JEM JOURNAL of EDUCATIONAL MEASUREMENT NCME national council on measurement in education
DOI: 10.1111/jedm.12334 *Journal of Educational Measurement*

Toward Argument-Based Fairness with an Application to AI-Enhanced Educational Assessments

A. Corinne Huggins-Manley
University of Florida
Brandon M. Booth and Sidney K. D'Mello
University of Colorado

The field of educational measurement places validity and fairness as central concepts of assessment quality. Prior research has proposed embedding fairness arguments within argument-based validity processes, particularly when fairness is conceived as comparability in assessment properties across groups. However, we argue that a more flexible approach to fairness arguments that occurs outside of and complementary to validity arguments is required to address many of the views on fairness that a set of assessment stakeholders may hold. Accordingly, we focus this manuscript on two contributions: (a) introducing the argument-based fairness approach to complement argument-based validity for both traditional and artificial intelligence (AI)-enhanced assessments and (b) applying it in an illustrative AI assessment of perceived hireability in automated video interviews used to prescreen job candidates. We conclude with recommendations for further advancing argument-based fairness approaches.

Introduction

The field of educational measurement places validity and fairness as central concepts of assessment quality (AERA et al., 2014). As discussed in this Special Issue, argument-based validity (Kane, 1992) is being adapted and introduced to the context of artificial intelligence (AI) enhanced assessments (heretofore termed AI assessment) to improve validity of inferences and uses, and adoption of such practices are already occurring in some assessment systems (e.g., Burstein et al., 2021). In short, argument-based validity provides a framework for approaching validity as a property of score inferences rather than assessments themselves and as a matter of plausible validity stemming from an argument chain explicitly linking score inferences and uses to evidence (Chapelle, 2021; Kane, 1992, 2013). Prior research has proposed embedding fairness arguments in such argument-based validity processes, particularly when fairness is a concern of comparability in assessment properties across groups (Chapelle, 2021; Xi, 2010). However, we argue that a more flexible approach to fairness arguments that occurs outside of and complementary to validity arguments is required to address some views on fairness that a set of assessment stakeholders may hold.

Meanwhile, researchers and practitioners are increasingly developing educational and psychological assessments that utilize AI to automate and enhance assessment processes (von Davier et al., 2019; D'Mello et al., 2021). In alignment with the National Artificial Intelligence Initiative Act (2020), we define AI as a machine-based system that can "make predictions, recommendations, or decisions influencing real

or virtual environments" (p. 5), and thus our definition includes a variety of methods that may be used to achieve these actions, such as machine learning, neural networks, and natural language processing. Examples of AI assessments include automated essay scoring (Yan et al., 2020), assessment of collaboration skills (Pugh et al., 2021), and so called "stealth" assessments of various competencies (Shute & Ventura, 2013). Meanwhile, AI applications are increasingly met with challenges of unfairness, such as using AI-based gender classification (Buolamwini & Gebru, 2018), AI-based risk assessments for criminal justice decision making (Hübner, 2021), and more (Benjamin, 2019). One issue is that there are no clear standards or frameworks in place for evaluating degrees of fairness in AI, particularly considering fairness issues beyond statistical bias; only methods are being developed and new ones proposed (Kizilcec & Lee, forthcoming; von Davier et al., 2019). Hence, developers of AI-enhanced educational and psychological assessment can benefit from a framework that can provide guidance in evaluating issues of fairness.

Accordingly, we propose an argument-based fairness approach to complement argument-based validity. The argument-based fairness approach can be applied to both traditional and AI-enhanced assessments. Hence, we focus this manuscript on two contributions to the field: introducing the argument-based fairness approach for any type of assessment, and then applying argument-based fairness to the unique context of AI assessment. First, we discuss three views of fairness in relation to validity, followed by a justification for the need for our argument-based fairness approach that can be applied under any of the three views of fairness or even additional views. We then detail a structure for fairness argumentation outside of validity arguments. Then, we summarize the status of fairness investigations in AI assessments and propose applying argument-based fairness to such assessments. We provide an illustrative case study of such an application, and end with a discussion of the potential and future of this approach.

Three Views on Fairness in Relation to Validity

There are many ways to view fairness in a broader societal sense. Focusing on assessment, Xi (2010) discusses three views on fairness with respect to how fairness is and is not related to validity (Figure 1). The first view is that fairness is an assessment quality that is independent of validity, as might be represented by the Code of Fair Testing Practices in Education (Joint Committee on Testing Practices, 2004) and the Standards for Quality and Fairness (Educational Testing Service, 2002). Here, assessment fairness can be viewed as having relationships to or impacts on certain components of validity, but it is not defined by those relationships. Rather, fairness issues are most often aligned to stages of measurement (Xi, 2010), such as providing accommodations during test administration for students with disabilities.

The second view is that fairness is an overarching assessment quality that subsumes validity. Here, validity is a prerequisite to achieving fair assessment, but fair assessment requires more than just validity evidence. Xi (2010) provides the work of Kunnan (2000, 2004) as an example of this viewpoint. Kunnan (2000) defines fairness as a property of assessment that can be present if there is evidence of validity, access, and social justice in the assessment system. In this manuscript, we follow the

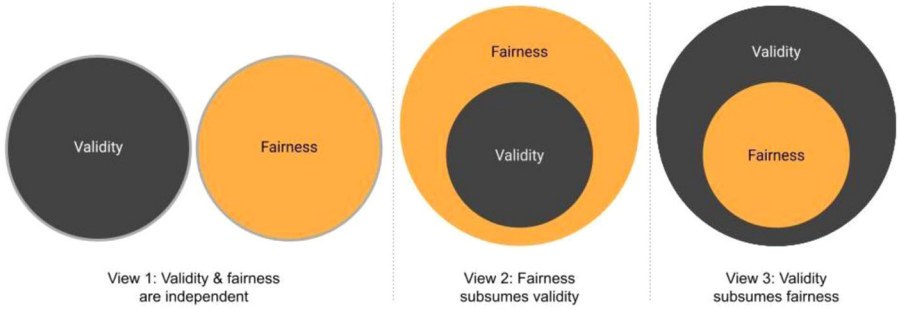

Figure 1. Three views of validity by Xi (2010). [Color figure can be viewed at wileyonlinelibrary.com]

SAGE Encyclopedia of Educational Research, Measurement, and Evaluation (Frey, 2018), which defines social justice as moves to achieve equity through a framework of human rights and diversity, and we use the terms social justice and equity in a broad sense that is open to various definitions of stakeholders in assessment. This viewpoint on fairness puts emphasis on the promotion of social justice and equity through assessment, with validity being a necessary but insufficient means to that end.

The third view is that fairness is subsumed by validity, with problems of fairness necessarily threatening validity. This viewpoint follows a premise that validity is the ultimate aim for assessment quality and problems related to fairness are threats to that aim. Xi (2010) argues that this view on fairness is aligned with the 1999 *Standards for Educational and Psychological Testing* (AERA et al., 1999). In those *Standards* there are three characteristics of fairness: (a) a lack of bias, (b) equitable treatment of all examinees, and (c) equity in opportunity to learn. Notably, all three of these characteristics can be viewed as a matter of comparability (across subgroups, e.g., racial groups). That is, under this view, fairness can be achieved or evidenced if there is comparability in psychometric features, comparability in treatment during the assessment process, and comparability in learning opportunities prior to assessment (Willingham & Cole, 1997; Xi, 2010). Similarly, under this view, one may consider fairness within issues of comparable consequences of testing (AERA et al., 1999, 2014), which allows the placement of fairness rebuttals into a validity argument surrounding consequential validity evidence.

