Fossat, Y. (2019). Do you understand the words that are comin outta my mouth? Voice assistant comprehension of medication names. *Npj Digital Medicine*, *2*(1), 1–6. https://doi.org/10.1038/s41746-019-0133-x

Pan, P.-Y., Jonsson, U., Şahpazoğlu Çakmak, S. S., Häge, A., Hohmann, S., Nobel Norrman, H., Buitelaar, J. K., Banaschewski, T., Cortese, S., Coghill, D., & Bölte, S. (2022). Headache in ADHD as comorbidity and a side effect of medications: a systematic review and meta-analysis. *Psychological Medicine*, *52*(1), 14–25. https://doi.org/10.1017/S0033291721004141

Park, G., Andrew Schwartz, H., Eichstaedt, J. C., Kern, M. L., Kosinski, M., Stillwell, D. J., Ungar, L. H., & Seligman, M. E. P. (2015). Automatic personality assessment through social media language. *Journal of Personality and Social Psychology*, *108*(6), 934–952. https://doi.org/10.1037/pspp0000020

Parker, D. R., & Boutelle, K. (2009). Executive function coaching for college students with learning disabilities and ADHD: A new approach for fostering self-determination. *Learning Disabilities Research & Practice*, *24*(4), 204–215. https://doi.org/10.1111/j.1540-5826.2009.00294.x

Parker, J. D. A., Majeski, S. A., & Collin, V. T. (2004). ADHD symptoms and personality: relationships with the five-factor model. *Personality and Individual Differences*, *36*(4), 977–987. https://doi.org/10.1016/S0191-8869(03)00166-1

Patton, E. (2009). When diagnosis does not always mean disability: The challenge of employees with Attention Deficit Hyperactivity Disorder (ADHD). *Journal of Workplace Behavioral Health*, *24*(3), 326–343. https://doi.org/10.1080/15555240903176161

Paunonen, S. V., Ashton, M. C., & Jackson, D. N. (2001). Nonverbal assessment of the big five personality factors. *European Journal of Personality*, *15*(1), 3–18. https://doi.org/10.1002/per.385

Paunonen, S. V., Jackson, D. N., & Keinonen, M. (1990). The structured nonverbal assessment of personality. *Journal of Personality*, *58*(3), 481–502. https://doi.org/10.1111/j.1467-6494.1990.tb00239.x

Peer, E., Brandimarte, L., Samat, S., & Acquisti, A. (2017). Beyond the Turk: Alternative platforms for crowdsourcing behavioral research. *Journal of Experimental Social Psychology*, *70*, 153–163. https://doi.org/10.1016/J.JESP.2017.01.006

Pletzer, J. L., Oostrom, J. K., & de Vries, R. E. (2021). HEXACO personality and organizational citizenship behavior: A domain-and facet-level meta-analysis. *Human Performance*, *34*(2), 126–147. https://doi.org/10.1080/08959285.2021.1891072

Ployhart, R. E., & Holtz, B. C. (2008). The diversity-validity dilemma: Strategies for reducing racioethnic and sex subgroup differences and adverse impact in selection. *Personnel Psychology*, *61*(1), 153–172. https://doi.org/10.1111/j.1744-6570.2008.00109.x

Polanczyk, G., De Lima, M. S., Horta, B. L., Biederman, J., & Rohde, L. A. (2007). The worldwide prevalence of ADHD: a systematic review and metaregression analysis. *The American Journal of Psychiatry*, *164*(6), 942–948. https://doi.org/10.1176/AJP.2007.164.6.942

Powers, D. M. W. (2020). *Evaluation: From precision, recall and F-measure to ROC, informedness, markedness and correlation*.

Prevatt, F. (2016). Coaching for college students with ADHD. *Current Psychiatry Reports*, *18*(12), 110. https://doi.org/10.1007/s11920-016-0751-9

Prevatt, F., & Yelland, S. (2015). An empirical evaluation of ADHD coaching in college students. *Journal of Attention Disorders*, *19*(8), 666–677. https://doi.org/10.1177/1087054713480036

PricewaterhouseCoopers. (2017). *Women in tech: Time to close the gender gap*.

PSI. (2018). *Understanding adverse impact in the hiring process*.

Putka, D. J., Beatty, A. S., & Reeder, M. C. (2018). Modern prediction methods: New perspectives on a common problem. *Organizational Research Methods*, *21*(3), 689–732. https://doi.org/10.1177/1094428117697041

Quiroga, M. Á., Escorial, S., Román, F. J., Morillo, D., Jarabo, A., Privado, J., Hernández, M., Gallego, B., & Colom, R. (2015). Can we reliably measure the general factor of intelligence (g) through commercial video games? Yes, we can! *Intelligence*, *53*, 1–7. https://doi.org/10.1016/j.intell.2015.08.004

Quiroga, M. Á., Román, F. J., De La Fuente, J., Privado, J., & Colom, R. (2016). The measurement of intelligence in the XXI century using video games. *The Spanish Journal of Psychology*, *19*, E89. https://doi.org/10.1017/sjp.2016.84

Raghavan, M., & Barocas, S. (2019). *Challenges for mitigating bias in algorithmic hiring*. https://www.brookings.edu/research/challenges-for-mitigating-bias-in-algorithmic-hiring/

Raghavan, M., Barocas, S., Kleinberg, J., & Levy, K. (2020). Mitigating bias in algorithmic hiring: Evaluating claims and practices. *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, 469–481. https://doi.org/10.1145/3351095.3372828

Raji, I. D., Smart, A., White, R. N., Mitchell, M., Gebru, T., Hutchinson, B., Smith-Loud, J., Theron, D., & Barnes, P. (2020). Closing the AI accountability gap: Defining an end-to-end framework for internal algorithmic auditing. *FAT* 2020 - Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, 33–44. https://doi.org/10.1145/3351095.3372873

Ramus, F. (2014). Neuroimaging sheds new light on the phonological deficit in dyslexia. *Trends in Cognitive Sciences*, *18*(6), 274–275. https://doi.org/10.1016/J.TICS.2014.01.009

Raudys, S. J., & Jain, A. K. (1991). Small sample size effects in statistical pattern recognition: recommendations for practitioners. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *13*(3), 252–264. https://doi.org/10.1109/34.75512

Rello, L., & Baeza-Yates, R. (2013). Good fonts for dyslexia. *Proceedings of the 15th International ACM SIGACCESS Conference on Computers and Accessibility, ASSETS 2013*. https://doi.org/10.1145/2513383.2513447

Rello, L., & Baeza-Yates, R. (2017). How to present more readable text for people with dyslexia. *Universal Access in the Information Society*, *16*(1), 29–49. https://doi.org/10.1007/s10209-015-0438-8

Rello, L., & Bigham, J. P. (2017). Good background colors for readers: A study of people with and without dyslexia. *ASSETS 2017 - Proceedings of the 19th International ACM SIGACCESS Conference on Computers and Accessibility*, 72–80. https://doi.org/10.1145/3132525.3132546

Richlan, F., Kronbichler, M., & Wimmer, H. (2009). Functional abnormalities in the dyslexic brain: A quantitative meta-analysis of neuroimaging studies. *Human Brain Mapping*, *30*(10), 3299. https://doi.org/10.1002/HBM.20752

Rickerson, N. (2009). Universal Design: Principles and Practice for People with Disabilities. In *International Handbook of Occupational Therapy Interventions* (pp. 159–165). Springer New York. https://doi.org/10.1007/978-0-387-75424-6_14

Rixom, J. M., Jackson, M., & Rixom, B. A. (2022). Mandating diversity on the board of directors: Do investors feel that gender quotas result in tokenism or added value for firms? *Journal of Business Ethics*, *1*, 1–19. https://doi.org/10.1007/s10551-021-05030-9

Roberts, B. W., Kuncel, N. R., Shiner, R., Caspi, A., & Goldberg, L. R. (2007). The Power of Personality: The comparative validity of personality traits, socioeconomic status, and cognitive ability for predicting important life outcomes. *Perspectives on Psychological Science*, *2*(4), 313–345. https://doi.org/10.1111/j.1745-6916.2007.00047.x

Roberts, K., DeQuinzio, J. A., Taylor, B. A., & Petroski, J. (2021). Using behavioral skills training to teach interview skills to young adults with autism. *Journal of Behavioral Education*, *30*(4), 664–683. https://doi.org/10.1007/s10864-020-09389-z

Romualdez, A. M., Heasman, B., Walker, Z., Davies, J., & Remington, A. (2021). "People might understand me better": Diagnostic disclosure experiences of autistic individuals in the workplace. *Autism in Adulthood*, *3*(2), 157–167. https://doi.org/10.1089/aut.2020.0063

Romualdez, A. M., Walker, Z., & Remington, A. (2021). Autistic adults' experiences of diagnostic disclosure in the workplace: Decision-making and factors associated with outcomes. *Autism & Developmental Language Impairments*, *6*, 239694152110229. https://doi.org/10.1177/23969415211022955

Rong, Y., Yang, C. J., Jin, Y., & Wang, Y. (2021). Prevalence of attention-deficit/hyperactivity disorder in individuals with autism spectrum disorder: A meta-analysis. *Research in Autism Spectrum Disorders*, *83*, 101759. https://doi.org/10.1016/J.RASD.2021.101759

Rosales, R., & Whitlow, H. (2019). A component analysis of job interview training for young adults with autism spectrum disorder. *Behavioral Interventions*, *34*(2), 147–162. https://doi.org/10.1002/BIN.1658

Rosenbusch, H., Soldner, F., Evans, A. M., & Zeelenberg, M. (2021). Supervised machine learning methods in psychology: A practical introduction with annotated R code. *Social and Personality Psychology Compass*, *15*(2). https://doi.org/10.1111/spc3.12579

Rothmann, S., & Coetzer, E. P. (2003). The big five personality dimensions and job performance. *SA Journal of Industrial Psychology*, *29*(1), 68–74. https://doi.org/10.4102/sajip.v29i1.88

Rottman, C., Gardner, C., Liff, J., Mondragon, N., & Zuloaga, L. (2023). New strategies for addressing the diversity–validity dilemma with big data. *Journal of Applied Psychology*. https://doi.org/10.1037/APL0001084

Russell, G., Stapley, S., Newlove-Delgado, T., Salmon, A., White, R., Warren, F., Pearson, A., & Ford, T. (2022). Time trends in autism diagnosis over 20 years: a UK population-

based cohort study. *Journal of Child Psychology and Psychiatry*, *63*(6), 674–682. https://doi.org/10.1111/JCPP.13505

Ryan, A. M., & Ployhart, R. E. (2014). A Century of selection. *Annual Review of Psychology*, *65*(1), 693–717. https://doi.org/10.1146/annurev-psych-010213-115134

Sackett, P. R. (1987). Assessment centers and content validity: Some neglected issues. *Personnel Psychology*, *40*(1), 13–25. https://doi.org/10.1111/j.1744-6570.1987.tb02374.x

Sackett, P. R., & Wilk, S. L. (1994). Within-group norming and other forms of score adjustment in preemployment testing. *American Psychologist*, *49*(11), 929–954. https://doi.org/10.1037/0003-066X.49.11.929

Salgado, J. F. (2018). Transforming the area under the normal curve (AUC) into Cohen's d, pearson's rpb, odds-ratio, and natural log odds-ratio: Two conversion tables. *European Journal of Psychology Applied to Legal Context*, *10*(1), 35–47. https://doi.org/10.5093/ejpalc2018a5

Salgado, J. F., & Táuriz, G. (2014). The Five-Factor Model, forced-choice personality inventories and performance: A comprehensive meta-analysis of academic and occupational validity studies. *European Journal of Work and Organizational Psychology*, *23*(1), 3–30. https://doi.org/10.1080/1359432X.2012.716198

Samajima, F. (1994). Estimation of Reliability Coefficients Using the Test Information Function and Its Modifications. *Http://Dx.Doi.Org/10.1177/014662169401800304*, *18*(3), 229–244. https://doi.org/10.1177/014662169401800304

Sarkis, E. (2014). Addressing Attention-Deficit/Hyperactivity Disorder in the Workplace. *Postgraduate Medicine*, *126*(5), 25–30. https://doi.org/10.3810/pgm.2014.09.2797

Sasson, N. J., & Morrison, K. E. (2019). First impressions of adults with autism improve with diagnostic disclosure and increased autism knowledge of peers. *Autism*, *23*(1), 50–59. https://doi.org/10.1177/1362361317729526

Sauter, D. L., & McPeek, D. (1993). Dyslexia in the workplace: Implications of the Americans with disabilities act. *Annals of Dyslexia*, *43*(1), 271–277. https://doi.org/10.1007/BF02928186

Sayal, K., Prasad, V., Daley, D., Ford, T., & Coghill, D. (2018). ADHD in children and young people: prevalence, care pathways, and service provision. *The Lancet Psychiatry*, *5*(2), 175–186. https://doi.org/10.1016/S2215-0366(17)30167-0

Schaffer, C. (1993). Selecting a classification method by cross-validation. *Machine Learning*, *13*(1), 135–143. https://doi.org/10.1007/BF00993106

Schlosser, A. E. (2020). Self-disclosure versus self-presentation on social media. *Current Opinion in Psychology*, *31*, 1–6. https://doi.org/10.1016/J.COPSYC.2019.06.025

Schmidt, F. L., & Hunter, J. E. (1998). The validity and utility of selection methods in personnel psychology: Practical and theoretical implications of 85 years of research findings. *Psychological Bulletin*, *124*(2), 262–274. https://doi.org/10.1037/0033-2909.124.2.262

Schmidt, F. L., Oh, I.-S., & Shaffer, J. A. (2016a). *The validity and utility of selection methods in personnel psychology: Practical and theoretical Implications of 100 Years*. https://doi.org/10.13140/RG.2.2.18843.26400

Schmidt, F. L., Oh, I.-S., & Shaffer, J. A. (2016b). *The validity and utility of selection methods in personnel psychology: Practical and theoretical Implications of 100 Years*. https://doi.org/10.13140/RG.2.2.18843.26400

Schmitt, D. P., Allik, J., McCrae, R. R., & Benet-Martínez, V. (2007). The Geographic Distribution of Big Five Personality Traits. *Journal of Cross-Cultural Psychology*, *38*(2), 173–212. https://doi.org/10.1177/0022022106297299

Schmitt, N. (2014). Personality and cognitive ability as predictors of effective performance at work. *Annual Review of Organizational Psychology and Organizational Behavior*, *1*(1), 45–65. https://doi.org/10.1146/annurev-orgpsych-031413-091255

Schur, L., Nishii, L., Adya, M., Kruse, D., Bruyère, S. M., & Blanck, P. (2014). Accommodating Employees With and Without Disabilities. *Human Resource Management*, *53*(4), 593–621. https://doi.org/10.1002/HRM.21607

Schwartz, A. H., Eichstaedt, J. C., Dziurzynski, L., Kern, M. L., Seligman, M. E. P., Ungar, L. H., Blanco, E., Kosinski, M., & Stillwell, D. (2013). Toward personality insights from language exploration in social media. *AAAI Spring Symposium Series*, 72–79. http://sentiment.christopherpotts.net/code-data/

Schwartz, R., Vassilev, A., Greene, K., Perine, L., Burt, A., & Hall, P. (2022). Towards a Standard for Identifying and Managing Bias in Artificial Intelligence. In *NIST Special Publication*. https://doi.org/10.6028/NIST.SP.1270

Sedgwick, J. A., Merwood, A., & Asherson, P. (2019). The positive aspects of attention deficit hyperactivity disorder: a qualitative investigation of successful adults with ADHD. *ADHD Attention Deficit and Hyperactivity Disorders*, *11*(3), 241–253. https://doi.org/10.1007/s12402-018-0277-6

Sharp, S. I., McQuillin, A., & Gurling, H. M. D. (2009). Genetics of attention-deficit hyperactivity disorder (ADHD). *Neuropharmacology*, *57*(7–8), 590–600. https://doi.org/10.1016/J.NEUROPHARM.2009.08.011

Shaywitz, S. E., Mody, M., & Shaywitz, B. A. (2006). Neural Mechanisms in Dyslexia. *Current Directions in Psychological Science*, *15*(6), 278–281. https://doi.org/10.1111/j.1467-8721.2006.00452.x

Shimao, H., Khern-am-nuai, W., & Kannan, K. N. (2021). Addressing fairness in machine learning predictions: Strategic best-response fair discriminant removed algorithm. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.3389631

Siau, K., & Wang, W. (2020). Artificial intelligence (AI) ethics: Ethics of AI and ethical AI. *Journal of Database Management*, *31*(2), 74–87. https://doi.org/10.4018/JDM.2020040105

Sigi Hale, T., Bookheimer, S., McGough, J. J., Phillips, J. M., & McCracken, J. T. (2007). Atypical brain activation during simple & complex levels of processing in adult ADHD:

An fMRI study. *Journal of Attention Disorders*, *11*(2), 125–139. https://doi.org/10.1177/1087054706294101

Skewes, J. C., Jegindø, E.-M., & Gebauer, L. (2015). Perceptual inference and autistic traits. *Autism*, *19*(3), 301–307. https://doi.org/10.1177/1362361313519872

Smith, M. J., Sherwood, K., Ross, B., Smith, J. D., DaWalt, L., Bishop, L., Humm, L., Elkins, J., & Steacy, C. (2021). Virtual interview training for autistic transition age youth: A randomized controlled feasibility and effectiveness trial. *Autism*, *25*(6), 1536–1552. https://doi.org/10.1177/1362361321989928

Smits, J., & Charlier, N. (2011). Game-based assessment and the effect on test anxiety: A case study. *Proceedings of the European Conference on Games-Based Learning*, 562–566.

Smyth, J. D., Dillman, D. A., Christian, L. M., & Stern, M. J. (2006). Comparing check-all and forced-choice question formats in Web surveys. *Public Opinion Quarterly*, *70*(1), 66–77. https://doi.org/10.1093/poq/nfj007

Society for Industrial and Organizational Psychology. (2018). *Principles for the validation and use of personnel selection procedures* (5th ed.). https://doi.org/10.1017/iop.2018.195

Society for Industrial and Organizational Psychology. (2022). *SIOP statement on the use of artificial intelligence (AI) for hiring: Guidance on the effective use of AI-based assessments*.

Society for Industrial and Organizational Psychology. (2023). *Considerations and recommendations for the validation and use of AI-based assessments for employee selection*. https://www.siop.org/Portals/84/SIOP-AI Guidelines-Final-010323.pdf

Soldz, S., & Vaillant, G. E. (1999). The big five personality traits and the life course: A 45-year longitudinal study. *Journal of Research in Personality*, *33*(2), 208–232. https://doi.org/10.1006/jrpe.1999.2243

Solmi, M., Song, M., Yon, D. K., Lee, S. W., Fombonne, E., Kim, M. S., Park, S., Lee, M. H., Hwang, J., Keller, R., Koyanagi, A., Jacob, L., Dragioti, E., Smith, L., Correll, C. U., Fusar-Poli, P., Croatto, G., Carvalho, A. F., Oh, J. W., … Cortese, S. (2022). Incidence,

prevalence, and global burden of autism spectrum disorder from 1990 to 2019 across 204 countries. *Molecular Psychiatry 2022 27:10*, *27*(10), 4172–4180. https://doi.org/10.1038/s41380-022-01630-7

Speer, A. B., & Delacruz, A. Y. (2021). Introducing a supervised alternative to forced-choice personality scoring: A test of validity and resistance to faking. *International Journal of Selection and Assessment*, *29*(3–4), 448–466. https://doi.org/10.1111/IJSA.12345

Stark, E., Stacey, J., Mandy, W., Kringelbach, M. L., & Happé, F. (2021). Autistic cognition: Charting routes to anxiety. *Trends in Cognitive Sciences*, *25*(7), 571–581. https://doi.org/10.1016/J.TICS.2021.03.014

Stark, R., Bauer, E., Merz, C. J., Zimmermann, M., Reuter, M., Plichta, M. M., Kirsch, P., Lesch, K. P., Fallgatter, A. J., Vaitl, D., & Herrmann, M. J. (2011). ADHD related behaviors are associated with brain activation in the reward system. *Neuropsychologia*, *49*(3), 426–434. https://doi.org/10.1016/J.NEUROPSYCHOLOGIA.2010.12.012

Stark, S., Chernyshenko, O. S., & Drasgow, F. (2004). Examining the effects of differential item (functioning and differential) test functioning on selection decisions: When are statistically significant effects practically important? *Journal of Applied Psychology*, *89*(3), 497–508. https://doi.org/10.1037/0021-9010.89.3.497

Steele, L. M., Pindek, S., & Margalit, O. (2021). Associated with Idea Generation at Work? *Creativity Research Journal*, *33*(3), 275–283. https://doi.org/10.1080/10400419.2021.1916368

Stevenor, B. A., Hickman, L., Zickar, M. J., Wimbush, F., & Beck, W. (2024). Validity evidence for personality scores from algorithms trained on low-stakes verbal data and applied to high-stakes interviews. *International Journal of Selection and Assessment*. https://doi.org/10.1111/IJSA.12480

Stobart, G., & Eggen, T. (2012). High-stakes testing – value, fairness and consequences. *Assessment in Education: Principles, Policy & Practice*, *19*(1), 1–6. https://doi.org/10.1080/0969594X.2012.639191

Stocco, C. S., Thompson, R. H., Hart, J. M., & Soriano, H. L. (2017). Improving the interview skills of college students using behavioral skills training. *Journal of Applied Behavior Analysis*, *50*(3), 495–510. https://doi.org/10.1002/JABA.385