One can imagine other ways of viewing this relationship that are not shown explicitly in Figure 1, such as having varying degrees of overlap between the fairness and validity constructs where one does not fully subsume another while they are still heavily interrelated. Yet we proceed here under Xi's (2010) three views as they suffice for the purposes of justifying the need for our proposed fairness arguments.

The Need for Argument-Based Fairness

Xi (2010) adopts the third view of fairness when proposing a structured process for making and evaluating fairness arguments, explicitly constraining matters of fairness to those of comparable validity. Because this viewpoint defines fairness

problems as threats to validity, the structured fairness arguments take the form of rebuttals to validity arguments (see an applied example of fairness rebuttals in Oliveri et al., 2015). Essentially, if a test score inference cannot be made comparably across various groups, then there is a rebuttal to the validity argument. This approach to fairness argumentation persists in current treatments of argument-based validity (Chapelle, 2021), and there are some advantages to this approach. First, the broad nature of fairness and its relationship to social justice and equity are narrowed, making the process of satisfying matters of fair assessment more achievable. Roughly stated, it may be easier to demonstrate comparable validity as fairness than it is to attempt to demonstrate satisfaction of a broader need for promoting equity or social justice through assessment. Second, as Xi (2010) notes, fairness investigations under this approach can advance when validity theories and approaches advance in the field.

However, we argue that there can be some major disadvantages to this approach. First, by definition of this approach and the nature of rebuttals, fairness is treated secondary to validity. This approach sends an implicit and explicit message that validity is the goal, with fairness being an issue only in that it might threaten validity. However, one can argue that fairness can or should be emphasized more than validity in a particular testing context (e.g., see Kunnan, 2000, 2004). For example, Sireci (2020) discusses when standardization during testing for validity purposes may need to be relaxed to ensure fairness, and that focusing on this issue of fairness primarily can actually be seen as improving validity secondarily when we are open to thinking more broadly about some of our notions of validity. Also, it can be argued that neither fairness nor validity should be so narrowly defined as to be placed under or above the other (see Kane's [2010] preference for broad definitions of fairness and validity). So, while in some cases it is appropriate to work fairness issues into a validity argument, this is not always the case and hence practitioners need a way to shift this relationship to bring fairness arguments to an equal footing as validity arguments, or even to treat fairness arguments as the primary concern. We argue in this manuscript that the lack of consideration for primary fairness arguments in the measurement literature base adds to the implicit notion that validity always comes first, and that this notion can constrain measurement practitioners when thinking more broadly about mechanisms for developing more fair assessment systems.

Second, while Xi (2010) notes that fairness arguments via rebuttals in validity arguments can be flexible as our notions of validity change, fairness approaches should be able to advance along with societal needs even if validity theories do not change. Notably, the 2014 *Standards for Educational and Psychological Measurement*, which came out after Xi's (2010) work, contained large changes to fairness (Jonson et al., 2019) and less changes to validity. This calls for readdressing how we handle fairness, yet there may not be a need for validity approaches to change.

Third, argument-based validity is complex, and adding a full set of fairness arguments embedded within it makes it even more complex. There is already an issue with assessment practitioners not using argument-based validity and those who are using it not focusing on fairness (Kunnan, 2010), and so any fairness arguments that must be couched within a validity argument require first that those are in place and

that argument-based validation is fully understood and embraced by the assessment developer.

Fourth, and important to the involvement of stakeholders in assessment programs, when fairness arguments are wrapped into complex validity arguments, how can such stakeholders other than psychometricians meaningfully engage in the process of improving fairness in assessments? Fairness is a social construct so fairness must be defined through a social process. When engaging in the fairness arguments requires first a full understanding and engagement with the notion of comparable validity, we are essentially requiring a level of knowledge on validity that surpasses the knowledge of many stakeholders in the field of assessment and certainly almost all the public stakeholders of assessment. The complexity of such fairness arguments negates the ability of all stakeholders of an assessment to work together toward building a fairness argument that satisfies various stakeholder groups. We need a way to make claims about fairness that all stakeholders can define and agree on together, and to provide evidence to those claims until stakeholders are satisfied that the fairness claim is supported. Also, there are a plethora of fairness guidelines available to assessment stakeholders that are understandable by such stakeholders, such as the Code of Fair Testing Practices in Education (Joint Committee on Testing Practices, 2004). One can imagine that working with stakeholders to build fairness arguments stemming from such readily understandable guidelines is more easily facilitated than working with stakeholders to build fairness rebuttals to complex validity arguments.

Lastly, Xi's (2010) approach embraces only the third view of fairness above and does not provide a clear mechanism for incorporating any other view of fairness. This is a problem in that fairness issues are constrained to matters of comparable validity, which may not align with a stakeholder's definition of fairness and is overall quite narrow as the world moves to embrace new definitions of fairness. For example, the movements toward culturally sensitive and culturally responsive assessment that align with issues of fairness (Randall, 2021) can be ignored in the rebuttal approach to fairness arguments, as a lack of such sensitivity or responsiveness does not necessarily translate as a matter of incomparable validity. In fact, it has been argued that focusing on issues of comparability across groups as defining fairness can contribute to systems of oppression (Randall, 2021). If the construct of assessment is geared toward the dominant culture of the test taking population, then comparable validity simply becomes a matter of having equal opportunity to work toward obtaining knowledge of constructs defined by and valued by the dominant culture, equal access to the dominant cultural material during the assessment, and equal psychometric properties when conditioning test takers on the dominant cultural construct. In this context, validity and fairness can be "achieved" without addressing matters of power and social justice, which is not acceptable under some views on fairness (Kunnan, 2000; Randall, 2021).

Overall, the fairness argument referenced by Xi (2010) can only take one form, and that is rebuttals to the more flexible validity arguments. We need a way to investigate and argue fairness that allows for the stakeholders of assessment to begin with the open-ended question of "what are claims about fairness that would satisfy the needs of the persons involved in this assessment system, and what evidence do we have or can we work toward to support those claims?" If stakeholders decide to focus on

matters of comparable validity as fairness, then our proposed approach can be used, or one can follow Xi (2010) and embed fairness into validity arguments. However, if stakeholders decide to make claims of fairness outside of group comparability, our proposed approach can still be used. Ultimately, we in the assessment community want to be able to make claims *with* our stakeholders about fairness that can embrace *any* view on fairness, and we need a way to structure those claims with supporting evidence in a way that stakeholders can engage and help shape. We desire to have valid inferences from a validity argument and simultaneously make claims about fairness that can go beyond comparable score inferences from the assessment. This is precisely what the proposed argument-based fairness approach aims to do.

Argument-Based Fairness

Our proposed argument-based fairness approach mirrors the process of argument-based validity in many ways as it follows parts of the Toulmin model of argument structure (1958/2003). To introduce the approach before couching it within AI assessment, we use some isolated examples from ongoing work developing a classroom-based reading assessment in the Institute of Education Science's funded Project DIMES (Huggins-Manley, Benedict, Goodwin, & Templin, 2019–2022, R305A190079).

In argument-based fairness, the overall fairness argument is formed by a series of claims related to fairness issues, broadly defined. The overall argument and the claims to support it should be informed by representatives of each stakeholder group in the assessment system, allowing the overall fairness argument to be socially constructed. So, deciding on an overall argument or set of arguments is an initial step. For example, a group of stakeholders may desire and expect to form an argument that a classroom-based elementary reading assessment is culturally inclusive to the Black American community (Huggins-Manley, 2021). Stakeholders can work together to specify claims that are important to them and, if supported by evidence, would cohesively support this argument. For example, consider the following fairness claim: Scores are derived from test content that is inclusive to the daily cultural experiences of Black American students. Stakeholders may agree that having Black American elementary students and their teachers write and review assessment items can serve as a form of evidence for this claim. This is exactly what we heard from teacher stakeholders in our ongoing work (Huggins-Manley, 2021).