Strazzulla, P. (2020). *524% rise in video interview programs as businesses adapt to COVID-19*. https://www.selectsoftwarereviews.com/blog/video-interview-software-interest-covid-19

Suen, H. Y., Chen, M. Y. C., & Lu, S. H. (2019). Does the use of synchrony and artificial intelligence in video interviews affect interview ratings and applicant attitudes? *Computers in Human Behavior*, *98*, 93–101. https://doi.org/10.1016/J.CHB.2019.04.012

Suen, H.-Y., Hung, K.-E., & Lin, C.-L. (2019). TensorFlow-based automatic personality recognition used in asynchronous video interviews. *IEEE Access*, *7*, 61018–61023. https://doi.org/10.1109/ACCESS.2019.2902863

Taber, K. S. (2018). The use of Cronbach's alpha when developing and reporting research instruments in science education. *Research in Science Education*, *48*(6), 1273–1296. https://doi.org/10.1007/s11165-016-9602-2

Tamboer, P., Vorst, H. C. M., Ghebreab, S., & Scholte, H. S. (2016). Machine learning and dyslexia: Classification of individual structural neuro-imaging scans of students with and without dyslexia. *NeuroImage: Clinical*, *11*, 508–514. https://doi.org/10.1016/J.NICL.2016.03.014

Tanwar, M., Duggal, R., & Khatri, S. K. (2015). Unravelling unstructured data: A wealth of information in big data. *2015 4th International Conference on Reliability, Infocom Technologies and Optimization: Trends and Future Directions, ICRITO 2015*, 1–6. https://doi.org/10.1109/ICRITO.2015.7359270

Tatman, R. (2017). Gender and dialect bias in YouTube's automatic captions. *Proceedings of the First ACL Workshop on Ethics in Natural Language Processing*, 53–59. https://doi.org/10.18653/v1/W17-1606

Tatman, R., & Kasten, C. (2017). Effects of talker dialect, gender & race on accuracy of bing speech and youtube automatic captions. *Proceedings of the Annual Conference of the*

International Speech Communication Association, INTERSPEECH, *2017-Augus*, 934–938. https://doi.org/10.21437/Interspeech.2017-1746

Tay, L., Woo, S. E., Hickman, L., Booth, B. M., & D'Mello, S. (2022). A conceptual framework for investigating and mitigating machine-learning measurement bias (MLMB) in psychological assessment. *Advances in Methods and Practices in Psychological Science*, *5*(1), 1–30. https://doi.org/10.1177/25152459211061337

Tene, O., & Polonetsky, J. (2013). A theory of creepy: Technology, privacy and shifting social norms. *Yale Journal of Law & Technology*, *16*(1), 1–32. https://heinonline.org/HOL/P?h=hein.journals/yjolt16&i=59

Thapar, A. (2018). Discoveries on the genetics of ADHD in the 21st century: New findings and their implications. *American Journal of Psychiatry*, *175*(10), 943–950. https://doi.org/10.1176/appi.ajp.2018.18040383

Thapar, A., & Rutter, M. (2021). Genetic advances in autism. *Journal of Autism and Developmental Disorders*, *51*(12), 4321–4332. https://doi.org/10.1007/s10803-020-04685-z

The Center for Universal Design. (1997). *The principles of universal design version 2.0*. https://design.ncsu.edu/wp-content/uploads/2022/11/principles-of-universal-design.pdf

The New York City Council. (2021). *Int 1894-2020 - Local Law 144*. https://legistar.council.nyc.gov/LegislationDetail.aspx?ID=4344524&GUID=B051915D-A9AC-451E-81F8-6596032FA3F9&Options=Advanced&Search

The Verge. (2018). *Google 'fixed' its racist algorithm by removing gorillas from its image-labeling tech*. https://www.theverge.com/2018/1/12/16882408/google-racist-gorillas-photo-recognition-algorithm-ai

Thomas, E., Molebogeng Diale, B., & Victor-Aigbodion, V. (2022). Workplace experiences of youth diagnosed with attention deficit hyperactivity disorder (ADHD). *Journal of Positive School Psychology*, *2022*(6), 1306–1317. https://www.journalppw.com/index.php/jpsp/article/view/7293

Thornton, D., & Kline, P. (1982). Reliability and validity of the Belief in Human Benevolence scale. *British Journal of Social Psychology*, *21*(1), 57–62. https://doi.org/10.1111/J.2044-8309.1982.TB00513.X

Thurstone, L. L. (1994). A law of comparative judgment. *Psychological Review*, *101*(2), 266–270. https://doi.org/10.1037/0033-295X.101.2.266

Tibshirani, R. (1996). Regression shrinkage and selection via the Lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, *58*(1), 267–288. https://doi.org/10.1111/j.2517-6161.1996.tb02080.x

Tilston, O., Krings, F., Roulin, N., Bourdage, J. S., & Fetzer, M. (2024). Reactions to asynchronous video interviews: The role of design decisions and applicant age and gender. *Human Resource Management*, *63*(2), 313–332. https://doi.org/10.1002/HRM.22202

Tippins, N. (2009). Adverse impact in employee selection procedures from the perspective of an organizational consultant. In *Adverse impact: Implications for organizational staffing and high stakes selection* (pp. 201–225). https://doi.org/10.4324/9780203848418

Tippins, N., Beaty, J., Drasgow, F., Gibson, W. M., Pearlman, K., Ssgall, D. O., & Shepherd, W. (2006). Unproctored internet testing in employment settings. *Personnel Psychology*, *59*(1), 189–225. https://doi.org/10.1111/j.1744-6570.2006.00909.x

Tippins, N., Oswald, F., & McPhail, S. M. (2021). Scientific, Legal, and Ethical Concerns About AI-Based Personnel Selection Tools: A Call to Action. *Personnel Assessment and Decisions*, *7*(2). https://doi.org/10.25035/pad.2021.02.001

Tops, W., Verguts, E., Callens, M., & Brysbaert, M. (2013). Do Students with Dyslexia Have a Different Personality Profile as Measured with the Big Five? *PLOS ONE*, *8*(5), e64484. https://doi.org/10.1371/JOURNAL.PONE.0064484

Tsetsi, E., & Rains, S. A. (2017). Smartphone Internet access and use: Extending the digital divide and usage gap. *Mobile Media and Communication*, *5*(3), 239–255. https://doi.org/10.1177/2050157917708329

Vabalas, A., Gowen, E., Poliakoff, E., & Casson, A. J. (2019). Machine learning algorithm validation with a limited sample size. *PLOS ONE*, *14*(11), e0224365. https://doi.org/10.1371/journal.pone.0224365

Valentino, N. A., Zhirkov, K., Hillygus, D. S., & Guay, B. (2021). The Consequences of Personality Biases in Online Panels for Measuring Public Opinion. *Public Opinion Quarterly*, *84*(2), 446–468. https://doi.org/10.1093/POQ/NFAA026

van de Mortel, T. F. (2008). Faking it: Social desirability response bias in selfreport research. *Australian Journal of Advanced Nursing*, *25*(4), 40–48. https://researchportal.scu.edu.au/discovery/delivery/61SCU_INST:ResearchRepository/1267228250002368?l#1367368500002368

Vandenberg, R. J., & Lance, C. E. (2000). *A review and synthesis of the measurement invariance literature: Suggestions, practices, and recommendations for organizational research*.

Ventura, M., & Shute, V. (2013). The validity of a game-based assessment of persistence. *Computers in Human Behavior*, *29*(6), 2568–2572. https://doi.org/10.1016/J.CHB.2013.06.033

Verheul, I., Rietdijk, W., Block, J., Franken, I., Larsson, H., & Thurik, R. (2016). The association between attention-deficit/hyperactivity (ADHD) symptoms and self-employment. *European Journal of Epidemiology*, *31*(8), 793–801. https://doi.org/10.1007/s10654-016-0159-1

Verma, S., Ernst, M., & Just, R. (2021). Removing biased data to improve fairness and accuracy. *CEUR Workshop Proceedings*, *2657*, 1–9.

Verma, S., & Rubin, J. (2018). Fairness definitions explained. *Proceedings - International Conference on Software Engineering*, *18*, 1–7. https://doi.org/10.1145/3194770.3194776

Vincent, J., & Fabri, M. (2022). The ecosystem of competitive employment for university graduates with autism. *International Journal of Disability, Development and Education*, *69*(5), 1823–1839. https://doi.org/10.1080/1034912X.2020.1821874

Wachter, S., Mittelstadt, B., & Russell, C. (2021). Why fairness cannot be automated: Bridging the gap between EU non-discrimination law and AI. *Computer Law & Security Review*, *41*, 105567. https://doi.org/10.1016/J.CLSR.2021.105567

Wagner, R. K., Edwards, A. A., Malkowski, A., Schatschneider, C., Joyner, R. E., Wood, S., & Zirps, F. A. (2019). Combining old and new for better understanding and predicting dyslexia. *New Directions for Child and Adolescent Development*, *2019*(165), 11–23. https://doi.org/10.1002/CAD.20289

Waisman-Nitzan, M., Gal, E., & Schreuer, N. (2021). "It's like a ramp for a person in a wheelchair": Workplace accessibility for employees with autism. *Research in Developmental Disabilities*, *114*, 103959. https://doi.org/10.1016/J.RIDD.2021.103959

Waldren, L. H., Clutterbuck, R. A., & Shah, P. (2021). Erroneous NICE guidance on autism screening. *The Lancet Psychiatry*, *8*(4), 276–277. https://doi.org/10.1016/S2215-0366(21)00065-1

Waldren, L. H., Livingston, L. A., Clutterbuck, R. A., Callan, M. J., Walton, E., & Shah, P. (2022). Using incorrect cut-off values in autism screening tools: The consequences for psychological science. *PsyArXiv*. https://doi.org/10.31234/osf.io/x4h7n

Wall, S., & Schellmann, H. (2021). *LinkedIn's job-matching AI was biased. The company's solution? More AI*. https://www.technologyreview.com/2021/06/23/1026825/linkedin-ai-bias-ziprecruiter-monster-artificial-intelligence/

Wang, R., Harper, F. M., & Zhu, H. (2020). Factors Influencing Perceived Fairness in Algorithmic Decision-Making: Algorithm Outcomes, Development Procedures, and Individual Differences. *Conference on Human Factors in Computing Systems - Proceedings*, 1–14. https://doi.org/10.1145/3313831.3376813

Wassink, A. B., Gansen, C., & Bartholomew, I. (2022). Uneven success: automatic speech recognition and ethnicity-related dialects. *Speech Communication*, *140*, 50–70. https://doi.org/10.1016/J.SPECOM.2022.03.009

Weber, C., Krieger, B., Häne, E., Yarker, J., & McDowall, A. (2022). Physical workplace adjustments to support neurodivergent workers: A systematic review. *Applied Psychology*. https://doi.org/10.1111/APPS.12431

Wehman, P., Brooke, V., Brooke, A. M., Ham, W., Schall, C., McDonough, J., Lau, S., Seward, H., & Avellone, L. (2016). Employment for adults with autism spectrum disorders: A retrospective review of a customized employment approach. *Research in Developmental Disabilities*, *53–54*, 61–72. https://doi.org/10.1016/J.RIDD.2016.01.015

Weinberg, A., & Doyle, N. (2017). Focus on strengths: Supporting people who experience difficulties at work. In K. Scott & L. M. Coulthard (Eds.), *Psychology at work: Improving wellbeing and productivity in the workplace* (pp. 43–62). British Psychological Society.

Weiss, B., & Feldman, R. S. (2006). Looking Good and Lying to Do It: Deception as an Impression Management Strategy in Job Interviews. *Journal of Applied Social Psychology*, *36*(4), 1070–1086. https://doi.org/10.1111/J.0021-9029.2006.00055.X

West, S. G., Finch, J. F., & Curran, P. J. (1995). Structural equation models with nonnormal variables: Problems and remedies. In R. H. Hoyle (Ed.), *Structural equation modeling: Concepts, issues, and applications* (pp. 56–75). SAGE Publications.

Willis, C., Powell-Rudy, T., Colley, K., & Prasad, J. (2021). Examining the use of game-based assessments for hiring autistic job seekers. *Journal of Intelligence*, *9*(4), 53. https://doi.org/10.3390/jintelligence9040053

Willmott, C. J., & Matsuura, K. (2005). Advantages of the mean absolute error (MAE) over the root mean square error (RMSE) in assessing average model performance. *Climate Research*, *30*(1), 79–82. https://doi.org/10.3354/CR030079

Wilson, A. C., & Bishop, D. V. M. (2021). “Second guessing yourself all the time about what they really mean…”: Cognitive differences between autistic and non-autistic adults in understanding implied meaning. *Autism Research*, *14*(1), 93–101. https://doi.org/10.1002/AUR.2345

Wilson, C., Ghosh, A., Jiang, S., Mislove, A., Baker, L., Szary, J., Trindel, K., & Polli, F. (2021). Building and auditing fair algorithms: A case study in candidate screening. *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, 666–677. https://doi.org/10.1145/3442188.3445928

Winsborough, D., & Chamorro-Premuzic, T. (2016). Talent identification in the digital world: New talent signals and the future of HR assessment. *People and Strategy*, *39*(2), 28. https://info.hoganassessments.com/hubfs/TalentIdentification.pdf

Woo, E. (2019). *Autism at Work: Encouraging Neurodiversity in the Workplace*. https://news.sap.com/2019/10/workplace-neurodiversity-autism-at-work-program/

Woodworth, B., Gunasekar, S., Ohannessian, M. I., & Srebro, N. (2017). Learning Non-Discriminatory Predictors. *Proceedings of Machine Learning Research*, *65*, 1–34.

Wu, F. Y., Mulfinger, E., Alexander, L., Sinclair, A. L., McCloy, R. A., & Oswald, F. L. (2022). Individual differences at play: An investigation into measuring Big Five personality facets with game-based assessments. *International Journal of Selection and Assessment*, *30*(1), 62–81. https://doi.org/10.1111/IJSA.12360

Wu, H., Albiero, V., Krishnapriya, K. S., King, M. C., & Bowyer, K. W. (2023). Face recognition accuracy across demographics: Shining a light into the problem. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops,* 1041–1050. https://openaccess.thecvf.com/content/CVPR2023W/Biometrics/html/Wu_Face_Recognition_Accuracy_Across_Demographics_Shining_a_Light_Into_the_CVPRW_2023_paper.html

Xu, Y., & Goodacre, R. (2018). On Splitting Training and Validation Set: A Comparative Study of Cross-Validation, Bootstrap and Systematic Sampling for Estimating the Generalization Performance of Supervised Learning. *Journal of Analysis and Testing*, *2*, 249–262. https://doi.org/10.1007/s41664-018-0068-2

Yan, T., Conrad, F. G., Tourangeau, R., & Couper, M. P. (2011). Should I stay or should I go: The effects of progress feedback, promised task duration, and length of questionnaire on completing web surveys. *International Journal of Public Opinion Research*, *23*(2), 131–147. https://doi.org/10.1093/ijpor/edq046

Yarkoni, T., & Westfall, J. (2017). Choosing prediction over explanation in psychology: Lessons from machine learning. *Perspectives on Psychological Science*, *12*(6), 1100–1122. https://doi.org/10.1177/1745691617693393

Zemel, R., Wu, Y., Swersky, K., Pitassi, T., & Dwork, C. (2013). Learning fair representations. *30th International Conference on Machine Learning, ICML 2013*, *PART 2*, 1362–1370.

Zhang, B., Sun, T., Drasgow, F., Chernyshenko, O. S., Nye, C. D., Stark, S., & White, L. A. (2020). Though forced, still valid: Psychometric equivalence of forced-choice and single-statement measures. *Organizational Research Methods*, *23*(3), 569–590. https://doi.org/10.1177/1094428119836486

Zhang, F., & Roeyers, H. (2019). Exploring brain functions in autism spectrum disorder: A systematic review on functional near-infrared spectroscopy (fNIRS) studies. *International Journal of Psychophysiology*, *137*, 41–53. https://doi.org/10.1016/J.IJPSYCHO.2019.01.003

Zhang, H., Zhang, J., Sang, J., & Xu, C. (2017). A demo for image-based personality test. *Lecture Notes in Computer Science: MultiMedia Modelling*, *10133*, 433–437. https://doi.org/10.1007/978-3-319-51814-5_36

Zhu, Y., Liu, W., Li, Y., Wang, X., & Winterstein, A. G. (2018). Prevalence of ADHD in publicly insured adults. *Journal of Attention Disorders*, *22*(2), 182–190. https://doi.org/10.1177/1087054717698815

Zolnoori, M., Vergez, S., Xu, Z., Esmaeili, E., Zolnour, A., Anne Briggs, K., Scroggins, J. K., Hosseini Ebrahimabad, S. F., Noble, J. M., Topaz, M., Bakken, S., Bowles, K. H., Spens, I., Onorato, N., Sridharan, S., & McDonald, M. V. (2024). Decoding disparities: evaluating automatic speech recognition system performance in transcribing Black and White patient verbal communication with nurses in home healthcare. *JAMIA Open*, *7*(4). https://doi.org/10.1093/jamiaopen/ooae130

Zuloaga, L. (2021). *Industry leadership: New audit results and decision on visual analysis*. https://www.hirevue.com/blog/hiring/industry-leadership-new-audit-results-and-decision-on-visual-analysis

Chapter 9. Appendices

Appendix A
Worked bias mitigation example for conscientiousness (Chapter 2)

The Google Colab version of this worked example can be seen here.

Bias mitigation worked example: Conscientiousness

This worked example supports Chapter 2 by demonstrating how three machine learning approaches to bias mitigation work with data from the validation of an image-based assessment of personality designed to be used in recruitment, which is described in greater detail in Chapter 3. Specifically, this applies Learning fair representations (Zemel et al., 2013); Prejudice remover regularizer (Kamishima et al., 2012); and Equalised odds (Hardt et al., 2016) using the Holistic AI library to implement the mitigations.

The assessment, described in greater detail in Chapter 3, presents candidates with 150 pairs of images designed to measure the Big Five and asks them to select the image in the pair that is most like them. The 150 image pairs are used to create 300 dummy variables, which are used to create a simple logistic regression that predicts images in the assessment as predictors in the model to predict personality scores on the IPIP. Given that the training data demonstrates group differences for mixed ethnicity test-takers coompared to black females for conscientiousness, this example focuses on mitigating the adverse impact against females for emotional stability in the predicted scores.

Since the Big Five traits are typically measured continuously and the bias mitigations, like many in computer science, are designed to be used with classification systems, the training data (IPIP Scores) was binarised using the median emotional stability score for the dataset as a threshold in line with the metric that must be used to calculate impact ratios for continuous systems under Local Law 144. The following tutorial builds a baseline classification algorithm before applying the mitigation strategies and demonstrating how they change the data, coefficients, and outputs.

For each mitigation approach, the tutorial shows how it changes the data, coefficients, or outputs. It concludes with a comparison of the effectiveness of each mitigation measure, finding that for this example, the training constraints approach is the most effective approach and effectively mitigated the violations to the adverse impact metrics that were present in the training data and baseline model.

In [1]:
```
#install the Holistic AI library
#!pip install holisticai
!pip install git+https://github.com/holistic-ai/holisticai.git@holisticai-v1
```

```
Installing collected packages: pybind11, holisticai
Successfully installed holisticai-1.0.0 pybind11-2.13.5
```

0. Import data

```
#import libraries
import pandas as pd
import numpy as np
import scipy as scipy
import matplotlib.pyplot as plt
from sklearn.linear_model import Lasso
from sklearn.model_selection import GridSearchCV
from sklearn.linear_model import Ridge
from sklearn.model_selection import train_test_split
from sklearn.metrics import r2_score
import warnings
warnings.filterwarnings("ignore")
```

```
#import the file from Google Drive
!gdown 1jBGZ2GZnVcTyFdkMRoVSP9S5yBoN6SYm
```

```
Downloading...
From: https://drive.google.com/uc?id=1jBGZ2GZnVcTyFdkMRoVSP9S5yBoN6SYm
```

```
To: /content/demo_data.csv
100% 218k/218k [00:00<00:00, 5.38MB/s]
```

In [4]:
```
#load the data and replace missing values with 0
df=pd.read_csv('demo_data.csv')
df = df.replace(np.nan, 0)

#print the first few rows to inspect the data
df.head()
```

Out[4]:

	Age	Gender	Ethnicity	O	C	E	A	ES	N	s_16_2	...	e_20_2	c_39_2	all_25_e	words_7_e	e_
0	Under 40 years old	Male	Asian	83	92	62	87	68	76	0.0	...	1.0	0.0	1.0	0.0	
1	Under 40 years old	Female	Black	91	88	67	80	77	67	0.0	...	1.0	1.0	1.0	0.0	
2	Under 40 years old	Female	Asian	81	107	92	104	102	42	0.0	...	1.0	1.0	0.0	0.0	
3	Under 40 years old	Male	Asian	94	111	69	118	93	51	0.0	...	1.0	1.0	0.0	0.0	
4	Under 40 years old	Female	Asian	93	95	91	102	87	57	1.0	...	0.0	1.0	1.0	1.0	

5 rows × 309 columns

1. Calculate success rates and adverse impact metrics for training data

Based on the adverse impact analysis from Chapter 3, black test-takers have a higher pass rate than mixed ethnicity test-takers in the training data and results in an adverse impact ratio of .36. Since the four-fifths rule as well as Cohen's D and the 2SD rule were violated for this group, it will serve as a test of the effectiveness of the three bias migiation approaches.

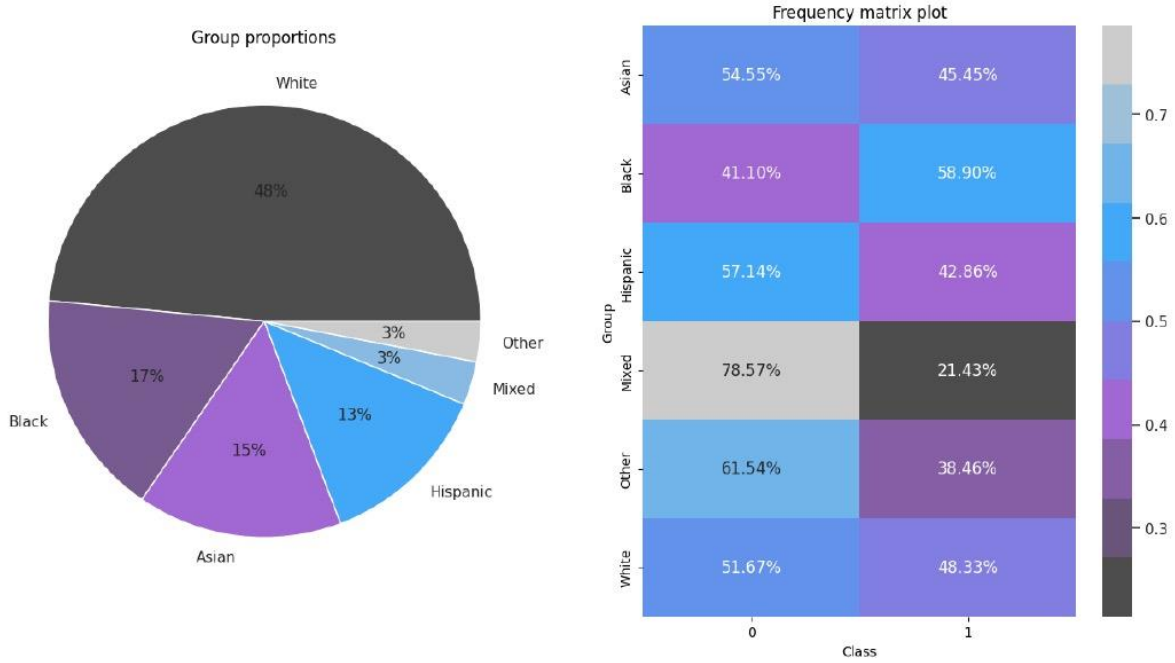
In [5]:
```
#binarise scores based on median
C_median=df['C'].median()
df['C_binarised'] = np.where(df['C'] > C_median, 1, 0)
```

In [6]:
```
#define functions to evaluate model performance

# efficacy metrics from sklearn
from sklearn import metrics
import pandas as pd

# dictionnary of metrics
metrics_dict={
        "Accuracy": metrics.accuracy_score,
        "Balanced accuracy": metrics.balanced_accuracy_score,
        "Precision": metrics.precision_score,
```

```
        "Recall": metrics.recall_score,
        "F1-Score": metrics.f1_score}

# efficacy metrics dataframe helper tool
def metrics_dataframe(y_pred, y_true, metrics_dict=metrics_dict):
    metric_list = [[pf, fn(y_true, y_pred)] for pf, fn in metrics_dict.items()]
    return pd.DataFrame(metric_list, columns=["Metric", "Value"]).set_index("Metric")
```

In [7]:
```
#create a figure showing the proportion of each ethnicity
from holisticai.bias.plots import group_pie_plot,frequency_matrix_plot
fig,axs=plt.subplots(1,2,figsize=(16,8))
group_pie_plot(df['Ethnicity'], ax=axs[0])

#create a figure showing the proportion of each ethnicity passing (scoring above the med
frequency_matrix_plot(df['Ethnicity'], df['C_binarised'], ax=axs[1], normalize='group')
```

Out[7]:
```
<Axes: title={'center': 'Frequency matrix plot'}, xlabel='Class', ylabel='Group'>
```



In [8]:
```
#set up data for model training

#X will be predictors and y is the target variable (binarised C scores)

#drop all data other than predictors from the dataframe
X=df.drop(['Age', 'Gender', 'Ethnicity', 'O', 'C', 'E', 'A',
           'ES', 'N', 'C_binarised'], axis=1)

#isolate the target variable
y=df['C_binarised']
```

In [9]:
```
#get success rate (proportion of test-takers passing in each group) based on ethnicity
from holisticai.bias.metrics import frequency_matrix
frequency_matrix(df['Ethnicity'], df['C_binarised'])
```

Out[9]:

	0	1
Asian	0.545455	0.454545
Black	0.410959	0.589041
Hispanic	0.571429	0.428571

Mixed	0.785714	0.214286
Other	0.615385	0.384615
White	0.516746	0.483254

In [10]:
```
#calculate bias metrics for the training data

from holisticai.bias.metrics import classification_bias_metrics
group_a = df["Ethnicity"]=='Black'
group_b = df["Ethnicity"]=='Mixed'
y_true  = y.copy()

trainingdata_bias=classification_bias_metrics(group_a, group_b, y)
trainingdata_bias
```

Out[10]:

Metric	Value	Reference
Statistical Parity	0.374755	0
Disparate Impact	2.748858	1
Four Fifths Rule	0.363787	1
Cohen D	0.780065	0
2SD Rule	2.573132	0

Based on the four-fiths rule, Cohen's d , and the 2SD rule, the training data shows evidence of bias since <.80, >|.20|, and >|2|, respectively.

2. Build classification model and examine performance

Given that the bias mitigation approaches were created for binary models, logistic regression would be used to predict levels of emotional stability in terms of high (1 - above the median) or low (0 - below the median).

In [11]:
```
#build a logistic model to predict C from image choices
from sklearn.linear_model import LogisticRegression
model = LogisticRegression(random_state=10)

#first train the model to create the regression line
model.fit(X, y)

#then apply the model to the data to predict the label
ypred = model.predict(X)
```

In [12]:
```
#save the baseline model coefficients and output for later

baseline_coef = list(model.coef_.reshape(-1)) + list(model.intercept_)
df_pred=pd.DataFrame(ypred,columns=['Baseline'])
```

In [13]:
```
#Determine the performance of the baseline model
baseline_metrics=metrics_dataframe(y, ypred)
baseline_metrics
```

Out[13]:

	Value

Metric	
Accuracy	0.916473
Balanced accuracy	0.917045
Precision	0.898058
Recall	0.925000
F1-Score	0.911330

These accuracy metrics range from 0 to 1, where scores closer to 1 indicate that the model performs well. From these values, this simple model performs well.

In [14]:
```
#calculate bias metrics for the predictions
baseline_bias = classification_bias_metrics(group_a, group_b, ypred, y, metric_type='equ
baseline_bias
```

Out[14]:

Metric	Value	Reference
Statistical Parity	0.361057	0
Disparate Impact	2.684932	1
Four Fifths Rule	0.372449	1
Cohen D	0.748464	0
2SD Rule	2.476450	0

As with the training data, the baseline model is biased against females, violating the four-fifths rule, 2 SD rule, and Cohen's *d*.

3. Use pre-processing to mitigate bias (Learning fair representations - Zemel et al., 2013)

In [15]:
```
from holisticai.bias.mitigation import LearningFairRepresentation
from sklearn.preprocessing import StandardScaler

# initialise the data transformation
lfr = LearningFairRepresentation(k=3, verbose=1, Ax=0.01, Ay=1, Az=50)
scaler1 = StandardScaler()
X_t = scaler1.fit_transform(X)
X_t = lfr.fit_transform(X_t, y, group_a, group_b)
```

In [16]:
```
#print the transformed training data
dframe= pd.DataFrame(X_t)
dframe
```

Out[16]:

	0	1	2	3	4	5	6	7	8	9
0	-0.715738	0.447836	0.842803	-0.425232	-0.440347	-0.660808	1.718758	0.708338	-0.947987	-0.712034
1	0.682272	0.412835	0.593774	0.521825	0.282570	0.505541	0.438010	0.610148	0.542691	0.349697
2	-0.715738	0.447836	-1.186516	-0.425232	-0.440347	-0.660808	1.718758	-1.411756	-0.947987	1.404426
3	-0.715738	0.447836	0.842803	-0.425232	-0.440347	-0.660808	1.718758	0.708338	-0.947987	-0.712034
4	1.397160	0.447836	-1.186516	-0.425232	-0.440347	1.513300	1.718758	-1.411756	1.054867	1.404426

...	...	...	...	...	...	...	...	...	...	..
426	-0.715738	0.447836	-1.186516	-0.425232	-0.440347	-0.660808	-0.581816	-1.411756	-0.947987	-0.71203
427	-0.715738	0.447836	0.842803	-0.425232	-0.440347	-0.660808	-0.581816	-1.411756	1.054867	-0.71203
428	0.653326	0.364548	0.588861	0.544387	0.246043	0.456960	0.424298	0.633478	0.549450	0.32645
429	0.699709	0.522176	0.612343	0.497571	0.388862	0.555133	0.405041	0.532836	0.492352	0.41605
430	-0.715738	-2.232960	0.842803	2.351660	2.270934	1.513300	1.718758	-1.411756	1.054867	-0.71203

431 rows × 300 columns

The data above is the training data once it has been transformed by the LearningFairRepresentation mitigation. Instead of having values of 0 or 1, values range from 0 to 1.

In [17]:
```
#train the model on the transformed data
model = LogisticRegression()
model.fit(X_t, y)
y_pred = model.predict(X_t)

#calculate model performance for the new model trained on the transformed data
prepro_metrics=metrics_dataframe(y, y_pred)
prepro_metrics
```

Out[17]:

Metric	Value
Accuracy	0.904872
Balanced accuracy	0.917004
Precision	1.000000
Recall	0.834008
F1-Score	0.909492

In [18]:
```
#calculate bias metrics for the new predictions
from holisticai.bias.metrics import classification_bias_metrics
prepro_bias=classification_bias_metrics(group_a=group_a, group_b=group_b,
                                        y_pred=y_pred, y_true=y, metric_type='equal_outc
prepro_bias
```

Out[18]:

Metric	Value	Reference
Statistical Parity	0.0	0
Disparate Impact	1.0	1
Four Fifths Rule	1.0	1
Cohen D	NaN	0
2SD Rule	NaN	0

Based on these metrics, the model based on the mitigated data performs similarly to th baseline model, with subgroup differences mitigated. However, given that the raw training data is binary (0s and 1s) and the mitigated training data has been transformed into floats, it is unlikely that the training data accurately represents whether images were chosen or not, which reduces the validity of the assessment.

4. Use training constraints to mitigate bias (Prejudice remover regularizer - Kamishima et al., 2012)

In [19]:
```
from holisticai.bias.mitigation import PrejudiceRemover
from holisticai.pipeline import Pipeline
from sklearn.preprocessing import StandardScaler

#note that eta (η) is the regularisation parameter used by this mitigation
# A value of 1000 was selected for eta as the impact ratio plateaued at a value of
#around 1000, as seen in section 4.1

inprocessing = PrejudiceRemover(eta=9e3, init_type='StandarLR')
inprocessing.transform_estimator()

#use the Holistic AI pipline to apply the training constraints
pipeline = Pipeline(steps=[
    ('scaler', StandardScaler()),
    ('bm_inprocessing',inprocessing)])

#apply the pipeline to the raw (unmitigated) training data
pipeline.fit(X, y, bm__group_a=group_a, bm__group_b=group_b)

y_pred = pipeline.predict(X, bm__group_a=group_a, bm__group_b=group_b)
```

In [20]:
```
#evalate model performance
training_metrics=metrics_dataframe(y, y_pred)
training_metrics
```

Out[20]:

Metric	Value
Accuracy	0.851508
Balanced accuracy	0.851811
Precision	0.873786
Recall	0.825688
F1-Score	0.849057

In [21]:
```
#calculate bias metrics
training_bias=classification_bias_metrics(group_a, group_b, y_pred, y)
training_bias
```

Out[21]:

Metric	Value	Reference
Statistical Parity	0.006849	0
Disparate Impact	1.013699	1
Four Fifths Rule	0.986486	1
Cohen D	0.013700	0
2SD Rule	0.046954	0
Equality of Opportunity Difference	0.434109	0
False Positive Rate Difference	-0.412121	0
Average Odds Difference	0.010994	0

Accuracy Difference	0.379648	0

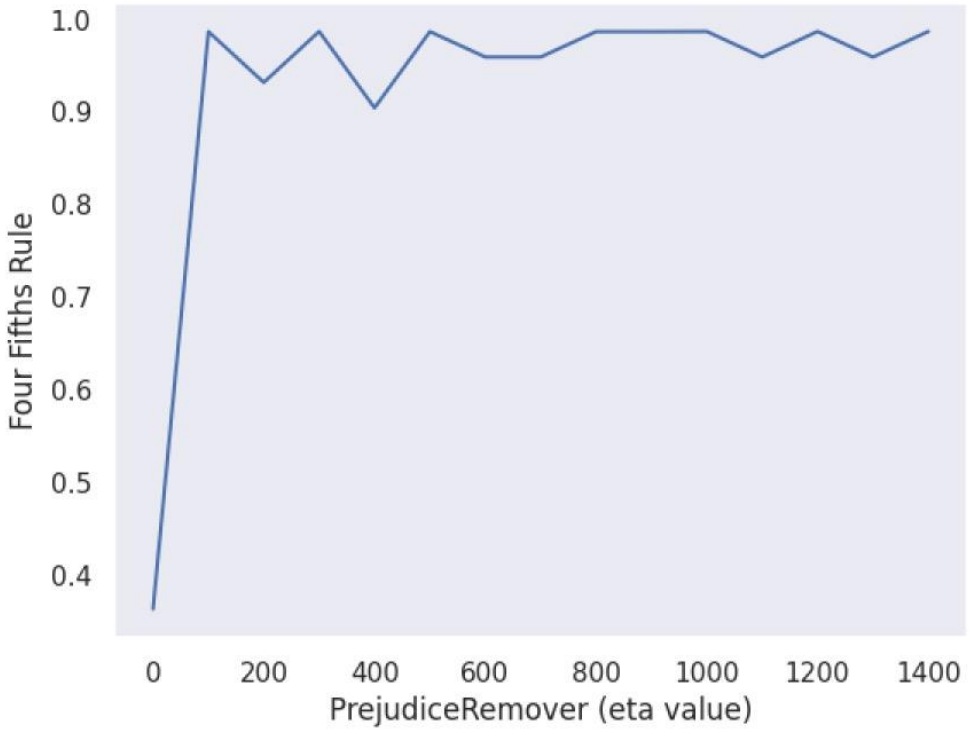
Given the above metrics, the mitigated model performs reasonable well, although not as well as the baseline model but greater than the pre-processing approach. There has also been successful mitigation of subgroup differences since the four-fifths rule is no longer violated.

4.1 Exploring eta parameter

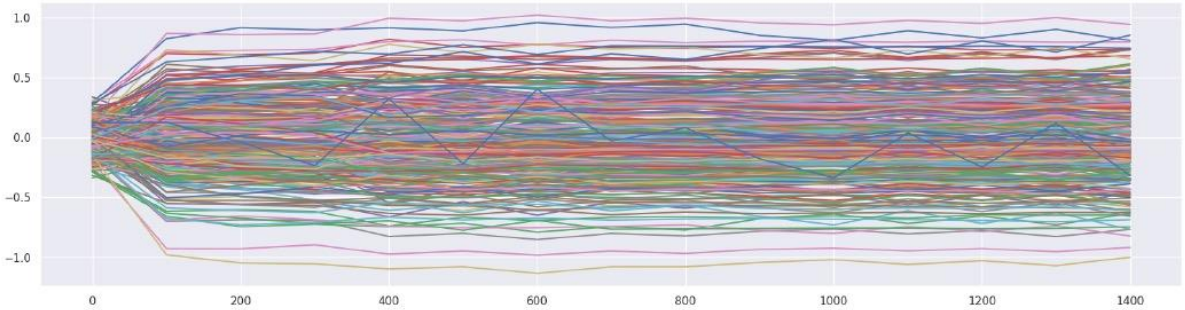
In [22]:
```
from tqdm import tqdm
import numpy as np
import warnings
warnings.filterwarnings("ignore")

etas = np.arange(0, 1500, 100)
metrics = []
coefficients = []
for eta in tqdm(etas):
    inprocessing = PrejudiceRemover(eta=eta, init_type='StandarLR').transform_estimator(
    pipeline = Pipeline(steps=[
        ('scaler', StandardScaler()),
        ('bm_inprocessing',inprocessing)])

    pipeline.fit(X, y, bm__group_a=group_a, bm__group_b=group_b)
    coefficients.append(inprocessing.estimator.coef)
    y_pred = pipeline.predict(X, bm__group_a=group_a, bm__group_b=group_b)

    eff = metrics_dataframe(y, ypred)
    bm = classification_bias_metrics(group_a, group_b, y_pred, y)
    metrics.append({'acc':eff.loc['Accuracy'], 'four_fifths':bm.loc['Four Fifths Rule'].
```

```
100%|██████████| 15/15 [00:12<00:00,  1.16it/s]
```

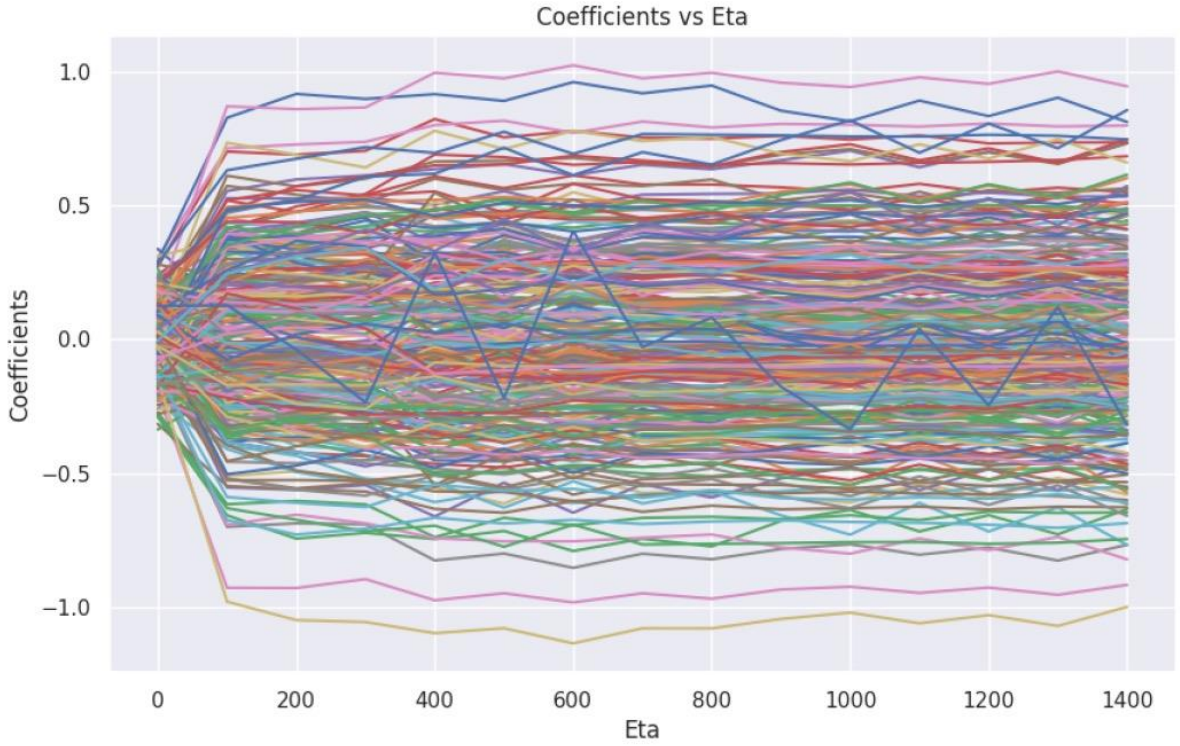
In [23]:
```
import matplotlib.pyplot as plt
plt.plot(etas, [m['four_fifths'] for m in metrics])
plt.ylabel('Four Fifths Rule')
plt.xlabel('PrejudiceRemover (eta value)')
plt.grid()
```



In [24]:
```
data = np.stack([coef[1] for coef in coefficients],axis=0).T
plt.figure(figsize=(20,5))
for i,d in enumerate(data):
    plt.plot(etas, d, label=f"coef[{i}]")
```



In [25]:
```
#plot the coefficients showing how the eta value impacts them
data = np.stack([coef[1] for coef in coefficients], axis=0).T
df_coeff = pd.DataFrame(data.T)
df_coeff.columns = [f"coeff[{c}]" for c in df_coeff.columns]
df_coeff = pd.concat([pd.DataFrame(etas, columns=['eta']), df_coeff], axis=1)


plt.figure(figsize=(10, 6))
for col in df_coeff.columns[1:]:
    plt.plot(df_coeff['eta'], df_coeff[col], label=col)
plt.xlabel('Eta')
plt.ylabel('Coefficients')
plt.title('Coefficients vs Eta')
plt.grid(True)
plt.show()
```



In [26]:
```
#print the coefficients of the mitigated model at different values of eta
#note that eta (η) is the regularisation parameter used by this mitigation
#print the first few rows to show the effect of eta on coefficients
df_coeff.head(10)
```

Out[26]:

	eta	coeff[0]	coeff[1]	coeff[2]	coeff[3]	coeff[4]	coeff[5]	coeff[6]	coeff[7]	coeff[8]	...	co
0	0	-0.054852	0.089808	0.084134	-0.174959	-0.260233	-0.089454	0.087617	0.099651	-0.107097	...	0
1	100	0.379613	-0.232811	-0.227651	-0.301827	-0.088798	-0.269194	0.327990	-0.211817	-0.510483	...	-0
2	200	0.311955	-0.236226	-0.211911	-0.296824	-0.072904	-0.241026	0.270021	-0.106865	-0.493277	...	-0
3	300	0.262303	-0.259830	-0.225701	-0.291614	-0.072906	-0.266835	0.251746	-0.054809	-0.520045	...	-0
4	400	0.266807	-0.345949	-0.202604	-0.324730	-0.176436	-0.316527	0.290289	-0.070708	-0.552841	...	-0
5	500	0.284337	-0.337914	-0.226689	-0.297893	-0.111333	-0.300804	0.293277	-0.066135	-0.613514	...	-0
6	600	0.265068	-0.335865	-0.163002	-0.341369	-0.219263	-0.254719	0.274988	-0.054808	-0.507587	...	-0
7	700	0.284725	-0.323200	-0.235156	-0.317345	-0.153828	-0.274609	0.278830	-0.058350	-0.573189	...	-0
8	800	0.304011	-0.353686	-0.214404	-0.311165	-0.182275	-0.267015	0.263295	-0.075624	-0.565985	...	-0
9	900	0.266933	-0.332149	-0.273843	-0.302593	-0.126826	-0.304386	0.271170	-0.057278	-0.588303	...	-0

10 rows × 302 columns

In [27]:
```
#compare coefficients for the baseline and mitigated model
#print only the row with eta set at 1000
baseline_coefs = df_coeff.loc[9:9]
baseline_coefs
```

Out[27]:

	eta	coeff[0]	coeff[1]	coeff[2]	coeff[3]	coeff[4]	coeff[5]	coeff[6]	coeff[7]	coeff[8]	...	coef
9	900	0.266933	-0.332149	-0.273843	-0.302593	-0.126826	-0.304386	0.27117	-0.057278	-0.588303	...	-0.