We recommend that claims associated with a fairness argument be structured in tables, flowcharts, or other information organizers that show support logic of each claim, just as seen in validity arguments (e.g., Chapelle, 2021; Oliveri et al., 2015). Structuring fairness arguments by systematic and visual means not only assists in mapping out the components of a Toulmin-based argument, but it may also assist in combining fairness arguments with similarly-structured validity arguments. Table 1 is a generic table format that could be used for this purpose. Once a fairness argument is decided on, the claims needed to support an argument would go in column 1 of Table 1. Table 2 provides a tabular example claim to support the fairness argument laid out above, that the reading assessment is culturally inclusive to the Black American community. Once a fairness claim is in place, the various warrants, or rules

Table 1
Structure of a Table of Claims in Support of a Fairness Argument

Claims Derived with Stakeholder Involvement	Warrants at an Associated Stage of Assessment	Assumptions	Evidence
Claim1: Here a claim is decided on by stakeholders as something that, if backed by evidence, would support the fairness argument.	Warrant 1: Here assessment developers delineate an assessment feature that would have to be present to support the claim.	Assumptions: Here assessment developers list the types of evidence needed to support the warrant.	Evidence: Here assessment developers list evidence that has been gathered to support the assumptions and, hence, the warrant.
	Additional rows for additional warrants of Claim 1	…	…
Additional rows for additional claims	…	…	…

Table 2
Reduced Example of a Fairness Claim, Warrant, Assumption, and Evidence That Would Be Embedded in a Larger Fairness Argument That the Classroom-Based Elementary Reading Assessment Is Culturally Inclusive to Black American Students

One Example Claim	One Example Warrant (Associated Stage of Assessment)	One Example Assumption	One Example Piece of Supporting Evidence from a Fairness Study
Scores are derived from reading assessment content that is inclusive to the daily cultural experiences of Black American students.	Content on the assessment reflects the lived experiences of Black American students. (Content Development stage)	The lived experiences of Black American students are captured by including a sample of such students in the assessment content development process.	Five Black American students from the intended population of assessment takers co-wrote text scenarios for the assessment items that aligned with their daily experiences.

for inferring claims (Kane, 2013), that would support that claim need to be decided. Warrants would be placed in column 2 of Tables 1 and 2.

The fairness argument phases commence once an initial model is in place. Once warrants are in place, assumptions made by the warrants themselves need to be delineated (see column 3 in Tables 1 and 2). Just as in argument-based validity, assumptions are the presumed types of evidence necessary to have in place to justify the warrant, allowing for further clarification about the type of evidence needed to support the claim. Indeed, after claims, assumptions, and warrants are in place, evidence that can support the claims and warrants are laid out (see column 4 in Tables 1 and 2).

A core benefit of defining and mapping claims, warrants, and assumptions is resulting clarity in the types of research or practical endeavors that would need to be conducted to support the fairness argument and claims within it. In the reading assessment example with a cultural inclusiveness claim, one piece of supporting evidence could be the inclusion of multiple Black American students in the generation of content for word problems assessment. This is shown in the last column of Table 2, and the whole of Table 2 shows an example of how the use of argument-based fairness allows test developers to approach fairness in a principled manner that does not relegate fairness to delayed issues of rebuttals to an already mapped out validity framework (Kunnan, 2010).

Once the argument table is laid out, rebuttals are to be welcomed from stakeholders. Using evidence in Table 2 as an example, one may rebut that if the Black American children attend a school in a dominant White culture, they may tend to write assessment items that align with that culture instead of their own. This rebuttal may be used to negate the argument, but also may be used to change the nature of the evidence gathered to address this rebuttal and, ultimately, strengthen the evidence of the fairness argument.

In summary, we posit that having a structured argument-based fairness framework alongside an argument-based validity framework allows for a simultaneous and complementary set of validation and fairness studies that culminate into an assessment program that makes explicit evidence-based claims of both validity and fairness. Whereas the proposed approach to argument-based fairness can be applied to any type of assessment, in the following sections we demonstrate its flexibility and utility by considering its use in AI assessments where issues of fairness are of recent concern.

Fairness in AI Assessments

We agree with a team of international leaders in educational measurement who state (in reference to computational psychometrics, which center on AI),

"While significant progress has been made on the research and development of holistic learning and assessment systems, more work is needed to refine the methodologies, to continuously evaluate them for fairness, efficacy, and validity, and to scale them up." (von Davier et al., 2019, p. 11).

To this point, some recent work has been building toward a framework for fairness in the development and use of AI assessments. Booth et al. (2021a) provided

an exposition of bias (which we define as a statistical manifestation that may indicate a fairness issue) and fairness related to AI assessments of emotion-related constructs. This work established methods and metrics for quantifying and studying bias, aligning such methods to demarked stages of such assessment development. More broadly, Tay et al. (2022) proposed a framework for conceptualizing and quantifying bias in AI assessments that use machine-learning (ML) as the core assessment engine, discussing some distinctions between fairness and bias and then focusing the framework on the latter. They, too, centered their approach to fairness on matters of bias only and aligned the bias concerns to stages of assessment development, which are somewhat different than traditional assessments as elaborated later and discussed in D'Mello et al. (2021).

Although such works are a step toward a framework for fairness in AI assessment, many issues remain. Importantly, several critical questions arise when implementing these recent advances in fairness evaluation: what is our goal for fairness, who gets to decide the goal, and how do we know when we have met our goal? If we compute some fairness metrics (e.g., calculate bias statistics) for a certain number of demographic or otherwise relevant groups, have we achieved fairness? A framework is needed if for no other reason than to establish achievable goals for fairness and provide a mechanism to work toward them.

We use a recently published description of the AI-enhanced assessment system associated with the Duolingo English Test (Burstein et al., 2021) to demonstrate the need for a framework for fairness. Burstein et al. (2021) shared a theoretical assessment ecosystem, which included frameworks for language assessment design, expanded evidence-centered design, computational psychometrics, and test security. They follow Chapelle et al. (2008) in developing a validity argument; however, they do not include fairness as rebuttals to the validity claims as recommended by Chapelle (2021) and Xi (2010). This aligns with Kunnan's 1997 review of validity argument literature in which fairness arguments are often not reached or otherwise included in the process (as discussed in Kunnan, 2010). Indeed, Burstein et al. (2021) recognize the lack of a systematic approach to integrating fairness into the assessment development and implementation. Citing Randall's (2021) work calling for broad changes in how the field views and tackles fairness in assessment, Burstein et al. (2021) state, "Assessment researchers and designers continue to investigate current thinking associated with fairness" (p. 12). They recommend that scholars work toward innovative solutions to integrate with and expand their approach to fairness in a variety of sociopolitical contexts. To this point, we aim to demonstrate the process of integrating structured fairness arguments in the following illustrative case study.