1 rows × 302 columns

In [28]:
```
#print the coefficients for the baseline model
mitigated_coefs = pd.DataFrame([baseline_coef], index=['Baseline'])
mitigated_coefs
```

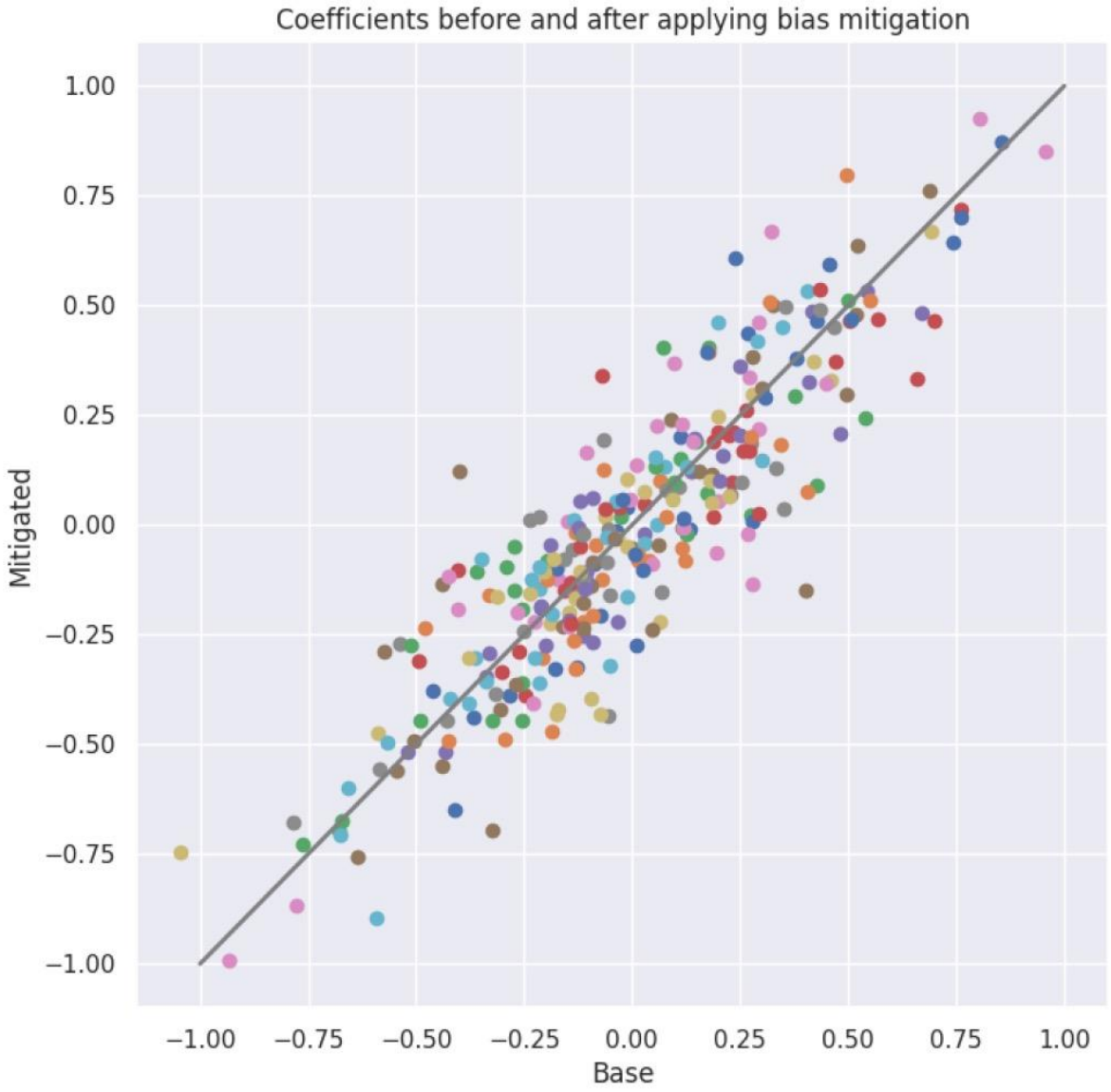
Out[28]:

	0	1	2	3	4	5	6	7	8	
Baseline	0.437588	-0.158197	-0.149287	-0.335417	-0.130331	-0.420376	0.336154	-0.434031	-0.474725	-0.30

1 rows × 301 columns

In [29]:
```
#reshape the data so that coefficients for the mitigated and unmitigated model can be co
coef_data = np.c_[baseline_coefs.iloc[:,1:].values.T, mitigated_coefs.values.T]
coef_data = np.c_[coef_data, np.arange(coef_data.shape[0])]
```

In [30]:
```
#plot coefficients before and after bias mitigation
df_coeff = pd.DataFrame(coef_data)
df_coeff.columns = ['base', 'mitigated', 'coeff']
df_coeff['coeff'] = df_coeff['coeff'].apply(lambda x: f"coeff[{int(x)}]")


plt.figure(figsize=(8, 8))

for coeff_value in df_coeff['coeff'].unique():
    subset = df_coeff[df_coeff['coeff'] == coeff_value]
    plt.scatter(subset['base'], subset['mitigated'], label=coeff_value)

plt.plot([-1, 1], [-1, 1], color='gray', linestyle='-', linewidth=2)

plt.xlabel('Base')
plt.ylabel('Mitigated')
plt.title('Coefficients before and after applying bias mitigation')
plt.grid(True)

plt.show()
```



5. Use post-processing to mitigate bias (Equalised odds - Hardt et al., 2016)

In [31]:

```
from holisticai.bias.mitigation import EqualizedOdds
from holisticai.pipeline import Pipeline
from sklearn.linear_model import LogisticRegression

#use the Holistic AI pipeline to implement the post-processing mitigation
pipeline = Pipeline(steps=[
    ('scaler', StandardScaler()),
    ('model',LogisticRegression()),
    ('bm_postprocessing', EqualizedOdds())])

#fit the pipeline on the raw (unmitigated) training data
pipeline.fit(X, y, bm__group_a=group_a, bm__group_b=group_b)

y_pred = pipeline.predict(X, bm__group_a=group_a, bm__group_b=group_b)
```

```
postpro_metrics=metrics_dataframe(y, y_pred)
postpro_metrics
```

Out[31]:

Metric	Value
Accuracy	0.928074
Balanced accuracy	0.929867
Precision	0.898058
Recall	0.948718
F1-Score	0.922693

In [32]:
```
#calculate bias metrics
postpro_bias=classification_bias_metrics(group_a, group_b, y_pred, y, metric_type='equal
postpro_bias
```

Out[32]:

Metric	Value	Reference
Statistical Parity	0.319961	0
Disparate Impact	2.493151	1
Four Fifths Rule	0.401099	1
Cohen D	0.657902	0
2SD Rule	2.194578	0

Although the performance of this model is similar to the baseline model, the adverse impact metrics continue to be violated.

In [33]:
```
#add these predictions to the baseline predictions to see difference
df_pred['Mitigated'] = y_pred.tolist()
ethnicity=df['Ethnicity']
df_pred=df_pred.join(ethnicity)
```

In [34]:
```
#print the outputs that have been changed by the mitigation
df_pred['Changed'] = np.where(df_pred['Baseline'] == df_pred['Mitigated'], 'No', 'Yes')
df_pred[df_pred['Changed']=='Yes'].iloc[:,:3]
```

Out[34]:

	Baseline	Mitigated	Ethnicity
7	1	0	Asian
22	0	1	Asian
52	1	0	Black
105	1	0	Black
162	0	1	White
173	1	0	White
234	1	0	White
395	1	0	Black
409	1	0	Hispanic

6. Comparing mitigation approaches

To examine which mitigation strategy performs best, the performance of each mitigation approach and presence of adverse impact can be compared.

In [35]:
```
#create a table to display performance metrics across all approaches for comparison
metrics=baseline_metrics
metrics=metrics.rename(columns={'Value':'Baseline'})
metrics['Pre-processing']=prepro_metrics['Value']
metrics['Training constraints']=training_metrics['Value']
metrics['Post-processing']=postpro_metrics['Value']
```

In [36]:
```
#Create a bias metrics table across all approaches for comparison
bias=baseline_bias
bias=bias.drop(columns=['Reference'])
bias=bias.rename(columns={'Value':'Baseline'})
bias['Training data']=trainingdata_bias['Value']
bias['Pre-processing']=prepro_bias['Value']
bias['Training contraints']=training_bias['Value']
bias['Post-processing']=postpro_bias['Value']
```

In [37]:
```
#print the performance metrics
metrics
```

Out[37]:

Metric	Baseline	Pre-processing	Training constraints	Post-processing
Accuracy	0.916473	0.904872	0.851508	0.928074
Balanced accuracy	0.917045	0.917004	0.851811	0.929867
Precision	0.898058	1.000000	0.873786	0.898058
Recall	0.925000	0.834008	0.825688	0.948718
F1-Score	0.911330	0.909492	0.849057	0.922693

In [38]:
```
#print the bias metrics
bias
```

Out[38]:

Metric	Baseline	Training data	Pre-processing	Training contraints	Post-processing
Statistical Parity	0.361057	0.374755	0.0	0.006849	0.319961
Disparate Impact	2.684932	2.748858	1.0	1.013699	2.493151
Four Fifths Rule	0.372449	0.363787	1.0	0.986486	0.401099
Cohen D	0.748464	0.780065	NaN	0.013700	0.657902
2SD Rule	2.476450	2.573132	NaN	0.046954	2.194578

As can be seen from the outputs above, the pre-processing and training constraint techniques were both effective at mitigating the bias that was present in the training data and baseline model. Neither approach considerably impacted model performance, although performance dipped more for training constraints compared to the baseline model. Given the considerable deviation from the adverse impact metrics in the training data and baseline model, both approaches can be said to have effectively mitigated bias. However,

the pre-processing approach resulted in a transformation of the data in a way that meant image choices were no longer represented well by the data.

On the other hand, the post-processing approach improved adverse impact but did not mitigate it such that the metrics were no longer violted.

Therefore, for this applied example, the training constraints were the most effective at mitigating bias and considerably reduced group differences.

v

Appendix B

Worked bias mitigation example for emotional stability (Chapter 2)

The Google Colab version of this worked example can be seen here.

Bias mitigation worked example: Emotional stability

This worked example supports Chapter 2 by demonstrating how three machine learning approaches to bias mitigation work with data from the validation of an image-based assessment of personality designed to be used in recruitment, which is described in greater detail in Chapter 3. Specifically, this applies Learning fair representations (Zemel et al., 2013); Prejudice remover regularizer (Kamishima et al., 2012); and Equalised odds (Hardt et al., 2016) using the Holistic AI library to implement the mitigations.

The assessment, described in greater detail in Chapter 3, presents candidates with 150 pairs of images designed to measure the Big Five and asks them to select the image in the pair that is most like them. The 150 image pairs are used to create 300 dummy variables, which are used to create a simple logistic regression that predicts images in the assessment as predictors in the model to predict personality scores on the IPIP. Given that the training data demonstrates group differences for females for emotional stability, this example focuses on mitigating the adverse impact against females for emotional stability in the predicted scores.

Since the Big Five traits are typically measured continuously and the bias mitigations, like many in computer science, are designed to be used with classification systems, the training data (IPIP Scores) was binarised using the median emotional stability score for the dataset as a threshold in line with the metric that must be used to calculate impact ratios for continuous systems under Local Law 144. The following tutorial builds a baseline classification algorithm before applying the mitigation strategies and demonstrating how they change the data, coefficients, and outputs.

For each mitigation approach, the tutorial shows how it changes the data, coefficients, or outputs. It concludes with a comparison of the effectiveness of each mitigation measure, finding that for this example, the training constraints approach is the most effective approach and effectively mitigated the violations to the adverse impact metrics that were present in the training data and baseline model.

In [1]:
```
#install the Holistic AI library
#!pip install holisticai
!pip install git+https://github.com/holistic-ai/holisticai.git@holisticai-v1
```

0. Import data

In [2]:
```
#import libraries
import pandas as pd
import numpy as np
import scipy as scipy
import matplotlib.pyplot as plt
from sklearn.linear_model import Lasso
from sklearn.model_selection import GridSearchCV
from sklearn.linear_model import Ridge
from sklearn.model_selection import train_test_split
from sklearn.metrics import r2_score
import warnings
warnings.filterwarnings("ignore")
```

In [3]:
```
#import the file from Google Drive
!gdown 1jBGZ2GZnVcTyFdkMRoVSP9S5yBoN6SYm
```

```
Downloading...
From: https://drive.google.com/uc?id=1jBGZ2GZnVcTyFdkMRoVSP9S5yBoN6SYm
To: /content/demo_data.csv
100% 218k/218k [00:00<00:00, 24.7MB/s]
```

In [4]:
```
#load the data and replace missing values with 0
df=pd.read_csv('demo_data.csv')
df = df.replace(np.nan, 0)

#print the first few rows to inspect the data
df.head()
```

Out[4]:

	Age	Gender	Ethnicity	O	C	E	A	ES	N	s_16_2	...	e_20_2	c_39_2	all_25_e	words_7_e	e_

											...				
0	Under 40 years old	Male	Asian	83	92	62	87	68	76	0.0	...	1.0	0.0	1.0	0.0
1	Under 40 years old	Female	Black	91	88	67	80	77	67	0.0	...	1.0	1.0	1.0	0.0
2	Under 40 years old	Female	Asian	81	107	92	104	102	42	0.0	...	1.0	1.0	0.0	0.0
3	Under 40 years old	Male	Asian	94	111	69	118	93	51	0.0	...	1.0	1.0	0.0	0.0
4	Under 40 years old	Female	Asian	93	95	91	102	87	57	1.0	...	0.0	1.0	1.0	1.0

5 rows × 309 columns

1. Calculate success rates and adverse impact metrics for training data

Based on the adverse impact analysis from Chapter 3, male test-takers have a higher pass rate than female test-takers in the training data and results in an adverse impact ratio of .70. Since the four-fifths rule as well as Cohen's D and the 2SD rule were violated for this group, it will serve as a test of the effectiveness of the three bias migiation approaches.

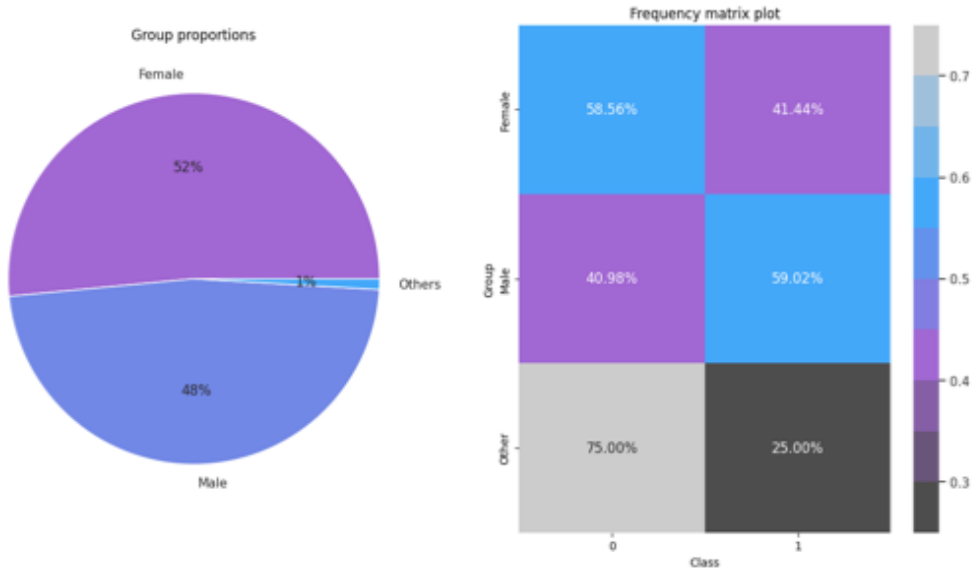
In [5]:
```
#binarise the training data based on the median score
ES_median=df['ES'].median()
df['ES_binarised'] = np.where(df['ES'] > ES_median, 1, 0)
```

In [6]:
```
#define functions to evaluate model performance
# efficacy metrics from sklearn
from sklearn import metrics
import pandas as pd

# dictionnary of metrics
metrics_dict={
        "Accuracy": metrics.accuracy_score,
        "Balanced accuracy": metrics.balanced_accuracy_score,
        "Precision": metrics.precision_score,
        "Recall": metrics.recall_score,
        "F1-Score": metrics.f1_score}

# efficacy metrics dataframe helper tool
def metrics_dataframe(y_pred, y_true, metrics_dict=metrics_dict):
    metric_list = [[pf, fn(y_true, y_pred)] for pf, fn in metrics_dict.items()]
    return pd.DataFrame(metric_list, columns=["Metric", "Value"]).set_index("Metric")
```

In [7]:
```
#create a figure showing the proportion of each gender in the dataset
from holisticai.bias.plots import group_pie_plot,frequency_matrix_plot
fig,axs=plt.subplots(1,2,figsize=(16,8))
group_pie_plot(df['Gender'], ax=axs[0])
```

```
#create a figure showing the proportion of each gender passing (scoring above the median
frequency_matrix_plot(df['Gender'], df['ES_binarised'], ax=axs[1], normalize='group')
```

Out[7]:
```
<Axes: title={'center': 'Frequency matrix plot'}, xlabel='Class', ylabel='Group'>
```



In [8]:
```
#set up data for model training

#X will be predictors and y is the target variable (binarised ES scores)

#drop all data other than predictors from the dataframe
X=df.drop(['Age', 'Gender', 'Ethnicity', 'O', 'C', 'E', 'A',
           'ES', 'N', 'ES_binarised'], axis=1)

#isolate the target variable
y=df['ES_binarised']
```

In [9]:
```
#get success rate (proportion of test-takers passing in each group) based on gender
from holisticai.bias.metrics import frequency_matrix
frequency_matrix(df['Gender'], df['ES_binarised'])
```

Out[9]:

	0	1
Female	0.585586	0.414414
Male	0.409756	0.590244
Other	0.750000	0.250000

In [10]:
```
#calculate bias metrics for the training data

from holisticai.bias.metrics import classification_bias_metrics
group_a = df["Gender"]=='Female'
group_b = df["Gender"]=='Male'
y_true  = y.copy()

trainingdata_bias=classification_bias_metrics(group_a, group_b, y)
trainingdata_bias
```

Out[10]:

	Value	Reference

Metric		
Statistical Parity	-0.175829	0
Disparate Impact	0.702107	1
Four Fifths Rule	0.702107	1
Cohen D	-0.357216	0
2SD Rule	-3.630466	0

Based on the four-fifths rule, Cohen's *d* , and the 2SD rule, the training data shows evidence of bias since <.80, >|.20|, and >|2|, respectively.

2. Build classification model and examine performance

Given that the bias mitigation approaches were created for binary models, logistic regression would be used to predict levels of emotional stability in terms of high (1 - above the median) or low (0 - below the median).

In [11]:
```
#build a logistic model to predict ES from image choices
from sklearn.linear_model import LogisticRegression
model = LogisticRegression(random_state=10)

#first train the model to create the regression line
model.fit(X, y)

#then apply the model to the data to predict the label
ypred = model.predict(X)
```

In [12]:
```
#save the baseline model coefficients and output for later

baseline_coef = list(model.coef_.reshape(-1)) + list(model.intercept_)
df_pred=pd.DataFrame(ypred,columns=['Baseline'])
```

In [13]:
```
#Determine the performance of the baseline model
baseline_metrics=metrics_dataframe(y, ypred)
baseline_metrics
```

Out[13]:

Metric	Value
Accuracy	0.902552
Balanced accuracy	0.902726
Precision	0.892523
Recall	0.909524
F1-Score	0.900943

These accuracy metrics range from 0 to 1, where scores closer to 1 indicate that the model performs well. From these values, this simple model performs well.

In [14]:
```
#calculate bias metrics for the predictions
baseline_bias = classification_bias_metrics(group_a, group_b, ypred, y, metric_type='equ
baseline_bias
```

Out[14]:

Metric	Value	Reference
Statistical Parity	-0.128170	0
Disparate Impact	0.769520	1
Four Fifths Rule	0.769520	1
Cohen D	-0.258526	0
2SD Rule	-2.646983	0

As with the training data, the baseline model is biased against females, violating the four-fifths rule, 2 SD rule, and Cohen's *d*.

3. Use pre-processing to mitigate bias (Learning fair representations - Zemel et al., 2013)

In [15]:
```
from holisticai.bias.mitigation import LearningFairRepresentation
from sklearn.preprocessing import StandardScaler

# initialise the data transformation
lfr = LearningFairRepresentation(k=3, verbose=1, Ax=0.01, Ay=1, Az=50)
scaler1 = StandardScaler()
X_t = scaler1.fit_transform(X)
X_t = lfr.fit_transform(X_t, y, group_a, group_b)
```

In [16]:
```
#print the transformed training data
dframe= pd.DataFrame(X_t)
dframe
```

Out[16]:

	0	1	2	3	4	5	6	7	8	9	...
0	0.370354	0.440823	0.486753	0.780887	0.258027	0.506964	0.195366	0.336316	0.325208	0.510149	...
1	0.335150	0.431111	0.459932	0.804736	0.221765	0.494249	0.204575	0.383373	0.289225	0.525597	...
2	0.344269	0.419548	0.463240	0.793182	0.229558	0.508857	0.205802	0.379968	0.306828	0.527018	...
3	0.373843	0.477751	0.498709	0.792257	0.265707	0.479321	0.185224	0.309213	0.307615	0.494766	...
4	0.378692	0.454547	0.496059	0.779601	0.267913	0.500795	0.190253	0.318043	0.327010	0.502090	...
...	...	...	...	...	...	...	...	...	...	...	...
426	0.335335	0.435558	0.461209	0.806289	0.222454	0.490784	0.203398	0.380383	0.286828	0.523823	...
427	0.356835	0.550722	0.505827	0.833437	0.257010	0.410761	0.169744	0.283492	0.244536	0.472318	...
428	0.338635	0.473926	0.473407	0.818357	0.230109	0.461872	0.192923	0.352601	0.268163	0.507948	...
429	0.397302	0.437865	0.504598	0.758663	0.284604	0.525050	0.190983	0.306780	0.358867	0.502325	...
430	0.352133	0.469256	0.481521	0.806008	0.243059	0.473493	0.191546	0.339796	0.286898	0.505258	...

431 rows × 300 columns

The data above is the training data once it has been transformed by the LearningFairRepresentation mitigation. Instead of having values of 0 or 1, values range from 0 to 1.

In [17]:
```
#train the model on the transformed data
```

```
model = LogisticRegression()
model.fit(X_t, y)
y_pred = model.predict(X_t)

#calculate model performance for the new model trained on the transformed data
prepro_metrics=metrics_dataframe(y, y_pred)
prepro_metrics
```

Out[17]:

Metric	Value
Accuracy	0.584687
Balanced accuracy	0.584651
Precision	0.574766
Recall	0.582938
F1-Score	0.578824

In [18]:
```
#calculate bias metrics for the new predictions
from holisticai.bias.metrics import classification_bias_metrics
prepro_bias=classification_bias_metrics(group_a=group_a, group_b=group_b,
                                        y_pred=y_pred, y_true=y, metric_type='equal_outc
prepro_bias
```

Out[18]:

Metric	Value	Reference
Statistical Parity	-0.142430	0
Disparate Impact	0.748291	1
Four Fifths Rule	0.748291	1
Cohen D	-0.287829	0
2SD Rule	-2.941237	0

Based on these metrics, the model based on the mitigated data does not perform as well as the baseline model, although the adverse impact has been mitigated. This is unsurprising since the raw training data is binary (0s and 1s), while the mitigated training data has been transformed into floats that no longer represent whether an image was chosen or not.

4. Use training constraints to mitigate bias (Prejudice remover regularizer - Kamishima et al., 2012)

In [19]:
```
from holisticai.bias.mitigation import PrejudiceRemover
from holisticai.pipeline import Pipeline
from sklearn.preprocessing import StandardScaler

#note that eta (η) is the regularisation parameter used by this mitigation
# A value of 1000 was selected for eta as the impact ratio plateaued at a value of
#around 1000, as seen in section 4.1

inprocessing = PrejudiceRemover(eta=1.1e3, init_type='StandarLR')
inprocessing.transform_estimator()
```

```
#use the Holistic AI pipline to apply the training constraints
pipeline = Pipeline(steps=[
    ('scaler', StandardScaler()),
    ('bm_inprocessing',inprocessing)])

#apply the pipeline to the raw (unmitigated) training data
pipeline.fit(X, y, bm__group_a=group_a, bm__group_b=group_b)

y_pred = pipeline.predict(X, bm__group_a=group_a, bm__group_b=group_b)
```

In [20]:
```
#evalate model performance
training_metrics=metrics_dataframe(y, y_pred)
training_metrics
```

Out[20]:

Metric	Value
Accuracy	0.849188
Balanced accuracy	0.850783
Precision	0.883178
Recall	0.825328
F1-Score	0.853273

In [21]:
```
#calculate bias metrics
training_bias=classification_bias_metrics(group_a, group_b, y_pred, y)
training_bias
```

Out[21]:

Metric	Value	Reference
Statistical Parity	-0.000176	0
Disparate Impact	0.999669	1
Four Fifths Rule	0.999669	1
Cohen D	-0.000352	0
2SD Rule	-0.003637	0
Equality of Opportunity Difference	0.110941	0
False Positive Rate Difference	0.143223	0
Average Odds Difference	0.127082	0
Accuracy Difference	-0.025577	0

Given the above metrics, the mitigated model performs reasonable well, although not as well as the baseline model but greater than the pre-processing approach. There has also been successful mitigation of subgroup differences since the four-fifths rule is no longer violated.

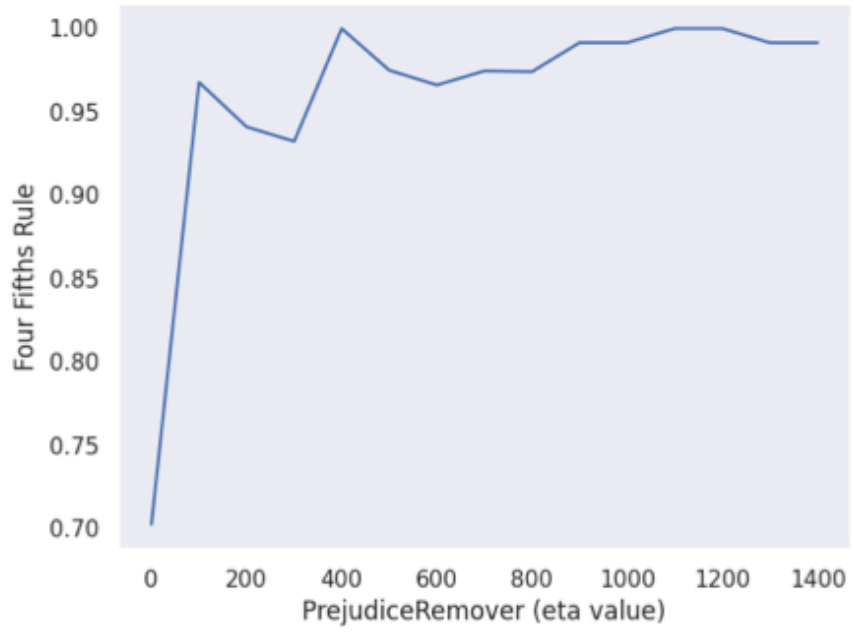
4.1 Exploring eta parameter

In [22]:
```
from tqdm import tqdm
import numpy as np
import warnings
warnings.filterwarnings("ignore")
```

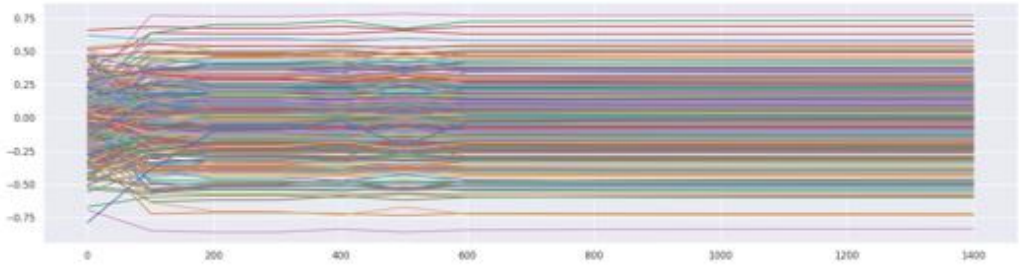
```
etas = np.arange(0, 1500, 100)
metrics = []
coefficients = []
for eta in tqdm(etas):
    inprocessing = PrejudiceRemover(eta=eta, init_type='StandarLR').transform_estimator(
    pipeline = Pipeline(steps=[
        ('scaler', StandardScaler()),
        ('bm_inprocessing',inprocessing)])

    pipeline.fit(X, y, bm__group_a=group_a, bm__group_b=group_b)
    coefficients.append(inprocessing.estimator.coef)
    y_pred = pipeline.predict(X, bm__group_a=group_a, bm__group_b=group_b)

    eff = metrics_dataframe(y, ypred)
    bm = classification_bias_metrics(group_a, group_b, y_pred, y)
    metrics.append({'acc':eff.loc['Accuracy'], 'four_fifths':bm.loc['Four Fifths Rule'].
```

```
100%|██████████| 15/15 [00:08<00:00,  1.73it/s]
```

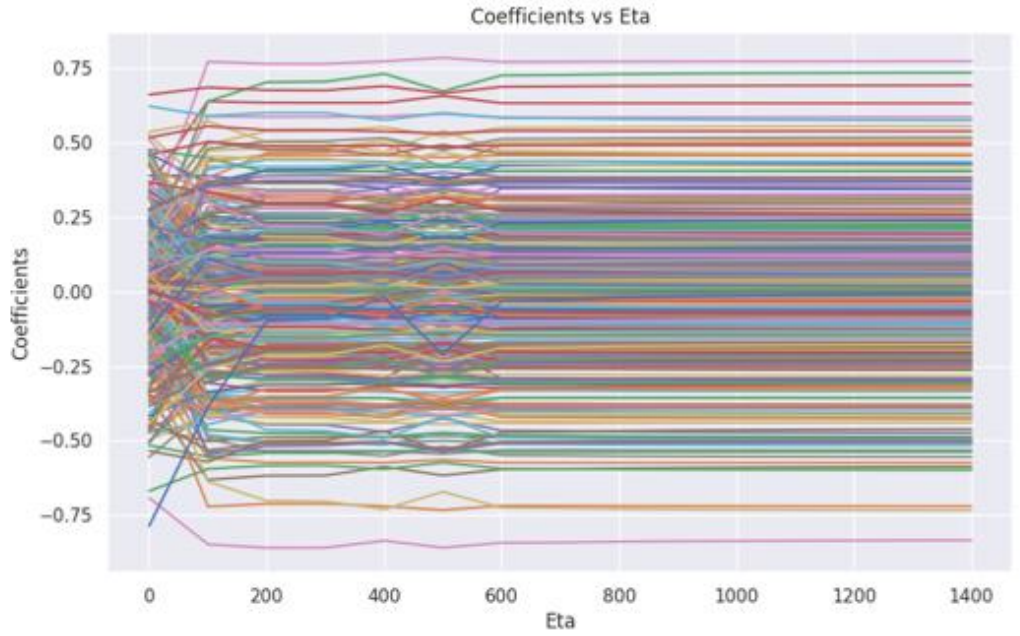
In [23]:
```
import matplotlib.pyplot as plt
plt.plot(etas, [m['four_fifths'] for m in metrics])
plt.ylabel('Four Fifths Rule')
plt.xlabel('PrejudiceRemover (eta value)')
plt.grid()
```



In [24]:
```
data = np.stack([coef[1] for coef in coefficients],axis=0).T
plt.figure(figsize=(20,5))
for i,d in enumerate(data):
   plt.plot(etas, d, label=f"coef[{i}]")
```



In [25]:
```
#plot the coefficients showing how the eta value impacts them
data = np.stack([coef[1] for coef in coefficients], axis=0).T
df_coeff = pd.DataFrame(data.T)
df_coeff.columns = [f"coeff[{c}]" for c in df_coeff.columns]
df_coeff = pd.concat([pd.DataFrame(etas, columns=['eta']), df_coeff], axis=1)


plt.figure(figsize=(10, 6))
for col in df_coeff.columns[1:]:
    plt.plot(df_coeff['eta'], df_coeff[col], label=col)
plt.xlabel('Eta')
plt.ylabel('Coefficients')
plt.title('Coefficients vs Eta')
plt.grid(True)
plt.show()
```



In [26]:
```
#print the coefficients of the mitigated model at different values of eta
#note that eta (η) is the regularisation parameter used by this mitigation
#print the first few rows to show the effect of eta on coefficients
df_coeff.head()
```

Out[26]:

	eta	coeff[0]	coeff[1]	coeff[2]	coeff[3]	coeff[4]	coeff[5]	coeff[6]	coeff[7]	coeff[8]	...	coe
0	0	0.058209	-0.257073	0.038747	0.312435	-0.234130	0.265073	0.469519	-0.049063	-0.133431	...	-0.1

1	100	0.332077	-0.145200	-0.015670	-0.192366	-0.415574	0.301151	0.044920	0.158283	0.085177	...	0.0
2	200	0.298544	-0.132277	-0.043632	-0.215808	-0.446177	0.323638	0.040653	0.172187	0.067823	...	-0.0
3	300	0.298285	-0.132358	-0.044360	-0.216953	-0.446895	0.324293	0.040178	0.173278	0.067686	...	-0.0
4	400	0.267592	-0.119155	-0.048336	-0.246537	-0.470883	0.298065	0.010565	0.184726	0.076960	...	-0.0

5 rows × 302 columns

In [27]:
```
#print only the row with eta set at 1000
baseline_coefs = df_coeff.loc[10:10]
baseline_coefs
```

Out[27]:

	eta	coeff[0]	coeff[1]	coeff[2]	coeff[3]	coeff[4]	coeff[5]	coeff[6]	coeff[7]	coeff[8]	...	coeff[2
10	1000	0.265276	-0.11772	-0.04946	-0.245087	-0.472055	0.29849	0.012364	0.185057	0.074707	...	-0.083

1 rows × 302 columns

In [28]:
```
#print the coefficients for the baseline model
mitigated_coefs = pd.DataFrame([baseline_coef], index=['Baseline'])
mitigated_coefs
```

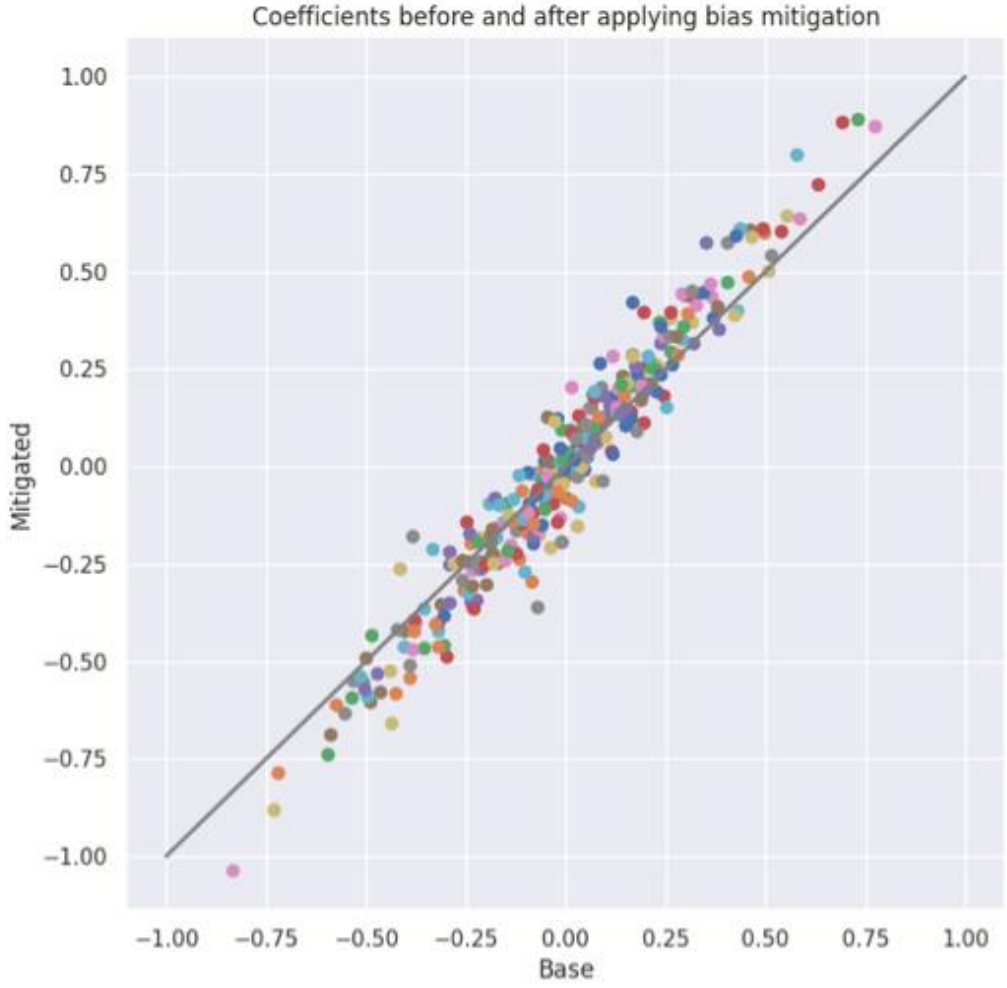
Out[28]:

	0	1	2	3	4	5	6	7	8	
Baseline	0.261354	-0.235236	-0.023399	-0.244702	-0.530512	0.439049	0.204978	0.224854	-0.035524	-0.364

1 rows × 301 columns

In [29]:
```
#reshape the data so that coefficients for the mitigated and unmitigated model can be co
coef_data = np.c_[baseline_coefs.iloc[:,1:].values.T, mitigated_coefs.values.T]
coef_data = np.c_[coef_data, np.arange(coef_data.shape[0])]
```

In [30]:
```
#plot coefficients before and after bias mitigation
df_coeff = pd.DataFrame(coef_data)
df_coeff.columns = ['base', 'mitigated', 'coeff']
df_coeff['coeff'] = df_coeff['coeff'].apply(lambda x: f"coeff[{int(x)}]")


plt.figure(figsize=(8, 8))

for coeff_value in df_coeff['coeff'].unique():
    subset = df_coeff[df_coeff['coeff'] == coeff_value]
    plt.scatter(subset['base'], subset['mitigated'], label=coeff_value)

plt.plot([-1, 1], [-1, 1], color='gray', linestyle='-', linewidth=2)

plt.xlabel('Base')
plt.ylabel('Mitigated')
plt.title('Coefficients before and after applying bias mitigation')
plt.grid(True)

plt.show()
```



5. Use post-processing to mitigate bias (Equalised odds - Hardt et al., 2016)

In [31]:

```
from holisticai.bias.mitigation import EqualizedOdds
from holisticai.pipeline import Pipeline
from sklearn.linear_model import LogisticRegression

#use the Holistic AI pipeline to implement the post-processing mitigation
pipeline = Pipeline(steps=[
    ('scaler', StandardScaler()),
    ('model',LogisticRegression()),
    ('bm_postprocessing', EqualizedOdds())])

#fit the pipeline on the raw (unmitigated) training data
pipeline.fit(X, y, bm__group_a=group_a, bm__group_b=group_b)

y_pred = pipeline.predict(X, bm__group_a=group_a, bm__group_b=group_b)
```

```
postpro_metrics=metrics_dataframe(y, y_pred)
postpro_metrics
```

Out[31]:

Metric	Value
Accuracy	0.914153
Balanced accuracy	0.914261
Precision	0.906542
Recall	0.919431
F1-Score	0.912941

In [32]:
```
#calculate bias metrics
postpro_bias=classification_bias_metrics(group_a, group_b, y_pred, y, metric_type='equal
postpro_bias
```

Out[32]:

Metric	Value	Reference
Statistical Parity	-0.142430	0
Disparate Impact	0.748291	1
Four Fifths Rule	0.748291	1
Cohen D	-0.287829	0
2SD Rule	-2.941237	0

Although the performance of this model is similar to the baseline model, the adverse impact metrics continue to be violated.

In [33]:
```
#add these predictions to the baseline predictions to see difference
df_pred['Mitigated'] = y_pred.tolist()
gender=df['Gender']
df_pred=df_pred.join(gender)
```

In [34]:
```
#print the outputs that have been changed by the mitigation
df_pred['Changed'] = np.where(df_pred['Baseline'] == df_pred['Mitigated'], 'No', 'Yes')
df_pred[df_pred['Changed']=='Yes'].iloc[:,:3]
```

Out[34]:

	Baseline	Mitigated	Gender
23	1	0	Male
26	0	1	Male
69	0	1	Male
81	1	0	Female
115	0	1	Female
126	1	0	Male
127	1	0	Male
131	0	1	Male
252	0	1	Female
253	1	0	Male
270	0	1	Male

271	0	1	Male
272	1	0	Male
302	1	0	Male
310	0	1	Male
338	1	0	Female
356	0	1	Male
360	1	0	Female
412	0	1	Male

Comparing mitigation approaches

To examine which mitigation strategy performs best, the performance of each mitigation approach and presence of adverse impact can be compared.