Applying Argument-Based Fairness to AI Assessments

Whereas the above proposed approach to argument-based fairness holds for both traditional and AI assessments, it is the stages of assessment, types of claims, and types of evidence that can separate the latter from the former, at a minimum. As such, it is important to first identify the unique features of AI assessments as discussed extensively elsewhere (D'Mello et al., 2021; Tay et al., 2022) and summarized below:

1. Items in AI assessments correspond to tasks or activities that the participant engages in. These are intended to emulate naturalistic behaviors (e.g., people collaboratively solve a task) rather than behaviors curated for the purpose of assessment (e.g., responding on a computer to hypothetical collaboration scenarios).
2. Participant responses to tasks/activities often yield unstructured (or differentially structured) data (e.g., video, audio, language products, eye gaze) compared to traditional item responses.
3. Transformation of data to high-level abstract representations called features is automatically computed by machines. This may entail using intermediate AI algorithms and the features may be explicit (i.e., exist outside of the assessment itself) or implicitly contained within the algorithmic representations (i.e., in deep learning; Le Cun et al., 2015).
4. Mapping between features and construct (i.e., assessment target) is automatically learned by machine learning algorithms (a workhorse of AI).
5. The algorithms are often selected to learn complex associations (i.e., nonlinear, interactive, time-delayed) among features and often cannot be reduced to an additive combination of item weights.
6. The corresponding assessments generally have good prediction power but lower interpretability compared to traditional psychometric assessments.

It is in the feature computing (item 3) and feature mapping (item 4) stages where "AI" is introduced into the assessment. Typically, this is done using a form of machine learning (ML) called supervised learning where the ML algorithm uses a human-provided supervisory signal to learn the mappings from low-level signals into higher-level abstractions and then to the construct(s) (e.g., identifying a face in an image; identifying a furrowed brow in a face; learning that a furrowed brow can indicate confusion). Due to the heavy reliance on ML, these assessments should be termed ML-enhanced assessments as in Tay et al. (2022), but we use the broader term AI assessments here.

Based on current approaches to fairness in AI assessment (Booth et al., 2021a; Tay et al., 2022) and based on many guidelines for fair assessment (e.g., *Standards* [APA et al., 2014], Code of Fair Testing Practices in Education (Joint Committee on Testing Practices, 2004), Standards for Fairness and Quality (Educational Testing Service, 2002), we recommend that warrants in fairness arguments be couched within the stages of assessment. We adapt Tay et al.'s (2022) analysis when considering the stages shown in Figure 2, aligned with the construct, task, and evidence models espoused by evidence-centered design (ECD) (Mislevy & Haertel, 2006). The stages include:

1. Construct Definition stage: A technical and specific definition of the construct(s) of interest is developed.
2. Content Development stage: Activities are designed that elicit participant behaviors (or provide a context to exhibit behaviors) that relate to the construct(s).
3. Data Collection stage: Sensors and other instruments collect data (e.g., click stream responses, videos, text) from a sample of participants based on the developed content that is relevant for assessing the construct(s).

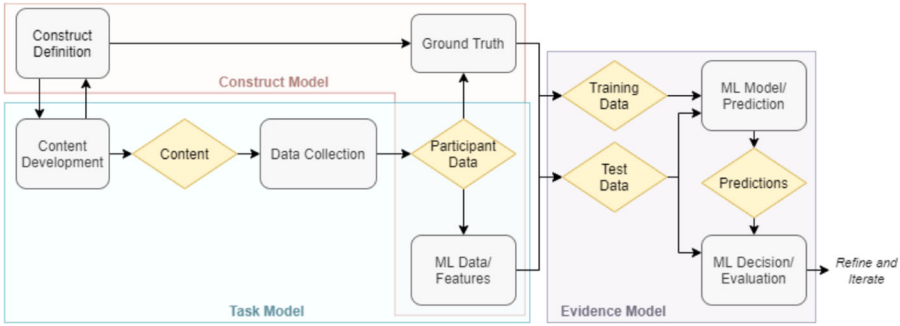

Figure 2. The relationship between different AI-assessment stages (light gray) that comprise a study involving human participants and machine learning for construct prediction and/or decision-making. Arrows represent a flow of information and indicate the order of stage operations. The "Task Model," "Construct Model," and "Evidence Model" boxes encapsulating different portions of the pipeline depict the correspondence between these stages and the *evidence-centered design* framework (Mislevy & Haertel, 2006). [Color figure can be viewed at wileyonlinelibrary.com]

4. Ground Truth stage: Measurements of the construct(s) (e.g., self-reports, expert annotations) are obtained from the collected data and/or observation of the participants while engaged with the developed content.
5. ML Features stage: Characteristic *features* (high-level representations of signals such as specific types of words in a text) are procedurally extracted from signals in the collected data. Note that this step can be skipped in some of the more modern deep learning AI techniques where the features are implicit to the ML itself rather than being externally defined (Le Cun et al., 2015).
6. ML Model/Prediction stage: Samples of *features* and their associated ground truth labels are paired and provided to a supervised learning algorithm. The algorithm inspects a subset of these samples (often deemed the *training data*) to identify and learn common relationships (mappings) between combinations of *features* and ground truth labels. The process can be iterated until a suitable mapping is found or sufficient time has elapsed. The product of machine learning is a computational *model* (a computer program), which can predict the ground truth labels from input samples of *features*.
7. ML Model Testing stage: The remaining samples of *features* (often deemed the *test data*) are provided as input to the *model* and result in a set of *predictions* (AI-based ground-truth predictions), which are then compared to their corresponding ground truth values, resulting in one or more measures of prediction quality. Researchers then decide whether to refine and repeat one or more of the past stages.

Case Study Applying Argument-Based Fairness to AI Assessment

We present an illustrative example of the proposed argument-based fairness approach in a case study involving prerecorded mock interviews collected as part of a preemployment screening study. In these types of AI-assessments (also called

"automated video interviews," or AVIs [Hickman et al., 2021]), machines are used to screen a potentially large pool of job applicants and generate a short list of the top candidates. One may view this as an AI version of a psychological assessment used to screen potential employees. In this context, job applicants are asked to record and submit a video of themselves answering a set of interview questions selected by the employer. Some Fortune 500 companies are already using these systems to expedite the hiring process (HireVue, 2019), and the decisions made by these systems can have profound impacts on which applicants are considered for follow-up interviews. Concerns about fairness in AVIs are especially important because of the ethical and potential legal ramifications, thus AVIs are an ideal domain for structured fairness arguments.

In this case study, college-aged students were recruited to participate in a mock AVI study where an AI-enhanced assessment was trained to predict their "hireability" based on video responses to interview questions. This study was conceived with the aim of understanding if and how group biases and fairness concerns arise in AI assessment, so the mock job and hireability construct were designed to be as group-agnostic as possible. The job was presented as a generic managerial position with no specific qualifications necessary, meaning that participants with certain traits that may be perceived as more desirable for certain jobs (e.g., agreeableness for a call center manager) would not necessarily be preferred. Because the management job was nondescript, hireability was defined in terms of trained annotators' subjective perceptions rather than specific knowledge, skills, abilities, and other characteristics (KSAOs) as would be recommended by the Society for Industrial Organizational Psychology (2018) for real-world application. Accordingly, ground truth for perceived hireability (hereafter called "hireability") was established using a panel of human raters who were given frame-of-reference training for hireability (e.g., does this person seem like they would put forth consistent effort and work well with others) and asked to rate participants by watching their videos. We emphasize that the present focus on hireability is for illustrative purposes only and we make no claims regarding the validity of its use in real-world application screenings.

In the fairness argument for this case study, we will examine issues related to gender groups. Because the job role was non-descript and hireability was framed in terms of perceived consistency of effort and ability to get along with others, there should be no theoretical link between a participant's gender and perceived hireability. Thus, as stakeholders in this study, we assert that there should be no ground truth differences in the distributions of hireability across genders. In studies focused on other constructs, the same statement would not necessarily hold true if gender differences were known to exist (e.g., agreeableness, extraversion; Weisberg et al., 2011), which would change the approach toward some of the bias issues we note below.