In [35]:
```
#create a table to display performance metrics across all approaches for comparison
metrics=baseline_metrics
metrics=metrics.rename(columns={'Value':'Baseline'})
metrics['Pre-processing']=prepro_metrics['Value']
metrics['Training constraints']=training_metrics['Value']
metrics['Post-processing']=postpro_metrics['Value']
```

In [36]:
```
#Create a bias metrics table across all approaches for comparison
bias=baseline_bias
bias=bias.drop(columns=['Reference'])
bias=bias.rename(columns={'Value':'Baseline'})
bias['Training data']=trainingdata_bias['Value']
bias['Pre-processing']=prepro_bias['Value']
bias['Training']=training_bias['Value']
bias['Post-processing']=postpro_bias['Value']
```

In [37]:
```
#print the performance metrics
metrics
```

Out[37]:

Metric	Baseline	Pre-processing	Training constraints	Post-processing
Accuracy	0.902552	0.584687	0.849188	0.914153
Balanced accuracy	0.902726	0.584651	0.850783	0.914261
Precision	0.892523	0.574766	0.883178	0.906542
Recall	0.909524	0.582938	0.825328	0.919431
F1-Score	0.900943	0.578824	0.853273	0.912941

In [38]:
```
#print the bias metrics
bias
```

Out[38]:

Metric	Baseline	Training data	Pre-processing	Training	Post-processing
Statistical Parity	-0.128170	-0.175829	-0.142430	-0.000176	-0.142430
Disparate Impact	0.769520	0.702107	0.748291	0.999669	0.748291

Four Fifths Rule	0.769520	0.702107	0.748291	0.999669	0.748291
Cohen D	-0.258526	-0.357216	-0.287829	-0.000352	-0.287829
2SD Rule	-2.646983	-3.630466	-2.941237	-0.003637	-2.941237

As can be seen from the outputs above, the pre-processing and training constraint techniques were both effective at mitigating the bias that was present in the training data and baseline model. However, the pre-processing approach did so at the expense of model accuracy, likely due to the data no longer representing binary image choices.

On the other hand, while the post-processing approach improved model performance compared to the baseline model, adverse impact was not mitigated. Furthermore, changing scores based on subgroup membership could be illegal if used for recruitment tools.

Therefore, for this applied example, the training constraints were the most effective at mitigating bias.

Appendix C
Cohen's d Values for the 150 Items Selected in Study One (Chapter 3)

Mapped trait(s)	*n* respondents selecting		Cohen's *d*					Image retained in model				
	Image 1	Image 2	O	C	E	A	N	O	C	E	A	ES
O	183	249	-0.69	0.00	-0.17	-0.39	-0.13	1		1, 2		
A/ES	249	180	0.34	0.03	0.27	0.54	-0.30	1			1	
C/A	193	237	-0.21	0.31	-0.37	-0.27	0.32	1				
ES	189	240	-0.21	-0.17	-0.72	0.00	0.73	1				
O/C	179	252	0.48	-0.21	0.56	-0.21	-0.35	1				
O	104	325	-0.76	0.31	-0.33	-0.35	0.12	2				
O	325	105	0.83	-0.23	0.36	-0.25	-0.24	2				
O	209	222	-0.54	-0.37	0.22	0.06	0.10	2				
C/ES	144	284	-0.26	-0.20	-0.60	-0.24	0.57	2				
A/ES	242	189	-0.09	0.19	-0.43	-0.43	0.27	1, 2				
ES/C	243	188	0.16	-0.46	-0.33	-0.37	0.34	1, 2				
A	144	285	-0.14	-0.85	0.42	-1.03	-0.01		1		2	1
E	78	352	-0.29	-0.66	-0.94	0.27	0.50		1			
C	281	147	-0.04	1.51	0.27	0.04	-0.54		1			
O	314	116	0.57	0.33	-0.06	0.20	-0.41		1			
C	189	240	0.31	-0.91	-0.93	0.11	0.75		2	2	2	1
ES	263	166	0.30	0.30	1.07	0.18	-0.77		2			2
O/C	206	224	0.92	-0.42	-0.05	-0.13	0.24		2			2
E	131	297	-0.34	-0.30	-0.94	0.00	0.47		2			
C	320	111	-0.30	0.61	0.22	0.09	-0.65		2			
O/ES	175	256	0.34	-0.39	0.55	0.17	-0.56		2			
A	323	107	0.24	0.59	0.66	0.76	-0.24		1, 2	1, 2	1, 2	
ES	158	273	0.08	-0.51	-0.71	0.12	0.84		1, 2	1,2		1,2
C	267	165	0.10	0.81	0.49	-0.18	-0.59		1, 2			1,2
C	273	158	0.07	0.77	-0.10	0.57	-0.06		1, 2			
C	132	299	0.14	-0.75	-0.01	-0.59	0.05		1, 2			
C	205	226	0.12	-0.86	-0.24	0.03	0.31		1, 2			
C	220	210	0.22	-0.74	0.41	-0.61	-0.05		1, 2			
E/C	176	255	0.19	-0.39	0.34	-0.18	-0.19		1, 2			
ES	180	248	-0.32	-0.38	-0.86	-0.07	0.76			1		2
ES/O	265	164	-0.47	-0.13	0.14	-0.15	0.25			1		

E	248	182	-0.15	-0.34	-1.00	0.15	0.43			1		
E	207	223	-0.02	0.14	1.19	0.27	-0.59			2		2
E	241	188	0.03	0.53	-0.85	0.51	0.11			2		2
C/ES	238	191	-0.12	-0.22	-0.95	-0.24	0.75			2		2
E	259	171	-0.38	-0.48	-1.27	-0.02	0.83			2		
E	243	187	-0.14	-0.28	-1.14	-0.13	0.57			2		
O/E	211	220	0.31	-0.21	-1.17	-0.31	0.44			2		
O/ES	243	188	-0.28	0.01	-0.76	0.25	0.61			1, 2		1,2
O	54	377	0.59	-0.26	-0.27	-1.77	0.51			1, 2		
C/E	299	132	-0.17	0.33	-0.72	0.27	-0.19			1, 2		
C/E	157	273	-0.01	-0.35	0.50	-0.62	-0.30			1, 2		
O/ES	202	229	0.43	0.05	0.53	-0.30	-0.49			1, 2		
E/C	141	289	-0.05	-0.22	0.72	-0.39	-0.27			1, 2		
O/ES	161	270	0.33	-0.02	1.35	-0.54	-0.93			1, 2		
E/O	174	257	-0.43	-0.02	0.97	0.09	-0.31			1, 2		
O	104	328	-0.78	-0.72	-0.39	-0.74	0.05				1	
O/E	222	207	0.71	0.24	-0.29	0.48	0.01				1	
O/A	91	339	0.25	0.07	-0.22	-0.71	0.22				1	
A	73	358	-0.53	-0.42	0.21	-1.36	-0.28				1	
A	346	84	-0.04	0.68	-0.01	0.67	0.11				1	
A/C	239	191	0.24	-0.85	0.24	0.55	0.13				1	
E/ES	155	277	0.18	0.00	0.73	-0.19	-1.28				1	
C	364	66	0.27	1.34	0.25	0.39	-0.19				2	
E/A	143	287	-0.17	0.14	0.47	-0.65	-0.49				2	
E/C	172	257	-0.18	-0.73	0.47	-0.40	0.06				2	
O	112	318	-0.54	-0.04	-0.43	-0.48	0.34				1, 2	
A	344	86	0.35	-0.01	0.02	1.09	0.01				1, 2	
A/E	343	88	-0.59	-0.07	-0.39	0.82	0.07				1, 2	
O/ES	140	291	0.34	0.26	-0.28	-0.12	-0.36				1, 2	
A/C	204	225	0.38	-0.59	-0.20	0.21	0.49					1
ES	164	265	-0.11	-0.49	-0.61	0.06	1.11					2
ES	164	267	-0.23	-0.36	-0.76	0.21	0.73					1,2
ES	278	153	-0.36	-0.39	-0.76	0.04	0.69					1,2
A	289	143	0.13	-0.21	-0.37	-0.52	0.29					1,2
ES	266	165	-0.03	0.16	0.43	0.00	-1.01					1,2
E/O	285	145	-0.54	0.03	0.44	0.02	-0.37					1,2

E/A	194	236	-0.13	0.18	0.60	-0.36	-0.79					1,2
A/O	281	149	-0.40	0.28	0.10	0.39	0.16					1,2
ES	328	104	0.19	-0.02	-0.49	0.31	0.71					
A	328	101	0.17	0.71	-0.33	0.63	0.11					
A	342	89	0.57	0.59	0.70	0.58	-0.72					
A	71	359	-0.24	-0.79	0.48	-0.61	-0.11					
A	364	65	-0.22	0.31	-0.13	0.97	-0.19					
A	359	70	0.34	0.68	0.48	0.53	-0.71					
E/ES	172	257	-0.05	-0.04	0.47	0.13	-0.40					
E/ES	253	179	-0.22	0.06	-0.61	0.01	0.61					
E/A	130	301	-0.13	-0.17	0.54	-0.41	-0.50					
ES/C	90	341	-0.20	-0.53	-0.81	-0.63	0.76					
E/ES	262	169	-0.35	0.16	0.73	0.23	-0.44					
E/O	224	208	-0.62	0.02	0.49	-0.06	0.25					
A/ES	177	253	0.16	0.00	0.61	0.28	-0.43					
O/A	198	233	0.39	-0.16	-0.17	-0.39	0.21					
O/E	167	263	0.64	-0.08	-0.24	-0.05	0.14					
A/ES	222	209	0.24	0.39	-0.29	0.24	-0.58					
A/ES	182	249	-0.02	-0.37	-0.35	-0.49	0.24					
C/ES	171	258	0.03	0.49	0.23	0.13	-0.37					
ES/C	254	177	-0.10	-0.31	-0.43	0.09	0.45					
A/C	173	258	0.29	-0.45	0.15	0.22	-0.06					
O	47	382	-0.57	-0.54	-0.69	-0.27	0.51					
C	281	148	0.02	1.06	0.47	0.19	-0.62					
A	178	253	0.21	-0.30	0.30	-0.55	-0.50					
O	108	323	-0.51	0.35	-0.25	0.24	-0.18					
C	98	333	0.03	-0.82	-0.14	-0.40	0.03					
O	109	321	0.54	0.10	0.21	-0.04	-0.13					
C	193	239	0.42	0.72	-0.04	0.38	0.23					
E	225	206	-0.13	-0.38	-1.12	-0.43	0.51					
E	312	120	-0.27	-0.13	-0.87	0.04	0.29					
E	292	138	0.19	0.39	1.08	0.14	-0.83					
E	226	202	0.20	-0.09	1.00	0.08	-0.49					
O	264	166	-0.48	-0.35	-0.66	0.02	0.49					
E	197	233	0.11	-0.29	-1.13	-0.12	0.67					
C	354	75	-0.44	0.96	-0.81	0.52	0.04					

E	224	208	0.17	-0.11	-0.94	0.20	0.37					
C	242	190	-0.04	0.78	0.48	0.36	-0.50					
O	71	359	-1.05	-0.15	-0.51	0.24	0.22					
E	342	86	-0.51	0.18	-1.03	0.49	0.45					
ES	284	146	0.03	0.35	0.53	0.04	-0.83					
ES	109	322	-0.41	-0.62	-1.04	-0.28	0.80					
ES	165	266	-0.19	-0.65	-0.94	0.04	1.17					
C	301	129	0.19	0.73	0.25	0.46	-0.39					
A	180	250	-0.45	-0.40	0.05	-0.70	-0.01					
C/A	280	150	0.51	-0.31	-0.22	0.23	0.44					
C/ES	302	128	0.03	-0.41	-0.65	0.06	0.56					
C/E	268	163	0.08	0.51	-0.28	-0.09	-0.19					
E/A	162	269	0.13	-0.03	0.30	-0.64	-0.26					
A/ES	281	150	0.53	0.22	0.59	0.47	-0.34					
ES/O	190	241	0.37	-0.37	0.00	0.35	-0.58					
E/A	252	179	-0.06	0.12	-0.89	0.26	0.12					
O/E	254	175	-0.35	0.06	0.94	0.27	-0.24					
C/O	269	161	-0.30	0.34	-0.40	-0.12	-0.04					
C/A	202	230	-0.14	0.39	0.32	-0.78	-0.36					
C/ES	222	209	-0.05	-0.38	-0.74	-0.10	0.40					
A/E	207	224	-0.34	0.20	0.24	-0.82	-0.22					
C/E	333	98	0.08	0.95	-0.48	0.69	-0.01					
O/ES	331	100	-0.67	0.36	-0.84	-0.05	0.40					
C/O	301	129	-0.58	0.69	-0.48	-0.14	0.14					
C/ES	195	235	-0.21	-0.36	-0.48	0.18	0.90					
A/E	269	161	0.38	0.15	-0.91	0.34	0.41					
O/A	116	313	0.37	-0.19	0.32	-0.68	-0.31					
O/A	144	287	0.28	-0.11	0.84	-0.82	-0.36					
E/ES	205	225	0.32	0.18	0.58	0.05	-0.49					
E/ES	182	248	-0.20	-0.14	0.94	-0.16	-0.69					
A/ES	312	119	0.18	0.25	0.57	1.02	-0.47					
E/C	196	235	-0.01	-0.51	0.53	0.05	-0.35					
E/O	254	176	-0.65	0.00	0.83	-0.45	-0.43					
E/ES	207	224	0.09	-0.19	0.83	-0.14	-0.35					
E/ES	150	281	0.51	0.16	0.61	0.16	-0.81					
C/A	290	141	-0.04	-0.22	-0.28	0.26	0.29					

ES/A	204	224	0.17	0.06	-0.56	-0.42	0.42					
O/A	202	230	0.30	0.00	0.19	-0.44	-0.22					
ES/A	164	267	0.23	-0.17	-0.52	-0.55	0.28					
O/A	142	290	0.75	-0.24	0.33	-0.29	-0.34					
O/A	196	234	0.35	-0.02	0.33	-0.82	-0.29					
O/C	184	245	0.42	-0.73	-0.26	-0.37	0.61					
A/C	175	254	0.27	-0.71	-0.29	-0.25	0.21					
O/C	209	222	0.91	-0.26	0.12	0.07	-0.31					
A/E	265	166	0.36	-0.17	-0.48	0.40	0.21					
O/E	230	202	0.62	0.13	-0.38	0.58	0.36					
ES/E	232	199	0.04	0.05	-1.05	-0.29	0.29					

Appendix D
Full adverse impact analysis for the image- and questionnaire-based assessments (Study Two; Chapter 3)

Table D1

*Adverse impact analysis for the image-based assessment based on the four-fifths rule, two standard deviations rule, and Cohen's d (*Accepted Adverse Impact Ratio: >.8. Accepted Cohen's D: < |.20|. Accepted 2 SD: <|2|).

Demographic	Subgroup	Group Size	*n* passing	Pass rate	Impact ratio	Cohen's *d*	2SD
Openness							
Age	Age 40 or older	75	35	0.47	0.92	-0.08	-0.61
Age	Under 40 years old	356	180	0.51	1.00		
Gender	Female	222	130	0.59	1.00		
Gender	Male	205	83	0.40	0.69	-0.37	-3.73
Ethnicity	Asian	66	19	0.29	0.47	-0.68	-3.72
Ethnicity	Black	73	38	0.52	0.85	-0.19	0.41
Ethnicity	Hispanic	56	31	0.55	0.90	-0.12	0.88
Ethnicity	Mixed	14	7	0.50	0.81	-0.23	0.01
Ethnicity	Other	13	8	0.62	1.00		0.85
Ethnicity	White	209	112	0.54	0.87	-0.16	1.49
Conscientiousness							
Age	Age 40 or older	75	42	0.56	1.00		
Age	Under 40 years old	356	173	0.49	0.87	0.15	-1.17
Gender	Female	222	122	0.55	1.00		
Gender	Male	205	93	0.45	0.83	-0.19	-1.98
Ethnicity	Asian	66	32	0.48	0.82	-0.21	-0.25
Ethnicity	Black	73	43	0.59	1.00		1.69
Ethnicity	Hispanic	56	23	0.41	0.70	-0.36	-1.41
Ethnicity	Mixed	14	3	0.21	0.36	-0.81	-2.16
Ethnicity	Other	13	4	0.31	0.52	-0.58	-1.40
Ethnicity	White	209	110	0.53	0.89	-0.13	1.11
Extraversion							
Age	Age 40 or older	75	38	0.51	1.00		
Age	Under 40 years old	356	177	0.50	0.98	0.02	-0.15
Gender	Female	222	112	0.50	1.00		
Gender	Male	205	102	0.50	0.99	-0.01	-0.14
Ethnicity	Asian	66	33	0.50	0.94	-0.07	0.02
Ethnicity	Black	73	39	0.53	1.00		0.66
Ethnicity	Hispanic	56	26	0.46	0.87	-0.14	-0.55
Ethnicity	Mixed	14	6	0.43	0.80	-0.21	-0.53
Ethnicity	Other	13	6	0.46	0.86	-0.14	-0.27
Ethnicity	White	209	105	0.50	0.94	-0.06	0.14

Agreeableness							
Age	Age 40 or older	75	40	0.53	1.00		
Age	Under 40 years old	356	175	0.49	0.92	0.08	-0.66
Gender	Female	222	136	0.61	1.00		
Gender	Male	205	79	0.39	0.63	-0.47	-4.69
Ethnicity	Asian	66	28	0.42	0.74	-0.29	-1.32
Ethnicity	Black	73	39	0.53	0.93	-0.07	0.66
Ethnicity	Hispanic	56	21	0.38	0.66	-0.39	-1.99
Ethnicity	Mixed	14	8	0.57	1.00		0.55
Ethnicity	Other	13	5	0.38	0.67	-0.37	-0.84
Ethnicity	White	209	114	0.55	0.95	-0.05	1.88
Emotional stability							
Age	Age 40 or older	75	39	0.52	1.00		
Age	Under 40 years old	356	176	0.49	0.95	0.05	-0.40
Gender	Female	222	98	0.44	0.78	0.25	-2.57
Gender	Male	205	116	0.57	1.00		
Ethnicity	Asian	66	38	0.58	1.00		1.36
Ethnicity	Black	73	36	0.49	0.86	-0.16	-0.11
Ethnicity	Hispanic	56	27	0.48	0.84	-0.19	-0.27
Ethnicity	Mixed	14	5	0.36	0.62	-0.44	-1.08
Ethnicity	Other	13	5	0.38	0.67	-0.38	-0.84
Ethnicity	White	209	104	0.50	0.86	-0.16	-0.05

Note. Violations in **bold**.

Table D2

Adverse impact analysis for the questionnaire-based assessment based on the four-fifths rule, two standard deviations rule, and Cohen's d (Accepted Adverse Impact Ratio: >.8. Accepted Cohen's D: < |.20|. Accepted 2 SD: <|2|).

Demographic	Subgroup	Group Size	*n* passing	Pass rate	Impact ratio	Cohen's *d*	2SD
Openness							
Age	Age 40 or older	75	29	0.39	0.75	-0.26	-2.05
Age	Under 40 years old	356	184	0.52	1.00		
Gender	Female	222	124	0.56	1.00		
Gender	Male	205	86	0.42	0.75	-0.28	-2.87
Ethnicity	Asian	66	21	0.32	0.51	-0.64	-3.11
Ethnicity	Black	73	39	0.53	0.85	-0.18	0.75
Ethnicity	Hispanic	56	35	0.63	1.00		2.10
Ethnicity	Mixed	14	8	0.57	0.91	-0.11	0.59
Ethnicity	Other	13	8	0.62	0.98	-0.02	0.89
Ethnicity	White	209	102	0.49	0.78	-0.28	-0.25
Conscientiousness							
Age	Age 40 or older	75	38	0.51	1.00		
Age	Under 40 years old	356	168	0.47	0.93	0.07	-0.55
Gender	Female	222	119	0.54	1.00		
Gender	Male	205	87	0.42	0.79	-0.22	-2.31
Ethnicity	Asian	66	30	0.45	0.77	-0.27	-0.41
Ethnicity	Black	73	43	0.59	1.00		2.08
Ethnicity	Hispanic	56	24	0.43	0.73	-0.32	-0.79
Ethnicity	Mixed	14	3	0.21	0.36	-0.81	-2.01
Ethnicity	Other	13	5	0.38	0.65	-0.41	-0.68
Ethnicity	White	209	101	0.48	0.82	-0.21	0.21
Extraversion							
Age	Age 40 or older	75	39	0.52	1.00		
Age	Under 40 years old	356	176	0.49	0.95	0.05	-0.40
Gender	Female	222	115	0.52	1.00		
Gender	Male	205	99	0.48	0.93	-0.07	-0.72
Ethnicity	Asian	66	34	0.52	0.90	-0.11	0.29
Ethnicity	Black	73	35	0.48	0.84	-0.18	-0.36
Ethnicity	Hispanic	56	28	0.50	0.88	-0.14	0.02
Ethnicity	Mixed	14	8	0.57	1.00		0.55
Ethnicity	Other	13	7	0.54	0.94	-0.06	0.29
Ethnicity	White	209	103	0.49	0.86	-0.15	-0.24
Agreeableness							
Age	Age 40 or older	75	42	0.56	1.00		
Age	Under 40 years old	356	168	0.47	0.84	0.18	-1.39

Gender	Female	222	129	0.58	1.00		
Gender	Male	205	81	0.40	0.68	-0.38	-3.84
Ethnicity	Asian	66	24	0.36	0.61	-0.48	-2.18
Ethnicity	Black	73	24	0.33	0.55	-0.56	-2.97
Ethnicity	Hispanic	56	26	0.46	0.78	-0.27	-0.37
Ethnicity	Mixed	14	5	0.36	0.60	-0.49	-0.99
Ethnicity	Other	13	6	0.46	0.77	-0.27	-0.19
Ethnicity	White	209	125	0.60	1.00		4.47
Emotional stability							
Age	Age 40 or older	75	44	0.59	1.00		
Age	Under 40 years old	356	170	0.48	0.81	0.22	-1.72
Gender	Female	222	92	0.41	0.70	0.36	-3.63
Gender	Male	205	121	0.59	1.00		
Ethnicity	Asian	66	29	0.44	0.82	-0.19	-1.01
Ethnicity	Black	73	36	0.49	0.92	-0.09	-0.06
Ethnicity	Hispanic	56	30	0.54	0.99	-0.01	0.63
Ethnicity	Mixed	14	7	0.50	0.93	-0.07	0.03
Ethnicity	Other	13	7	0.54	1.00		0.31
Ethnicity	White	209	105	0.50	0.93	-0.07	0.24

Note. Violations in **bold**.

Appendix E
Full adverse impact analysis for the different scoring approaches and predictor combinations in Study Three (Chapter 4)

Table E1

Adverse impact analysis for the Lasso models (Mapped/Intended) based on the four-fifths rule, two standard deviations rule, and Cohen's d (Accepted Adverse Impact Ratio: >.8. Accepted Cohen's D: < |.20|. Accepted 2 SD: <|2|).

Demographic	Subgroup	Group Size	*n* passing	Pass rate	Impact ratio	Cohen's *d*	2SD
Openness							
Age	Age 40 or older	75	33/34	.44/.45	.86/.89	-.14/-.11	-1.12/-.87
Age	Under 40 years old	356	182/181	.51/.51	1.00/1.00		
Gender	Female	222	115/110	.52/.50	1.00/.99	/.01	/-.14
Gender	Male	205	98/103	.48/.50	.92/1.00	-.08/	-.83/
Ethnicity	Asian	66	24/24	.36/.36	**.62/.60**	**-.46/-.50**	**-2.39/-2.39**
Ethnicity	Black	73	38/38	.52/.52	.88/.86	-.14/-.17	.41/.41
Ethnicity	Hispanic	56	33/34	.59/.61	1.00/1.00		1.45/1.74
Ethnicity	Mixed	14	6/7	.43/.50	**.73/.82**	**-.32/-.21**	-.53/.01
Ethnicity	Other	13	7/6	.54/.46	.91/**.76**	-.10/**-.29**	.29/-.27
Ethnicity	White	209	107/106	.51/.51	.87/.84	-.16/-.20	.53/.34
Conscientiousness							
Age	Age 40 or older	75	38/39	.51/.52	1.00/1.00		
Age	Under 40 years old	356	177/176	.5/.49	.98/.95	.02/.05	-.15/-.40
Gender	Female	222	119/121	.54/.55	1.00/1.00		
Gender	Male	205	96/94	.47/.46	.87/.84	-.14/-.17	-1.4/-1.79
Ethnicity	Asian	66	31/33	.47/.5	**.71/.79**	**-.38/-.26**	-.51/.02
Ethnicity	Black	73	48/46	.66/.63	1.00/1.00		**2.98/2.46**
Ethnicity	Hispanic	56	27/24	.48/.43	**.73/.68**	**-.36/-.41**	-.27/-1.13
Ethnicity	Mixed	14	4/6	.29/.43	**.43/.68**	**-.79/-.40**	-1.62/-.53
Ethnicity	Other	13	6/5	.46/.38	**.70/.61**	**-.39/-.49**	-.27/-.84
Ethnicity	White	209	99/101	.47/.48	**.72/.77**	**-.38/-.30**	-1.01/-.63
Extraversion							
Age	Age 40 or older	75	34/36	.45/.48	.89/.95	-.11/-.05	-.87/-.36
Age	Under 40 years old	356	181/179	.51/.50	1.00/1.00		
Gender	Female	222	106/106	.48/.48	.91/.91	.10/.10	-1.02/-1.02
Gender	Male	205	108/108	.53/.53	1.00/1.00		
Ethnicity	Asian	66	37/40	.56/.61	.87/1.00	-.16/	1.09/1.89
Ethnicity	Black	73	38/32	.52/.44	.81/**.72**	**-.24/-.34**	.41/-1.13
Ethnicity	Hispanic	56	31/29	.55/.52	.86/.85	-.18/-.18	.88/.31
Ethnicity	Mixed	14	9/8	.64/.57	1.00/.94	/-.07	1.10/.55

Ethnicity	Other	13	5/6	.38/.46	**.60/.76**	**-.51/-.29**	-.84/-.27
Ethnicity	White	209	95/100	.45/.48	**.71/.79**	**-.38/-.26**	-1.78/-.82
Agreeableness							
Age	Age 40 or older	75	44/44	.59/.59	1.00/1.00		
Age	Under 40 years old	356	171/171	.48/.48	.82/.82	**.21/.21**	-1.67/-1.67
Gender	Female	222	130/131	.59/.59	1.00/1.00		
Gender	Male	205	84/84	.41/.41	**.70/.69**	**-.36/-.37**	**-3.63/-3.72**
Ethnicity	Asian	66	27/27	.41/.41	**.66/.75**	**-.41/-.27**	-1.58/-1.58
Ethnicity	Black	73	38/38	.52/.52	.85/.95	-.19/-.05	.41/.41
Ethnicity	Hispanic	56	19/23	.34/.41	**.55/.75**	**-.56/-.27**	**-2.56**/-1.41
Ethnicity	Mixed	14	8/7	.57/.50	.93/.92	-.09/-.09	.55/.01
Ethnicity	Other	13	8/6	.62/.46	1.00/.85	/-.16	.85/-.27
Ethnicity	White	209	115/114	.55/.55	.89/1.00	-.13/	**2.07**/1.88
Emotional stability							
Age	Age 40 or older	75	40/38	.53/.51	1.00/1.00		
Age	Under 40 years old	356	175/177	.49/.50	.92/.98	.08/.02	-.66/-.15
Gender	Female	222	100/102	.45/.46	.81/.84	**.21/.17**	**-2.18**/-1.79
Gender	Male	205	114/112	.56/.55	1.00/1.00		
Ethnicity	Asian	66	37/37	.56/.56	1.00/1.00		1.09/1.09
Ethnicity	Black	73	36/38	.49/.52	.88/.93	-.13/-.08	-.11/.41
Ethnicity	Hispanic	56	26/27	.46/.48	.83/.86	-.19/-.16	-.55/-.27
Ethnicity	Mixed	14	5/5	.36/.36	**.64/.64**	**-.41/-.41**	-1.08/-1.08
Ethnicity	Other	13	6/5	.46/.38	.82/.69	-.19/**-.35**	-.27/-.84
Ethnicity	White	209	105/103	.50/.49	.90/.88	-.12/-.14	.14/-.24

Note. Adverse impact analysis for the All Lasso models can be seen in Appendix D. Violations in **bold**.

Table E2

Adverse impact analysis for the Ridge models (All/Mapped/Intended) based on the four-fifths rule, two standard deviations rule, and Cohen's d (Accepted Adverse Impact Ratio: >.8. Accepted Cohen's D: < |.20|. Accepted 2 SD: <|2|).

Demographic	Subgroup	Group Size	*n* passing	Pass rate	Impact ratio	Cohen's *d*	2SD
Openness							
Age	Age 40 or older	75	35/33/34	.47/.44/.45	.92/.86/.89	-.08/-.14/-.11	-.61/-1.12/-.87
Age	Under 40 years old	356	180/182/181	.51/.51/.51	1.00/1.00/1.00		
Gender	Female	222	130/115/110	.59/.52/.50	1.00/1.00/.99	//.01	//-.14
Gender	Male	205	83/98/103	.40/.48/.50	**.69**/.92/1.00	**-.37**/-.08/	**-3.73**/-.83/
Ethnicity	Asian	66	19/24/24	.29/.36/.36	**.47/.62/.60**	**-.68/-.46/-.50**	**-3.72/-2.39/-2.39**
Ethnicity	Black	73	38/38/38	.52/.52/.52	.85/.88/.86	-.19/-.14/-.17	.41/.41/.41
Ethnicity	Hispanic	56	31/33/34	.55/.59/.61	.90/1.00/1.00	-.12//	.88/1.45/1.74
Ethnicity	Mixed	14	7/6/7	.50/.43/.50	.81/**.73**/.82	**-.23/-.32/-.21**	.01/-.53/.01
Ethnicity	Other	13	8/7/6	.62/.54/.46	1.00/.91/**.76**	/-.10/**-.29**	.85/.29/-.27
Ethnicity	White	209	112/107/106	.54/.51/.51	.87/.87/.84	-.16/-.16/-.20	1.49/.53/.34
Conscientiousness							
Age	Age 40 or older	75	42/38/39	.56/.51/.52	1.00/1.00/1.00		
Age	Under 40 years old	356	173/177/176	.49/.50/.49	.87/.98/.95	.15/.02/.05	-1.17/-.15/-.4
Gender	Female	222	122/119/121	.55/.54/.55	1.00/1.00/1.00		
Gender	Male	205	93/96/94	.45/.47/.46	.83/.87/.84	-.19/-.14/-.17	-1.98/-1.4/-1.79
Ethnicity	Asian	66	32/31/33	.48/.47/.50	.82/**.71/.79**	**-.21/-.38/-.26**	-.25/-.51/.02
Ethnicity	Black	73	43/48/46	.59/.66/.63	1.00/1.00/1.00		**1.69/2.98/2.46**
Ethnicity	Hispanic	56	23/27/24	.41/.48/.43	**.70/.73/.68**	**-.36/-.36/-.41**	-1.41/-.27/-1.13
Ethnicity	Mixed	14	3/4/6	.21/.29/.43	**.36/.43/.68**	**-.81/-.79/-.40**	**-2.16**/-1.62/-.53
Ethnicity	Other	13	4/6/5	.31/.46/.38	**.52/.70/.61**	**-.58/-.39/-.49**	-1.4/-.27/-.84
Ethnicity	White	209	110/99/101	.53/.47/.48	.89/**.72/.77**	-.13/**-.38/-.30**	1.11/-1.01/-.63
Extraversion							
Age	Age 40 or older	75	38/34/36	.51/.45/.48	1.00/.89/.95	/-.11/-.05	/-.87/-.36
Age	Under 40 years old	356	177/181/179	.50/.51/.50	.98/1.00/1.00	.02//	-.15//
Gender	Female	222	112/106/106	.50/.48/.48	1.00/.91/.91	/.10/.10	/-1.02/-1.02
Gender	Male	205	102/108/108	.50/.53/.53	.99/1.00/1.00	-.01//	-.14//
Ethnicity	Asian	66	33/37/40	.50/.56/.61	.94/.87/1.00	-.07/-.16/	.02/1.09/1.89
Ethnicity	Black	73	39/38/32	.53/.52/.44	1.00/.81/.72	**/-.24/-.34**	.66/.41/-1.13
Ethnicity	Hispanic	56	26/31/29	.46/.55/.52	.87/.86/.85	-.14/-.18/-.18	-.55/.88/.31
Ethnicity	Mixed	14	6/9/8	.43/.64/.57	.80/1.00/.94	**-.21**//-.07	-.53/1.1/.55
Ethnicity	Other	13	6/5/6	.46/.38/.46	.86/**.60/.76**	-.14/**-.51/-.29**	-.27/-.84/-.27
Ethnicity	White	209	105/95/100	.50/.45/.48	.94/**.71/.79**	-.06/**-.38/-.26**	.14/-1.78/-.82
Agreeableness							
Age	Age 40 or older	75	40/44/44	.53/.59/.59	1.00/1.00/1.00		

Age	Under 40 years old	356	175/171/171	.49/.48/.48	.92/.82/.82	.08/**.21/.21**	-.66/-1.67/-1.67
Gender	Female	222	136/130/131	.61/.59/.59	1.00/1.00/1.00		
Gender	Male	205	79/84/84	.39/.41/.41	**.63/.70/.69**	**-.47/-.36/-.37**	**-4.69/-3.63/-3.72**
Ethnicity	Asian	66	28/27/27	.42/.41/.41	**.74/.66/.75**	**-.29/-.41/-.27**	-1.32/-1.58/-1.58
Ethnicity	Black	73	39/38/38	.53/.52/.52	.93/.85/.95	-.07/-.19/-.05	.66/.41/.41
Ethnicity	Hispanic	56	21/19/23	.38/.34/.41	**.66/.55/.75**	**-.39/-.56/-.27**	-1.99/-**2.56**/-1.41
Ethnicity	Mixed	14	8/8/7	.57/.57/.50	1.00/.93/.92	/-.09/-.09	.55/.55/.01
Ethnicity	Other	13	5/8/6	.38/.62/.46	**.67**/1.00/.85	**-.37**//-.16	-.84/.85/-.27
Ethnicity	White	209	114/115/114	.55/.55/.55	.95/.89/1.00	-.05/-.13/	1.88/**2.07**/1.88
Emotional stability							
Age	Age 40 or older	75	39/40/38	.52/.53/.51	1.00/1.00/1.00		
Age	Under 40 years old	356	176/175/177	.49/.49/.5	.95/.92/.98	.05/.08/.02	-.40/-.66/-.15
Gender	Female	222	98/100/102	.44/.45/.46	**.78**/.81/.84	**.25/.21/.17**	**-2.57/-2.18**/-1.79
Gender	Male	205	116/114/112	.57/.56/.55	1.00/1.00/1.00		
Ethnicity	Asian	66	38/37/37	.58/.56/.56	1.00/1.00/1.00		1.36/1.09/1.09
Ethnicity	Black	73	36/36/38	.49/.49/.52	.86/.88/.93	-.16/-.13/-.08	-.11/-.11/.41
Ethnicity	Hispanic	56	27/26/27	.48/.46/.48	.84/.83/.86	-.19/-.19/-.16	-.27/-.55/-.27
Ethnicity	Mixed	14	5/5/5	.36/.36/.36	**.62/.64/.64**	**-.44/-.41/-.41**	-1.08/-1.08/-1.08
Ethnicity	Other	13	5/6/5	.38/.46/.38	**.67**/.82/**.69**	**'-.38**/-.19/**-.35**	-.84/-.27/-.84
Ethnicity	White	209	104/105/103	.50/.50/.49	.86/.90/.88	-.16/-.12/-.14	-.05/.14/-.24

Note. Violations in **bold**.

Table E3

Adverse impact analysis for the OLS models (All/Mapped/Intended) based on the four-fifths rule, two standard deviations rule, and Cohen's d (Accepted Adverse Impact Ratio: >.8. Accepted Cohen's D: < |.20|. Accepted 2 SD: <|2|).

Demographic	Subgroup	Group Size	*n* passing	Pass rate	Impact ratio	Cohen's *d*	2SD
Openness							
Age	Age 40 or older	75	32/37/33	0.43/0.49/0.44	.83/.99/.86	-.18/-.01/-.14	-1.38/-.1/-1.12
Age	Under 40 years old	356	183/178/182	0.51/0.5/0.51	1.00/1.00/1.00		
Gender	Female	222	126/121/120	0.57/0.55/0.54	1.00/1.00/1.00		
Gender	Male	205	87/91/93	0.42/0.44/0.45	.75/.81/.84	**-.29**/-.20/-.17	**-2.96/-2.09**/-1.79
Ethnicity	Asian	66	26/22/21	0.39/0.33/0.32	**.64/.61/.56**	**-.44/-.44/-.52**	**-1.85/-2.92/-3.19**
Ethnicity	Black	73	35/40/38	0.48/0.55/0.52	**.78**/1.00/.91	**-.27**//-.10	-.36/.92/.41
Ethnicity	Hispanic	56	31/30/32	0.55/0.54/0.57	.90/.98/1.00	-.12/-.02/	.88/.59/1.16
Ethnicity	Mixed	14	8/7/7	0.57/0.5/0.5	**.93/.91/.88**	**-.09/-.09/-.14**	.55/.01/.01
Ethnicity	Other	13	8/6/7	0.62/0.46/0.54	1.00/.84/.94	/-.17/-.06	.85/-.27/.29
Ethnicity	White	209	107/110/110	0.51/0.53/0.53	.83/.96/.92	**-.21**/-.04/-.09	.53/1.11/1.11
Conscientiousness							
Age	Age 40 or older	75	43/41/36	0.57/0.55/0.48	1.00/1.00/.95	//-.05	//-.36
Age	Under 40 years old	356	172/174/179	0.48/0.49/0.5	.84/.89/1.00	.18/.12/	-1.42/-.91/
Gender	Female	222	122/121/122	0.55/0.55/0.55	1.00/1.00/1.00		
Gender	Male	205	92/94/93	0.45/0.46/0.45	.82/.84/.83	-.20/-.17/-.19	**-2.08**/-1.79/-1.98
Ethnicity	Asian	66	34/32/32	0.52/0.48/0.48	.80/.80/**.79**	**-.26/-.24/-.26**	.29/-.25/-.25
Ethnicity	Black	73	47/44/45	0.64/0.6/0.62	1.00/1.00/1.00		**2.72/1.95/2.2**
Ethnicity	Hispanic	56	21/25/22	0.38/0.45/0.39	**.58/.74/.64**	**-.55/-.31/-.46**	-1.99/-.84/-1.7
Ethnicity	Mixed	14	5/5/5	0.36/0.36/0.36	**.55/.59/.58**	**-.59/-.50/-.53**	-1.08/-1.08/-1.08
Ethnicity	Other	13	5/7/5	0.38/0.54/0.38	**.60**/.89/**.62**	**-.52**/-.13/**-.47**	-.84/.29/-.84
Ethnicity	White	209	103/102/106	0.49/0.49/0.51	**.77**/.81/.82	**-.31/-.23/-.22**	-.24/-.44/.34
Extraversion							
Age	Age 40 or older	75	38/32/33	0.51/0.43/0.44	1.00/.83/.86	/-.18/-.14	/-1.38/-1.12
Age	Under 40 years old	356	177/183/182	0.5/0.51/0.51	.98/1.00/1.00	.02//	-.15//
Gender	Female	222	109/106/105	0.49/0.48/0.47	.97/.91/.89	.03/.10/.12	-.34/-1.02/-1.21
Gender	Male	205	104/108/109	0.51/0.53/0.53	1.00/1.00/1.00		
Ethnicity	Asian	66	35/41/39	0.53/0.62/0.59	.98/1.00/1.00	-.02//	.56/**2.16**/1.63
Ethnicity	Black	73	35/38/37	0.48/0.52/0.51	.89/.84/.86	**-.12/-.20/-.17**	-.36/.41/.15
Ethnicity	Hispanic	56	25/26/25	0.45/0.46/0.45	.83/**.75/.76**	-.18/**-.32/-.29**	-.84/-.55/-.84
Ethnicity	Mixed	14	7/8/7	0.5/0.57/0.5	.93/.92/.85	-.07/-.10/-.18	.01/.55/.01
Ethnicity	Other	13	7/6/6	0.54/0.46/0.46	**1.00/.74/.78**	**/-.32/-.26**	.29/-.27/-.27
Ethnicity	White	209	106/96/101	0.51/0.46/0.48	.94/**.74**/.82	**-.06/-.33/-.22**	.34/-1.59/-.63
Agreeableness							
Age	Age 40 or older	75	43/44/43	0.57/0.59/0.57	1.00/1.00/1.00		

Age	Under 40 years old	356	172/171/172	0.48/0.48/0.48	.84/.82/.84	.18/**.21**/.18	-1.42/-1.67/-1.42
Gender	Female	222	129/130/127	0.58/0.59/0.57	1.00/1.00/1.00		
Gender	Male	205	85/84/87	0.41/0.41/0.42	**.71/.70/.74**	**-.34/-.36/-.30**	**-3.44/-3.63/-3.05**
Ethnicity	Asian	66	27/28/28	0.41/0.42/0.42	**.64/.74/.76**	**-.47/-.29/-.28**	-1.58/-1.32/-1.32
Ethnicity	Black	73	36/36/41	0.49/0.49/0.56	**.77**/.86/1.00	**-.30**/-.15/	-.11/-.11/1.18
Ethnicity	Hispanic	56	20/22/24	0.36/0.39/0.43	**.56/.69/.76**	**-.58/-.35/-.27**	**-2.27/-1.7/-1.13**
Ethnicity	Mixed	14	9/8/7	0.64/0.57/0.5	1.00/1.00/.89	//-.12	1.1/.55/.01
Ethnicity	Other	13	6/6/4	0.46/0.46/0.31	**.72**/.81/**.55**	**-.36/-.21/-.52**	-.27/-.27/-1.40
Ethnicity	White	209	117/115/111	0.56/0.55/0.53	.87/.96/.95	-.17/-.04/-.06	**2.46/2.07**/1.30
Emotional stability							
Age	Age 40 or older	75	40/39/34	0.53/0.52/0.45	1.00/1.00/.89	//-.11	//-.87
Age	Under 40 years old	356	175/176/181	0.49/0.49/0.51	.92/.95/1.00	.08/.05/	-.66/-.40/
Gender	Female	222	95/101/101	0.43/0.45/0.45	**.74**/.83/.83	**.31**/.19/.19	**-3.15**/-1.99/-1.99
Gender	Male	205	119/113/113	0.58/0.55/0.55	1.00/1.00/1.00		
Ethnicity	Asian	66	37/37/39	0.56/0.56/0.59	1.00/1.00/1.00		1.09/1.09/1.63
Ethnicity	Black	73	36/35/41	0.49/0.48/0.56	.88/.86/.95	-.13/-.16/-.06	-.11/-.36/1.18
Ethnicity	Hispanic	56	26/29/28	0.46/0.52/0.5	.83/.92/.85	-.19/-.09/-.18	-.55/.31/.02
Ethnicity	Mixed	14	6/5/5	0.43/0.36/0.36	**.76/.64/.60**	**-.26/-.41/-.47**	-.53/-1.08/-1.08
Ethnicity	Other	13	7/7/5	0.54/0.54/0.38	.96/.96/**.65**	-.04/-.04/**-.41**	.29/.29/-.84
Ethnicity	White	209	103/102/97	0.49/0.49/0.46	.88/.87/**.79**	-.14/-.14/**-.25**	-.24/-.44/-1.40

Note. Violations in **bold.**

Table E4

Adverse impact analysis for the summative approach based on the four-fifths rule, two standard deviations rule, and Cohen's d (Accepted Adverse Impact Ratio: >.8. Accepted Cohen's D: < |.20|. Accepted 2 SD: <|2|).