The details of this gender fairness study with respect to the stages depicted in Figure 2 are as such:

1. Construct Definition stage: This study focuses on "subjective perceived hireability" of candidates for a hypothetical management role at a non-descript company based solely on the information contained in their recorded interviews.

2. Content Development stage: Six interview questions were developed to elicit differences in candidate (participant) perceived hireability, such as "describe a long-term project that you managed" (see Booth et al., 2021b, for the full list).
3. Data Collection stage: Each participant ($N = 733$) was asked to create a 1- to 3-minute video recording of themselves responding to each question. There were 262 men, 465 women, and six who identified as non-binary. Since participants' genders are conceptually irrelevant to their hireability ratings in this study, a subset of women was used (see Booth et al. [2021b} for details) to yield a balanced sample of men and women, thereby avoiding disadvantaging one gender over another in the machine-learning models. Additionally, the small number of non-binary gender participants were dropped due to insufficient representation. This resulted in ($N = 524$) participant samples comprised of 262 men and 262 women. Below, we discuss the issue of excluding participants with nonbinary gender identities.
4. Ground Truth stage: A small male-female gender-balanced panel of annotators were recruited to review and rate each participant based on their six interview question responses. Following interview best practices, each annotator received 1–2 hours of frame-of-reference training per Campion et al. (1997) including: reviewing the construct definition, reviewing the hireability scale, practicing providing ratings, and discussing sources of (dis)agreement with other annotators. The ground truth was taken as the panel's average hireability score per participant based on two survey items in the form of 5-point Likert scale responses pertaining to hireability. On average, across all participants and assembled annotation panels, the interrater agreement (ICC[1, k]) was .67, which suggested a moderate level of agreement according to Koo and Li (2016).
5. ML Data/Features stage: Characteristic features were extracted from each participant's video, audio, and transcribed text resulting in verbal (i.e., what was said; e.g., n-grams, linguistic inquiry and word counts [Pennebaker et al., 2001]), paraverbal (i.e., how it was said; e.g., loudness, jitter, shimmer), and visual (i.e., body and facial expression; e.g., facial expression valence, upper body motion) information. More details about these features are available in Booth et al. (2021b).
6. ML Model/Prediction stage: The features per participant within each video were pooled together using a set of statistical functions (see Booth et al., 2021b) capturing the mean levels and variability of each feature. Nested fivefold cross-validation was used to tune hyperparameters (details available in Booth et al. (2021b) but not relevant here) of a random forest ML model presented with participant data samples (features) with corresponding ground truth values. Critically, care was taken to ensure that there was no overlap in participants' data across training and testing sets. The output of the model was an estimate of hireability, given a set of features from a held-out (or test) participant.
7. ML Decision/Evaluation stage: Four metrics were used to assess the accuracy and bias of the resulting ML predictions (assessments) with respect to self-identified women and men gender groups: (a) Spearman's correlation was used to measure accuracy; (b) accuracy of gender predictability (can the features distinguish among women and men) was used to assess the level of gender

blindness; (c) the difference in women and men Spearman's correlations was used as a bias measure (per Tay et al., 2022); and (d) the adverse impact ratio (percentage of women selected over the percentage of men) was also calculated. Spearman's correlation was selected over other metrics because it considers the rankings (relative order) of hireability, which is more relevant in hiring decision-making than hireability ratings. Gender blindness and adverse impact both provide relevant measures of fairness because participants' hireability for the non-descript management position in this particular study should not be affected by gender.

The AVI scoring team worked extensively to quantitatively evaluate issues of bias that may arise across men and women candidates, and it became clear that there was a need to understand how these bias assessments relate to an overall fairness goal. To address this need, the team began utilizing the proposed argument-based fairness approach. For this manuscript, we present a fairness argument in Table 3 containing claims pertaining to each stage in Figure 2. However, readers should note that some parts of Table 3 have not yet been completed (as indicated by asterisks in Table 3), and we have not yet engaged with other stakeholders to develop the claims themselves. Hence, the case study is for exposition only, and while the proposed fairness argument framework is generalizable, the specific claims in this table may not be useful in other domains.

The fairness argument for this case study is: The proposed AVI assessment is inclusive of gender-based experiences of women and men and does not disadvantage candidates based on their identified gender as women or men. This argument is subjective, as all fairness arguments are by the socially constructed nature of fairness, and hence is supported by claims, warrants, assumptions, and evidence that can be continuously developed with stakeholders (who may provide rebuttals and more) until there is some form of agreement that the argument is supported within reason. Table 3 displays these features in support of the overall argument. Readers should note that additional fairness arguments would be needed to address issues pertaining to other groups (e.g., race, individual fairness, non-binary genders), and also that complementary measures of hireability based on KSAOs (Society for Industrial and Organizational Psychology, 2018) rather than perceived hireability would be needed for real-world application. Hence, the following fairness argument is for illustrative purposes only. After presenting the argument we discuss some rebuttals that would call into question the strength of the argument; this is not fully complete but is presented solely to highlight how a rebuttal process might take place in a fairness argument.

While Table 3 shows much work toward supporting the stated fairness argument, there are three things that must be done for our team to feel confident enough in the strength of the argument to share it with the public. First, all asterisked evidence in Table 3 needs to be completed. Second, stakeholders need to be thoroughly involved in our process. For example, stakeholders can help us to decide if there are any claims that we would need to add to support the argument, and also if there are any rebuttals to the current evidence that would call it into question. Third, whether with stakeholders or without, it is important to formally state and evaluate rebuttals to the

Table 3
Argument-Based Support for the AVI Fairness Argument: The Proposed AVI Assessment Is Inclusive of Gender-Based Experiences of Women and Men and Does Not Disadvantage Candidates Based on Their Identified Gender as Women or Men

Claims Derived with Stakeholder Involvement	Warrants at an Associated Stage of AI-Assessment	Assumptions	Evidence
1. The content used to elicit video interview material is inclusive of the lived experiences of self-identified women and men.	a. The hireability construct is defined by both women and men experts with a focus on avoiding a definition centered on a gender-dominant perspective. (Construct definition stage)	Experts on hireability are involved in defining the construct. Experts are instructed to consider the construct definition in a manner that does not favor women or men.	Hireability in AVI was defined by a process of expert collaboration that included self-identified women and men and that was guided by specific instructions to ensure that the definition of the construct was *not* centered on a male-dominant perspective.*
	b. The interview question content allows for candidates to provide open-ended data about their experiences and behaviors. (Content development stage)	Interview questions are open-ended.	The interview questions specifically prompt candidates to describe prior experiences and behaviors without additional constraints, apart from a suggested time limit.*
	c. The data gathered for ground truth of hireability (all men and women, $N = 727$) does not favor men over women or vice versa. (Ground truth stage)[1]	Distributions of hireability scores are equivalent between women and men because the nature of the perceived hireability in this assessment does not depend on gender.	The distributions of ground truth hireability scores for women and men are not significantly different; a Kolmogorov-Smirov test fails to reject the null hypothesis ($D = .06$, $p = .63$)

(*Continued*)

Table 3
(Continued)