Demographic	Subgroup	Group Size	*n* passing	Pass rate	Impact ratio	Cohen's *d*	2SD
Openness							
Age	Age 40 or older	75	35	0.47	0.94	-0.06	-0.48
Age	Under 40 years old	356	177	0.50	1.00		
Gender	Female	222	111	0.50	1.00		
Gender	Male	205	98	0.48	0.96	-0.04	-0.45
Ethnicity	Asian	66	26	0.39	**0.61**	**-0.50**	-1.73
Ethnicity	Black	73	37	0.51	**0.79**	**-0.27**	0.28
Ethnicity	Hispanic	56	34	0.61	0.94	-0.07	1.85
Ethnicity	Mixed	14	9	0.64	1.00		1.15
Ethnicity	Other	13	8	0.62	0.96	-0.05	0.90
Ethnicity	White	209	98	0.47	**0.73**	**-0.35**	-0.93
Conscientiousness							
Age	Age 40 or older	75	35	0.47	1.00		
Age	Under 40 years old	356	160	0.45	0.96	0.03	-0.27
Gender	Female	222	107	0.48	1.00		
Gender	Male	205	87	0.42	0.88	-0.12	-1.19
Ethnicity	Asian	66	33	0.50	0.89	-0.12	0.84
Ethnicity	Black	73	41	0.56	1.00		**2.06**
Ethnicity	Hispanic	56	25	0.45	**0.79**	**-0.23**	-0.10
Ethnicity	Mixed	14	5	0.36	**0.64**	**-0.41**	-0.73
Ethnicity	Other	13	4	0.31	**0.55**	**-0.52**	-1.06
Ethnicity	White	209	87	0.42	**0.74**	**-0.29**	-1.46
Extraversion							
Age	Age 40 or older	75	31	0.41	0.89	-0.10	-0.79
Age	Under 40 years old	356	165	0.46	1.00		
Gender	Female	222	93	0.42	0.84	0.16	-1.63
Gender	Male	205	102	0.50	1.00		
Ethnicity	Asian	66	38	0.58	1.00		**2.15**
Ethnicity	Black	73	27	0.37	**0.64**	**-0.42**	-1.60
Ethnicity	Hispanic	56	26	0.46	0.81	**-0.22**	0.15
Ethnicity	Mixed	14	8	0.57	0.99	-0.01	0.89
Ethnicity	Other	13	6	0.46	0.80	**-0.22**	0.05
Ethnicity	White	209	91	0.44	**0.76**	**-0.28**	-0.78
Agreeableness							

Age	Age 40 or older	75	41	0.55	1.00		
Age	Under 40 years old	356	153	0.43	**0.79**	**0.23**	-1.85
Gender	Female	222	119	0.54	1.00		
Gender	Male	205	75	0.37	**0.68**	**-0.35**	**-3.53**
Ethnicity	Asian	66	28	0.42	0.86	-0.14	-0.46
Ethnicity	Black	73	35	0.48	0.97	-0.03	0.55
Ethnicity	Hispanic	56	19	0.34	**0.69**	**-0.31**	-1.79
Ethnicity	Mixed	14	5	0.36	**0.72**	**-0.27**	-0.71
Ethnicity	Other	13	4	0.31	**0.62**	**-0.38**	-1.05
Ethnicity	White	209	103	0.49	1.00		1.73
Emotional stability							
Age	Age 40 or older	75	41	0.55	1.00		
Age	Under 40 years old	356	166	0.47	0.85	0.16	-1.27
Gender	Female	222	109	0.49	1.00		
Gender	Male	205	98	0.48	0.97	-0.03	-0.27
Ethnicity	Asian	66	28	0.42	**0.72**	**-0.33**	-0.99
Ethnicity	Black	73	43	0.59	1.00		**2.04**
Ethnicity	Hispanic	56	25	0.45	**0.76**	**-0.29**	-0.54
Ethnicity	Mixed	14	6	0.43	**0.73**	**-0.32**	-0.39
Ethnicity	Other	13	5	0.38	**0.65**	**-0.41**	-0.70
Ethnicity	White	209	100	0.48	0.81	**-0.22**	-0.07

Note. Violations in **bold**.

Appendix F
Interviewee demographics (Study Four; Chapter 5)

Identifier	Employment status	Demographics	Diagnoses	Interview Modality	Experience with automated tools
BD	Employed	Female; Asian; 45-54	Dyspraxia	Video	No
RC	Employed; Self-employed	Male; Other ethnic group (Jewish); 35-44	Dyslexia, Dysgraphia	Video	No
MK	Employed	Male; White; 35-44	ADHD; Autism	Video	Yes – game-based assessments, video interviews
MH	Employed	Male; Hispanic; 35-44	ADHD; Autism	Video	Yes – CV scanner
TM	Employed	Male; Black; 25-34	Dyslexia	Live chat	No
DA	Employed	Male; White; 25-34	ADHD	Live Chat	No
NA	Employed; Student	Female; White; 45-54	ADHD, Dyslexia, Dyspraxia	Email	No
SS	Self-employed	Male; White; 25-34	Dyslexia; ADHD	Video	Yes – game-based assessments
SZ	Employed	Female; White; 25-34	ADD	Video	No
LC	Employed	Female; White; 35-44	ADHD; autism; anxiety disorder	Video	Yes – CV scanner, video interview
EH	Employed	Female; White; 18-24	Anxiety disorder	Email	No
SN	Student	Female; Black; 35-44	ADHD	Video	Yes – CV scanner, video interview

Appendix G
Interview Schedule (Study Four; Chapter 5)

Introduction

- Introduction to the research topic
- Pre-employment test/ selection assessment – an assessment taken during the job application process. Includes interviews, assessments of personality and cognitive ability, assessment centres etc
- Automated or algorithmic pre-employment tool – pre-employment test where performance is judged by an algorithm. Includes game-based assessments, chatbots, asynchronous video interviews etc
- Recorded but only I will see it and no identifying information will be included and transcripts will be anonymised
- Responses will be analysed for themes and used to inform hypothesis for future research
- Free to withdraw at any point for any reason, including after the interview
- Questions will be added to the chat as they are asked but are there any other adjustments required?

Question 1 – previous experiences with pre-employment tests in general

- What are your perceptions of pre-employment assessments?
    - Are they more positive or negative?
    - Why this this?

Question 2- Experiences with automated tools

- Have you had any experience with automated tools? This includes an asynchronous video interview where you record answers to predefined questions and an algorithm

analyses responses, game-based assessments which are usually completed on a smartphone and scored by an algorithm, or a s CV screening tool to determine suitability for a position

- If yes:
  - How did you find the tools?
  - Was your experience positive or negative?
  - Why?
- If no:
  - based on my description or any other knowledge you might have, how do you feel about algorithmic recruitment tools?
  - Are your perceptions positive or negative?
  - Why?

Question 3 – automated vs traditional

- Compared to traditional pre-employment tests, including face-to-face interviews, assessment centres, questionnaire based psychometric tests, is your perception of algorithmic recruitment tools more positive or negative?
  - Why?

Question 4 – barriers

- Are there any barriers that you feel prevent you from performing well on pre-employment tests?
  - Do you feel that algorithmic recruitment tools would make these barriers worse or help to alleviate them?
  - Why is this?

Question 5 – any other thoughts that you would like to share?

Appendix H

Full adverse impact analysis for the two assessment formats (Study Six; Chapter 6)

Table H1. *Adverse impact analysis for the image-based assessment based on the four-fifths rule, two standard deviations rule, and Cohen's d (Accepted Adverse Impact Ratio: >.8. Accepted Cohen's D: < |.20|. Accepted 2 SD: <|2|).*

Group	Group Size	*n* passing	Pass rate	Impact ratio	2SD	Cohen's d
Openness						
Age						
Older	123	61	0.50	0.99	-0.10	-0.01
Younger	443	222	0.50	1.00	0.10	0.00
Gender						
Female	275	135	0.49	1.00	0.05	0.00
Male	270	132	0.49	1.00	-0.05	0.00
Ethnicity						
Asian	22	9	0.41	0.61	5.19	-0.52
Black	76	43	0.57	0.85	14.00	-0.21
Hispanic/Latino	24	16	0.67	1.00	16.52	0.00
Mixed	9	5	0.56	0.83	1.12	-0.22
White	425	204	0.48	0.72	9.97	-0.38
Neurodivergent						
Neurotypical	146	73	0.50	1.00	0.00	0.00
Neurodivergent	420	210	0.50	1.00	0.00	0.00
Dyslexic neurotype						
Non-dyslexic	408	190	0.47	0.79	-2.62	0.25
Dyslexic	158	93	0.59	1.00	2.62	0.00
ADHD neurotype						
Non-ADHD	335	170	0.51	1.00	0.43	0.00
ADHD	231	113	0.49	0.96	-0.43	-0.04
Autism neurotype						
Non-autistic	377	193	0.51	1.00	0.80	0.00
Autistic	189	90	0.48	0.93	-0.80	-0.07
Achievement						
Age						
Older	123	71	0.58	1.00	1.94	0.00
Younger	443	212	0.48	0.83	-1.94	0.20
Gender						
Female	275	133	0.48	0.93	-0.90	0.08
Male	270	141	0.52	1.00	0.90	0.00
Ethnicity						
Asian	22	13	0.59	0.90	7.21	-0.14
Black	76	50	0.66	1.00	16.12	0.00

Hispanic/Latino	24	13	0.54	0.82	14.87	-0.24
Mixed	9	4	0.44	0.68	0.51	-0.42
White	425	198	0.47	0.71	8.98	-0.39
Neurodivergent						
Neurotypical	146	97	0.66	1.00	4.61	0.00
Neurodivergent	420	186	0.44	0.67	-4.61	-0.46
Dyslexic neurotype						
Non-dyslexic	408	203	0.50	0.98	-0.19	0.02
Dyslexic	158	80	0.51	1.00	0.19	0.00
ADHD neurotype						
Non-ADHD	335	194	0.58	1.00	4.53	0.00
ADHD	231	89	0.39	0.67	-4.53	-0.39
Autism neurotype						
Non-autistic	377	198	0.53	1.00	1.69	0.00
Autistic	189	85	0.45	0.86	-1.69	-0.15
Extraversion						
Age						
Older	123	52	0.42	0.81	-1.94	-0.20
Younger	443	231	0.52	1.00	1.94	0.00
Gender						
Female	275	123	0.45	0.81	-2.44	0.21
Male	270	149	0.55	1.00	2.44	0.00
Ethnicity						
Asian	22	13	0.59	0.90	7.21	-0.14
Black	76	50	0.66	1.00	16.46	0.00
Hispanic/Latino	24	11	0.46	0.70	12.87	-0.40
Mixed	9	5	0.56	0.84	1.20	-0.20
White	425	198	0.47	0.71	9.74	-0.39
Neurodivergent						
Neurotypical	146	91	0.62	1.00	3.46	0.00
Neurodivergent	420	192	0.46	0.73	-3.46	-0.34
Dyslexic neurotype						
Non-dyslexic	408	194	0.48	0.84	-1.87	0.18
Dyslexic	158	89	0.56	1.00	1.87	0.00
ADHD neurotype						
Non-ADHD	335	180	0.54	1.00	2.14	0.00
ADHD	231	103	0.45	0.83	-2.14	-0.18
Autism neurotype						
Non-autistic	377	211	0.56	1.00	4.01	0.00
Autistic	189	72	0.38	0.68	-4.01	-0.36
Agreeableness						
Age						

Older	123	65	0.53	1.00	0.71	0.00
Younger	443	218	0.49	0.93	-0.71	0.07
Gender						
Female	275	142	0.52	1.00	0.90	0.00
Male	270	129	0.48	0.93	-0.90	-0.08
Ethnicity						
Asian	22	11	0.50	0.75	6.35	-0.34
Black	76	45	0.59	0.89	14.48	-0.15
Hispanic/Latino	24	16	0.67	1.00	16.96	0.00
Mixed	9	4	0.44	0.67	0.46	-0.44
White	425	202	0.48	0.71	9.89	-0.39
Neurodivergent						
Neurotypical	146	93	0.64	1.00	3.84	0.00
Neurodivergent	420	190	0.45	0.71	-3.84	-0.38
Dyslexic neurotype						
Non-dyslexic	408	197	0.48	0.89	-1.31	0.12
Dyslexic	158	86	0.54	1.00	1.31	0.00
ADHD neurotype						
Non-ADHD	335	177	0.53	1.00	1.62	0.00
ADHD	231	106	0.46	0.87	-1.62	-0.14
Autism neurotype						
Non-autistic	377	211	0.56	1.00	4.01	0.00
Autistic	189	72	0.38	0.68	-4.01	-0.36
Emotional stability						
Age						
Older	123	70	0.57	1.00	1.73	0.00
Younger	443	213	0.48	0.84	-1.73	0.18
Gender						
Female	275	117	0.43	0.72	-3.90	0.34
Male	270	160	0.59	1.00	3.90	0.00
Ethnicity						
Asian	22	13	0.59	0.98	7.53	-0.03
Black	76	46	0.61	1.00	15.21	0.00
Hispanic/Latino	24	13	0.54	0.89	14.87	-0.13
Mixed	9	4	0.44	0.73	0.47	-0.32
White	425	201	0.47	0.78	9.85	-0.27
Neurodivergent						
Neurotypical	146	112	0.77	1.00	7.49	0.00
Neurodivergent	420	171	0.41	0.53	-7.49	-0.78
Dyslexic neurotype						
Non-dyslexic	408	203	0.50	0.98	-0.19	0.02
Dyslexic	158	80	0.51	1.00	0.19	0.00

ADHD neurotype						
Non-ADHD	335	194	0.58	1.00	4.53	0.00
ADHD	231	89	0.39	0.67	-4.53	-0.39
Autism neurotype						
Non-autistic	377	213	0.56	1.00	4.37	0.00
Autistic	189	70	0.37	0.66	-4.37	-0.40

Table H2. *Adverse impact analysis for the questionnaire-based assessment based on the four-fifths rule, two standard deviations rule, and Cohen's d (Accepted Adverse Impact Ratio: >.8. Accepted Cohen's D: < |.20|. Accepted 2 SD: <|2|).*

Group	Group Size	*n* passing	Pass rate	Impact ratio	2SD	Cohen's d
Openness						
Age						
Older	123	54	0.44	0.95	-0.47	-0.05
Younger	443	205	0.46	1.00	0.47	0.00
Gender						
Female	275	142	0.52	1.00	2.99	0.00
Male	270	105	0.39	0.75	-2.99	-0.26
Ethnicity						
Asian	22	8	0.36	0.62	5.40	-0.44
Black	76	32	0.42	0.72	11.52	-0.32
Hispanic/Latino	24	14	0.58	1.00	15.59	0.00
Mixed	9	4	0.44	0.76	0.56	-0.27
White	425	194	0.46	0.78	9.58	-0.25
Neurodivergent						
Neurotypical	146	55	0.38	0.78	-2.28	0.22
Neurodivergent	420	204	0.49	1.00	2.28	0.00
Dyslexic neurotype						
Non-dyslexic	408	183	0.45	0.93	-0.70	0.07
Dyslexic	158	76	0.48	1.00	0.70	0.00
ADHD neurotype						
Non-ADHD	335	145	0.43	0.88	-1.42	0.12
ADHD	231	114	0.49	1.00	1.42	0.00
Autism neurotype						
Non-autistic	377	163	0.43	0.85	-1.70	0.15
Autistic	189	96	0.51	1.00	1.70	0.00
Achievement						
Age						
Older	123	65	0.53	1.00	1.96	0.00
Younger	443	190	0.43	0.81	-1.96	0.20
Gender						

Female	275	134	0.49	1.00	1.87	0.00
Male	270	110	0.41	0.84	-1.87	-0.16
Ethnicity						
Asian	22	13	0.59	1.00	8.71	0.00
Black	76	34	0.45	0.76	12.61	-0.29
Hispanic/Latino	24	11	0.46	0.78	12.87	-0.26
Mixed	9	5	0.56	0.94	1.34	-0.07
White	425	187	0.44	0.74	8.59	-0.30
Neurodivergent						
Neurotypical	146	77	0.53	1.00	2.17	0.00
Neurodivergent	420	178	0.42	0.80	-2.17	-0.21
Dyslexic neurotype						
Non-dyslexic	408	183	0.45	0.98	-0.15	0.01
Dyslexic	158	72	0.46	1.00	0.15	0.00
ADHD neurotype						
Non-ADHD	335	169	0.50	1.00	3.11	0.00
ADHD	231	86	0.37	0.74	-3.11	-0.27
Autism neurotype						
Non-autistic	377	169	0.45	0.99	-0.15	0.01
Autistic	189	86	0.46	1.00	0.15	0.00
Extraversion						
Age						
Older	123	64	0.52	1.00	1.31	0.00
Younger	443	201	0.45	0.87	-1.31	0.13
Gender						
Female	275	124	0.45	0.92	-0.97	0.08
Male	270	133	0.49	1.00	0.97	0.00
Ethnicity						
Asian	22	11	0.50	0.90	7.35	-0.11
Black	76	34	0.45	0.81	12.80	-0.21
Hispanic/Latino	24	10	0.42	0.75	12.05	-0.27
Mixed	9	5	0.56	1.00	1.16	0.00
White	425	201	0.47	0.85	9.85	-0.16
Neurodivergent						
Neurotypical	146	85	0.58	1.00	3.20	0.00
Neurodivergent	420	180	0.43	0.74	-3.20	-0.31
Dyslexic neurotype						
Non-dyslexic	408	183	0.45	0.86	-1.51	0.14
Dyslexic	158	82	0.52	1.00	1.51	0.00
ADHD neurotype						
Non-ADHD	335	166	0.50	1.00	1.57	0.00
ADHD	231	99	0.43	0.86	-1.57	-0.13

Autism neurotype						
Non-autistic	377	201	0.53	1.00	4.37	0.00
Autistic	189	64	0.34	0.64	-4.37	-0.40
Agreeableness						
Age						
Older	123	68	0.55	1.00	1.68	0.00
Younger	443	207	0.47	0.85	-1.68	0.17
Gender						
Female	275	152	0.55	1.00	3.13	0.00
Male	270	113	0.42	0.76	-3.13	-0.27
Ethnicity						
Asian	22	10	0.45	0.86	6.10	-0.14
Black	76	40	0.53	1.00	14.31	0.00
Hispanic/Latino	24	10	0.42	0.79	12.51	-0.22
Mixed	9	4	0.44	0.84	0.42	-0.16
White	425	206	0.48	0.92	10.04	-0.08
Neurodivergent						
Neurotypical	146	81	0.55	1.00	1.93	0.00
Neurodivergent	420	194	0.46	0.83	-1.93	-0.19
Dyslexic neurotype						
Non-dyslexic	408	196	0.48	0.96	-0.42	0.04
Dyslexic	158	79	0.50	1.00	0.42	0.00
ADHD neurotype						
Non-ADHD	335	158	0.47	0.93	-0.82	0.07
ADHD	231	117	0.51	1.00	0.82	0.00
Autism neurotype						
Non-autistic	377	200	0.53	1.00	3.00	0.00
Autistic	189	75	0.40	0.75	-3.00	-0.27
Emotional stability						
Age						
Older	123	76	0.62	1.00	3.45	0.00
Younger	443	196	0.44	0.72	-3.45	0.36
Gender						
Female	275	112	0.41	0.72	-3.72	0.32
Male	270	153	0.57	1.00	3.72	0.00
Ethnicity						
Asian	22	9	0.41	0.71	5.11	-0.34
Black	76	44	0.58	1.00	14.74	0.00
Hispanic/Latino	24	13	0.54	0.94	14.87	-0.07
Mixed	9	4	0.44	0.77	0.52	-0.26
White	425	197	0.46	0.80	9.70	-0.23
Neurodivergent						

Neurotypical	146	118	0.81	1.00	9.20	0.00
Neurodivergent	420	154	0.37	0.45	-9.20	-1.00
Dyslexic neurotype						
Non-dyslexic	408	205	0.50	1.00	1.67	0.00
Dyslexic	158	67	0.42	0.84	-1.67	-0.16
ADHD neurotype						
Non-ADHD	335	196	0.59	1.00	5.99	0.00
ADHD	231	76	0.33	0.56	-5.99	-0.53
Autism neurotype						
Non-autistic	377	207	0.55	1.00	4.61	0.00
Autistic	189	65	0.34	0.63	-4.61	-0.42