Claims Derived with Stakeholder Involvement	Warrants at an Associated Stage of AI-Assessment	Assumptions	Evidence
2. The ML model produces hireability scores that are sufficiently accurate for employment prescreening.	a. Ground truth hireability assessments are accurate representations of the hireability construct.[2] (Decision/evaluation stage)	The accuracy measures between ML predictions and hireability are sufficiently large.	The ML predictions achieve a Spearman's correlation of .4 with ground truth measures, corresponding to a moderate effect size (Cohen, 1992).
	b. The features used to train the ML model are accurate representations of utterances during the interview. (ML Data/Features stage)	Computer-derived transcripts sufficiently align with human-produced transcripts.	The computer-derived transcript moderately aligns with human-produced transcripts in over 130 cases with an average word error rate of .51.
3. The ML model produces hireability scores that are unbiased across self-identified women and men.	a. Annotators reviewing participant videos are trained in avoiding bias in ratings-based men- or women-normative perspectives. (Ground truth stage)	A training is developed for this purpose and provided to all annotators.	Annotators were provided with a 30-minute training on typical biases that arise when evaluating women candidates as compared to men candidates.*
	b. Items used in the survey for the ground truth data demonstrate measurement invariance across self-identified women and men. (Ground truth stage)	There is an absence of differential item functioning across self-identified women and men in the ground truth survey items.	The ground truth surveys have previously been tested for differential item functioning across self-identified women and men.*

(*Continued*)

Table 3
(Continued)

Claims Derived with Stakeholder Involvement	Warrants at an Associated Stage of AI-Assessment	Assumptions	Evidence
	c. The quantity of missing information in the features is comparable between self-identified women and men, and missing data is due to factors independent of participant characteristics that may be related to gender (e.g., voice, appearance, behavior). (ML data/feature stage)	If missing features data is present, an analysis to evaluate the independence of missingness from self-identified gender is conducted.	There is no missing data in the features for either self-identified men or women. If missing features data is present in future iterations, an analysis to evaluate the independence of missingness from self-identified gender must be conducted.
	d. During training, the model is presented with equal quantities of examples of self-identified women and men with similar hireability scores across the range of possible scores. (ML model/prediction stage)	Equal quantities of women and men data are provided to the ML model, and if their hireability score distributions are different, an analysis is conducted to identify a subset of data where the men and women hireability distributions are sufficiently similar to each other and representative of the hireability scores in the entire data.[3] A sufficient range of hireability scores are present in the data among both men and women.	A bipartite matching process ensures that features from an equal number of self-identified women and men are included where the distributions of hireability scores for these men and women are similar. After matching, no significant difference in the distributions is detected (KS-test: $D = .06$, $p = .57$).

(*Continued*)

Table 3
(Continued)

Claims Derived with Stakeholder Involvement	Warrants at an Associated Stage of AI-Assessment	Assumptions	Evidence
	e. The accuracy measures between ML predictions and ground truth for self-identified women and men is comparable. (Decision/evaluation stage)	Correlations between ML predictions and the ground truth are similar for self-identified women and men.	Spearman's correlations between ML predictions and the ground truth are similar for self-identified women and men. The difference in average Spearman's correlation computed separately for men and women is .01, a negligible effect.
	f. The ML model is trained on a subset of features which provides the most utility for hireability assessment and also which obfuscates the gender identity (men or women) of the candidates. (ML model/prediction stage)	A feature selection process leaves out the features which enable the ML model to reliably predict gender.	An iterative feature exclusion process leaves out the features which enable the ML model to predict gender. The area under the receiver operating characteristic curve (AUROC) of gender predictability with the resulting feature set is .5 (i.e., random chance).

(*Continued*)

Table 3
(Continued)

Claims Derived with Stakeholder Involvement	Warrants at an Associated Stage of AI-Assessment	Assumptions	Evidence
4. ML predictions can be used in employment prescreening production environments in a manner that satisfies U.S. discrimination policy.	a. Hireability decisions are based on the ML predictions alone, i.e., without any knowledge of gender. (Decision/evaluation stage)	Decision makers have access to, and make decisions on, final assessment scores only.	The top *n* candidates are selected by employers using only assessment scores.*
	b. Prescreening decisions based on the ML predictions adhere to U.S. legal requirements related to the 4/5ths rule. (Decision/evaluation stage)	The adverse impact ratio between women and men is .8 on average for 100 independently trained ML models using different training and test data partitions (i.e., cross-validation).	The adverse impact ratio between women and men is acceptable on average for 100 independently trained ML models using different training and test data partitions (i.e., cross-validation). The average AI = .85, which falls within the acceptable 4/5ths rule per U.S. law (AI threshold of .8).

*Included as a hypothetical example; this evidence has not yet been gathered by the assessment team.

evidence in Table 3. We currently are considering three rebuttals that we believe need to be addressed in our fairness argument:

1. Rebuttal to all evidence in the argument: What happens to fairness with respect to participants who were excluded in the AVI assessment? We had a small sample of non-binary participants, preventing us from gathering evidence of fairness across these groups. This limits the fairness argument to those holding binary gender identities, so future work needs to be done to overcome this limitation. Additionally, the number of women participants in the original sample was much greater than men, suggesting the presence of self-selection and sampling biases which threaten the generalizability of the AI assessment and fairness findings. Addressing these issues would, for example, entail collecting more representative data from these groups.
2. Rebuttal to Claim 2's evidence for the Decision/Evaluation Stage warrant: The Spearman's correlation of .4 is substantially lower than would be desired if this fairness argument were applied to a high-stakes AI screening assessment, where perhaps .7 or higher would be more reasonable. Achieving this level of hireability prediction accuracy may require substantial changes to the ML model, which may in turn affect the evidence pertaining to other claims in the fairness argument.
3. Rebuttal to Claim 3's evidence for the ML model/prediction stage warrant: Some of the extracted features were derived from external sources, some of which were AI-based assessments and may have their own bias concerns (e.g., less accurate feature computation for women vs. men). How can we evaluate bias in external AI systems used within the AVI?

We believe that by working with stakeholders and addressing all above and ongoing rebuttals, the AVI can eventually have strong support for the overall fairness argument: The proposed AVI assessment is inclusive of gender-based experiences of women and men and does not disadvantage candidates based on their identified gender as women or men. And, in making additional fairness arguments and supporting them through the proposed argument-based fairness process, AVI developers can work toward having a set of well-supported fairness arguments that stakeholders approve and can be confidently provided to future assessment users and to the public. Hence, the proposed approach to fairness allowed the assessment developers to express concrete fairness goals and systematically work to achieve them, which contrasts heavily to trying to work toward an ill-defined goal of having a "fair AI assessment."

Conclusion

The desire to develop "fair assessments" is clear in the research literature in education, psychology, and beyond. However, fairness is a social construct, which poses two major issues for aiming to develop a "fair assessment." First, social constructs are subjective by nature. Hence, there really is no such thing as a fair assessment, as anyone could look at the assessment or system and decide that it is not fair from their perspective. Second, for the same reason, two or more people can enter into

a project to develop a "fair assessment" only to realize that they have very different goals. Putting these two related issues together, AI assessment developers, and all assessment developers in general, searching for a set of procedures or methods that can be implemented to achieve "fair AI assessments" are likely to continue this search indefinitely. One may say they are searching for a holy grail (Davies, 2010).

In this manuscript, we propose an argument-based fairness approach to define and work toward fairness goals through fairness arguments developed in conjunction with stakeholders of an assessment. We designed the approach to be used with any type of assessment, and then we applied it to the context of AI assessment. The approach encourages developers and stakeholders to define what fairness means to them in arguments but also requires that the fairness goal be achievable by building claims and warrants within arguments that are supported by evidence. One critical result of this process is that an assessment community can be specific about arguments that they think are supported without overreaching toward an impossible goal of being able to state that fairness has been objectively achieved in the assessment system.

Fairness arguments are not new in the field of education assessment (Chapelle, 2021; Xi, 2010). However, we do not know of any assessment researchers or developers who have worked toward supporting formal fairness arguments that are not couched within and constrained by validity arguments. We discussed above why we believe there are advantages to making and supporting fairness arguments independent of, and ideally complementary to, validity arguments. The case study we presented demonstrated the potential to reap some of these advantages.

1. Fairness arguments should be flexible beyond the validity arguments made for an assessment, allowing goals for fairness to change as views on fairness change and society changes, and ensuring that fairness is not constrained to be secondary to validity or to be a matter of comparable validity only. Our case study showed that we can incorporate both traditional (e.g., AERA et al., 2014) and new (e.g., Randall, 2021) considerations of fairness within a single framework.
2. Assessment developers and stakeholders should be able to work toward fairness goals through arguments even if they are not familiar or comfortable with the full framework of argument-based validity. Our case study showed that there is no need for a stakeholder to understand, for example, the different types of score inferences within argument-based validity frameworks. This is in stark contrast to fairness arguments that come in the form of rebuttals to validity arguments.

Despite these advantages, we do consider the proposed argument-based fairness process to be in its infancy. We believe there are several areas of research that could help argument-based fairness to come to maturity. Importantly, argument-based fairness needs to be applied and studied in multiple assessment environments, including AI assessment environments, with a variety of stakeholder groups. There may be, for example, particular parts of the argument structure that impede the group-based construction of a fairness argument, requiring changes to the argumentation process or changes to how it is implemented with stakeholders. For example, we believe that stakeholders of our AVI assessment can easily engage with our claims but will

struggle to evaluate the evidence under our fairness argument, as much of the evidence is technical. How can we revise the proposed approach of argument-based fairness to address this issue?

Another way in which argument-based fairness methods may be advanced through future work is to develop a mechanism for categorizing fairness arguments or claims, similar to categorizations of validity claims under various types of score inferences and uses in argument-based validity (Kane, 2013). The benefit of such categorizations to AI assessments (or any assessments) could be twofold. First, AI assessment programs may be able to borrow fairness arguments and claims from other AI assessment programs, based on common desires. For example, if multiple assessment programs desire to become more inclusive of gender identities and were making arguments, there may be common experiences that can be shared easily if groups are sharing a common classification structure to their fairness arguments or claims. Second, developing a set of categorizations for fairness may allow the field of AI assessments (or a different field of assessment) to come to a consensus over time as to how many and which types of fairness arguments tend to satisfy the needs of assessment stakeholders. The proposed argument-based fairness approach is designed to be flexible, but that does not mean that there might not be some ways to categorize the work that can generalize and be useful to others trying to make similar and comprehensive sets of fairness arguments.

In addition, it may be helpful (or even necessary) to align fairness arguments to a particular theory on fairness or to a set of fairness standards or guidelines. In this manuscript, we have introduced fairness arguments in a broad way to demonstrate their flexibility and focus on the overall concept and its importance, and we have encouraged working with stakeholders for the development of the arguments. However, in practice and in future development of a framework for fairness arguments, it seems that aligning such arguments to an external standard may bring some practical benefits. For example, if we were to align the case study argument to fairness issues as conceptualized in the *Principles for the Validation and Use of Personnel Selection Procedures* (Society for Industrial and Organizational Psychology, 2018), it may be easier to decide which fairness arguments are most important, when a fairness argument is complete, and how a fairness argument can be structured to speak to a particular industry or professional community. Similarly, in the reading assessment shown in Tables 1 and 2, it would likely be easier to evaluate the strength of the full argument, once it is complete, if the claims were aligned to a particular theory on culturally responsive assessment. Thus, while fairness arguments can be flexible, aligning an argument to a set of fairness standards or theory can bring some coherence to an otherwise endless list of possible arguments and claims.

Ultimately, fairness in assessment has been an elusive goal for quite some time, and we agree with Lee Cronbach (1976) in his statement about fairness in selection processes based on assessment scores: "Make no mistake. The issues will not be settled by mathematical specialists" (p. 31). Our proposed argument-based fairness approach acknowledges that not all aspects of fairness will come down to statistical challenges, that fairness claims likely need support from a variety of evidence types (e.g., statistical evidence; qualitative evidence), and that assessment stakeholders holding a variety of expertise inside and outside of psychometrics need

to be involved if we are to achieve some version of fairness in any assessment. With continued research and applications of this process across the assessment community, we are hopeful that we can develop a mature framework for argument-based fairness that allows for achievable strides toward fairness in assessment programs, including but not limited to those using AI.

Acknowledgment

The research reported here was supported by the Institute of EducationSciences, U.S. Department of Education, through Grant R305A190079 to theUniversity of Florida. The opinions expressed are those of the authors and donot represent views of the Institute or the U.S. Department of Education.

Notes

[1]Recall this was expected given the irrelevance of gender to the hireability construct in this case study. This warrant and its associated claims and evidence may not hold in other pre-employment screening studies involving a job where, for example, traits such as agreeableness with known gender differences (Weisberg et al., 2011) were deemed important for job performance (Sackett & Walmsley, 2014). In this scenario, a better warrant might be that the mean differences in ground truth distributions for men and women are consistent with known differences from prior research, and it would still support the claim about gender inclusiveness. Similarly, the direction and degree of gender differences in the AI predictions should match what is observed in the ground truth using similar metrics (e.g., Cohen's *d*, Kolmogorov-Smirnov test).

[2]We note that this may not be true in our particular case study due to the rating protocol employed or because the ground truth measure is based on perceived hireability rather than knowledge, skills, abilities, and other characteristics (Society for Industrial and Organizational Psychology, 2018).

[3]In a different scenario involving fairness arguments for multiple groups (e.g., race and gender) where membership is irrelevant to the construct (e.g., hireability), it may still be important for each group to be represented in equal quantities, otherwise the ML model may learn to accept a higher assessment error rate for minority subgroups (e.g., Black women) over others (e.g., White men) to reduce the overall error rate. Stakeholders should be involved in determining whether this is necessary to achieve the (subjective) fairness goals.

References

American Educational Research Association (AERA), American Psychological Association (APA), and National Council on Measurement in Education (NCME). (1999). *Standards for educational and psychological testing*. Washington, DC: Author.

American Educational Research Association (AERA), American Psychological Association (APA), and National Council on Measurement in Education (NCME). (2014). *Standards for educational and psychological testing*. Washington, DC: Author.

Benjamin, R. (2019). Assessing risk, automating racism. *Science*, *366*(6464), 421–422.

Booth, B. M., Hickman, L., Subburaj, S. K., Tay, L., Woo, S. E., & D'Mello, S. K. (2021a). Integrating psychometrics and computing perspectives on bias and fairness in affective computing: A case study of automated video interviews. *Signal Processing Magazine*, Special Issue on Affective Computing. IEEE, 2021.

Booth, B. M., Hickman, L., Subburaj, S. K., Tay, L., Woo, S. E., & D'Mello, S. K. (2021b). Bias and fairness in multimodal machine learning: A case study of automated video interviews. Paper presented at the annual meeting of the International Conference on Multimodal Interaction in Montreal, Canada.

Buolamwini, J., & Gebru, T. (2018). Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Proceedings of the 1st Conference on Fairness, Accountability and Transparency*, PMLR 81:77–91, 2018.

Burstein, J., LaFlair, G. T., Kunnan, A. J., & von Davier, A. A. (2021). A theoretical assessment ecosystem for a digital-first assessment—The Duolingo English Test. (Duolingo Research Report DRR-21-04). Retrieved from englishtest.duolingo.com/research.

Campion M. A., Palmer D. K., & Campion, J. E. *(*1997*)*. A review of structure in the selection interview. *Personnel Psychology*, *50*, 655–702.

Chapelle, C. A. (2021). *Argument-based validation in testing and assessment*. Thousand Oaks, CA: Sage.

Chapelle, C., Enright, M., & Jamieson, J. (Eds.) (2008). *Building a validity argument for the test of English as a foreign language*. NY: Routledge.

Cohen, J. (1992). A power primer. *Psychological Bulletin*, *112*(1), 155.

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

Davies, A. (2010). Test fairness: A response. *Language Testing*, *27*, 171–176.

D'Mello, S. K., Tay, L., & Southwell, R. (2021). Psychological measurement in the information age: Machine-learned computational models. *Current Directions in Psychological Science*, 31, 76–87.

Educational Testing Service (ETS). (2002). *ETS standards for quality and fairness*. Princeton, NJ: Author.

Frey, B. (Ed.) (2018). *The SAGE encyclopedia of educational research, measurement, and evaluation*. Thousand Oaks, CA: Sage.

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*, online first.

HireVue (2019, May). *HireVue Surpasses Ten Million Video Interviews Completed Worldwide* [Press release]. Retrieved from https://www.hirevue.com/press-release/hirevue-surpasses-ten-million-video-interviews-completed-worldwide

Hübner, D. (2021). Two kinds of discrimination in AI-based penal decision-making. *ACM SIGKDD Explorations Newsletter*, *23*(1), 4–13.

Huggins-Manley, A. C. (2021, November). Argument-based fairness in educational assessment. Presentation at the annual meeting of the Florida Educational Research Association in St. Petersburg, FL.

Joint Committee on Testing Practices (2004). *Code of fair testing practices in education*. Washington, DC: Author.

Jonson, J. L., Trantham, P., & Usher-Tate, B. J. (2019). An evaluative framework for reviewing fairness standards and practices in educational tests. *Educational Measurement: Issues and Practice*, *38*, 6–19.

Kane, M. T. (1992). An argument-based approach to validity. *Psychological Bulletin*, *112*, 527–535.

Kane, M. T. (2010). Validity and fairness. *Language Testing*, *27*, 177–182.

Kane, M. T. (2013). Validating the interpretations and uses of test scores. *Journal of Educational Measurement*, *50*, 1–73.

Koo, T. K., & Li, M. Y. (2016). A guideline of selecting and reporting intraclass correlation coefficients for reliability research. *Journal of Chiropractic Medicine*, *15*, 155–163.

Kunnan, A. J. (1997). Connecting validation and fairness in language testing. In A. Huhta (Ed.), *Current developments and alternatives in language assessment* (pp. 85–105). Jyväskylä, Finland: University of Jyväskylä.

Kunnan, A. J. (2000). Fairness and justice for all. In A. J. Kunnan (Ed.), *Fairness and validation in language assessment* (pp. 1–14). Cambridge, UK: Cambridge University Press.

Kunnan, A. J. (2004). Test fairness. In M. Milanovic & C. Weir (Eds.), *European language testing in a global context: Proceedings of the ALTE Barcelona Conference* (pp. 27–48). Cambridge, UK: Cambridge University Press.

Kunnan, A. J. (2010). Test fairness and Toulmin's argument structure. *Language Testing*, *27*, 183–189.

Kizilcec, R. F., & Lee, H. (forthcoming). Algorithmic fairness in education. In W. Holmes & K. Porayska-Pomsta (Eds.), *Ethics in artificial intelligence in education*. London: Taylor & Francis.

Le Cun, Y., Bengio, Y., & Hinton, G. E. (2015). Deep learning. *Nature*, *521*, 436–444.

National Artificial Intelligence Initiative Act, H.R. 6216, 116th Cong. (2020). https://www.congress.gov/116/bills/hr6216/BILLS-116hr6216ih.pdf

Oliveri, M. E., Lawless, R., & Young, J. W. (2015). *A validity framework for the use and development of exported tests (ETS Office of Professional Standards Series)*. Princeton, NJ: Educational Testing Service.

Mislevy, R. J., & Haertel, G. D. (2006). Implications of evidence-centered design for educational testing. *Educational Measurement: Issues and Practice*, *25*(4), 6–20.

Pennebaker, J. W., Francis, M. E., & Booth, R. J. (2001). *Linguistic inquiry and word count: LIWC*. Mahwah, NJ: Lawrence Erlbaum Associates.

Pugh, S. L., Subburaj, S. K., Rao, A. R., Stewart, A. E., Andrews-Todd, J., & D'Mello, S. K. (2021). Say what? Automatic modeling of collaborative problem solving skills from student speech in the wild. International Educational Data Mining Society, Paper presented at the International Conference on Educational Data Mining (EDM) (14th, Online, Jun 29–Jul 2, 2021).

Randall, J. (2021). "Color-neutral" is not a thing: Redefining construct definition and representation through a justice-oriented critical antiracist lens. *Educational Measurement: Issues and Practice*, online first: https://doi.org/10.1111/emip.12429

Sackett, P. R., & Walmsley, P. T. (2014). Which personality attributes are most important in the workplace? *Perspectives on Psychological Science*, *9*, 538–551.

Shute, V., & Ventura, M. (2013). *Measuring and supporting learning in games: Stealth assessment*. The MIT Press.

Sireci, S. G. (2020). Standardization and *UNDERSTAND*ardization in educational assessment. *Educational Measurement: Issues and Practice*, *39*, 100–105.

Society for Industrial Organizational Psychology (2018). Principles for the validation and use of personnel selection procedures. *Industrial and Organizational Psychology*, *11*, 1–97. https://doi.org/10.1017/iop.2018.195

Tay, L., Woo, S., Hickman, L., Booth, B. M., & D'Mello, S. K. (2022). A conceptual framework for investigating and mitigating machine learning bias. *Advances in Methods and Practices in Psychological Science*. https://doi.org/10.1177/25152459211061337

Toulmin, S. (1958/2003). *The uses of argument*. Cambridge, UK: Cambridge University Press.

von Davier, A. A., Deonovic, B., Yudelson, M., Polyak, S. T., & Woo, A. (2019). Computational psychometrics approach to holistic learning and assessment systems. *Frontiers in Education*, *4*, 1–12.

Weisberg, Y. J., DeYoung, C. G., & Hirsh, J. B. (2011). Gender differences in personality across the ten aspects of the Big Five. *Frontiers in Psychology*, *2*, 178.

Willingham, W. W., & Cole, N. (1997). *Gender and fair assessment*. Mahwah, NJ: LawrenceErlbaum.

Xi, X. (2010). How do we go about investigating test fairness? *Language Testing*, *27*, 147–170.

Yan, D., Rupp, A. A., & Foltz, P. W. (2020). *Handbook of automated scoring: Theory into practice*. Boca Raton, FL: Chapman and Hall/CRC.

Authors

CORINNE HUGGINS-MANLEY is Associate Professor of Research and Evaluation Methodology at the University of Florida, 1602 Norman Hall, Gainesville, FL 32611; amanley@coe.ufl.edu. Her research interests include fairness and validity in educational measurement.

BRANDON BOOTH is a postdoctoral research associate in the Emotive Computing lab within the Institute of Cognitive Science at the University of Colorado Boulder, 594 UCB – Boulder, CO 80309; brandon.m.booth@gmail.com. His research focuses on using multi-modal machine learning techniques to model human perception, behavior, and experiences and developing frameworks to reduce the impact of inadvertent human biases and errors.

SIDNEY DMELLO is a Professor at the Institute for Cognitive Science and Department of Computer Science at the University of Colorado Boulder, 594 UCB, Boulder CO 80309; sidney.dmello@colorado.edu. His research interests include developing fair and equitable technologies to improve assessment and learning outcomes.