The State of the Situation and Policy Recommendations for Algorithmic Bias
By Ryan S. Baker, M. Aaron Hawn, Seiyon Lee

Abstract: This chapter discusses the current state of the evidence on algorithmic bias in education. After defining algorithmic bias and its possible origins, it reviews the existing international evidence about algorithmic bias in education, which has focused on gender and race, but has also involved some other demographic categories. The chapter concludes with a few recommendations, notably to ensure that privacy requirements do not prevent researchers and developers from identifying bias, so that it can be addressed.

1. Introduction

Concern about the problem of algorithmic bias has increased in the last decade. Algorithmic bias occurs when an algorithm encodes (typically unintentionally) the biases present in society, producing predictions or inferences that are clearly discriminatory towards specific groups (Executive Office of the President, 2014; O'Neil, 2017; Zuiderveen Borgesius, 2018). This concern has emerged across domains from criminal justice (Angwin et al., 2016), to medicine (O'Reilly-Shah et al., 2020), to computer vision (Klare et al., 2012), to hiring (Garcia, 2016).

Research has demonstrated that algorithmic bias is a problem for algorithms used in education as well. Academics have been warning about possible uneven effectiveness and lack of generalizability across populations in educational algorithms for several years (e.g. Bridgeman et al., 2009; Ocumpaugh et al., 2014). In education, algorithmic bias can manifest in several ways. For instance, an algorithm used in testing to identify English language proficiency may systematically underrate the proficiency of learners from some countries (Wang et al., 2018; Loukina et al., 2019), denying them access to college admission. To give another example, an algorithm identifying if learners are at risk of failing a course may underestimate the risk of learners in specific demographic groups (Hu & Rangwala, 2020; Kung & Yu, 2020; Yu et al., 2020), denying them access to needed support.

This concern has led to increasing interest in addressing algorithmic bias in education, both in academia and industry. A rapidly increasing number of publications in this area is a testimony to the increasing academic interest in this topic. Active, even fervent debate, is ongoing about how best to measure algorithmic bias (Caton & Haas, 2020; Mehrabi et al., 2021; Verma & Rubin, 2018) and which technical approaches can correct bias (Kleinberg et al., 2017; Loukina et al., 2019; Lee & Kizilcec, 2020). Within industry and the NGO sector, efforts such as the Prioritizing Racial Equity in AI Design Product Certification from Digital Promise (Digital Promise, 2022) demonstrate the efforts being made to systematize the process of reducing algorithmic bias, and several companies have actively published evidence about the algorithmic bias in their tools and platforms, sometimes in cooperation with academics (Bridgeman et al., 2009, 2012; Christie et al., 2019; Zhang et al., 2022). There has not yet been comparable interest in addressing

algorithmic bias in education within the policy space -- if anything, current directions in policy are towards adopting privacy regulations that will make it impossible to fix the problem of algorithmic bias in education, by making it impossible to collect the data needed to identify if it is occurring and to apply common methods for fixing it when it occurs (see review of this issue in Baker, in press).

Despite the increasing concern about algorithmic bias in education, however, work to determine its scope and address it has remained limited. While an increasing number of papers look into algorithmic bias in education, as this review will illustrate, this research is highly uneven in focus. The overwhelming majority of work on algorithmic bias in education focuses on the impacts on a small number of racial and ethnic groups and on sex (Baker & Hawn, 2022), with most effort going into the demographic variables that are most conveniently available to researchers (Belitz et al., 2022). Work in this area is also extremely focused on algorithms used in a single country, the United States of America (Baker & Hawn, 2022). The work that exists shows clear evidence that groups already disadvantaged societally are further disadvantaged by current educational technologies, a problem that requires action. But we do not yet know the full extent of the problem.

In this chapter, we discuss the current state of the evidence on algorithmic bias in education, key obstacles to creating fair algorithms, and steps that can be taken to surpass these obstacles. We conclude with recommendations for policy-makers for what they can do to help resolve this still mostly-hidden societal problem.

2. What is Algorithmic Bias?

2.1 Defining algorithmic bias

A recent survey across 146 papers found a lack of clarity in how authors define and use the term *bias*, from gaps of explanation as to how exactly systems are biased to confusion about the eventual harms that bias might cause (Crawford, 2017; Blodgett et al., 2020). We will briefly discuss some of the issues in defining *algorithmic bias* before proposing a limited working definition applied in this survey.

Algorithmic bias in emerging use

The term *algorithmic bias* has been used to describe many problems of fairness in automated systems, only some of which map onto statistical or technical definitions of bias. Some researchers define the term broadly, referring to *biases* as the set of possible harms throughout the machine learning process, including any “unintended or potentially harmful” properties of the data that lead to “unwanted or societally unfavorable outcome[s]” (Suresh & Guttag, 2020, p. 1-

2). Others apply *algorithmic bias* in a more limited way to cases where a model's performance or behavior differs systematically between groups (Gardner et al., 2019; Mitchell et al., 2021). This second definition of algorithmic bias -- systematic skew in performance -- may or may not lead to harmful disparate impacts or discrimination, depending on how model results are applied.

Because of this potential for algorithmic bias to translate into unintended impacts, the machine learning process should be conducted with caution, anticipating some of the very real damages that may result from bias. A widely accepted framework for such harms categorizes them broadly into allocative and representational forms (Crawford, 2017; Suresh & Guttag, 2020).

Allocative harms result from the withholding or the unfair distribution of some opportunity across groups, with examples including gender bias in assigning credit limits (Knight, 2019; Telford, 2019); racial bias in sentencing decisions (Angwin et al., 2016), racial bias in identifying patients for additional health care (Obermeyer et al., 2019), and -- in education -- bias in standardized testing and its resulting impact on high stakes admission decisions (Dorans, 2010; Santelices & Wilson, 2010).

Representational harms, on the other hand, manifest as the systematic representation of some group in a negative light, or by withholding positive representation (Crawford, 2017). Multiple forms of representational harm have been uncovered in recent years, with Sweeney (2013) identifying varieties of *denigration* and *stereotyping*, where, for instance, the word "criminal" was more frequently returned in online ads after searches for black-identifying first names.

While there are clearly a range of ways that *algorithmic bias* is discussed, here we focus on algorithmic bias in situations where model performance is substantially better or worse across mutually exclusive groups (i.e. Gardner et al., 2019; Mehrabi et al., 2021; Mitchell et al., 2021). Other forms of algorithmic bias (such as the cases mentioned above) can be highly problematic, but -- as we discuss below -- the published research in education thus far has focused on this performance-related version of bias. In this review, we also home in on bias in algorithms, excluding the broader design of the learning or educational systems that use these algorithms. Bias can also emerge in the design of learning activities, leading to differential impact for different populations (Finkelstein et al., 2013), but that is a much broader topic, beyond what this review covers.

How this type of algorithmic bias is identified

Though the origins of algorithmic bias are complex, and fixing it can in some cases be challenging, identifying this form of algorithmic bias is relatively straightforward. Doing so requires only two steps: 1) obtaining data on student identity; 2) checking model performance for students belonging to different groups.

The first step poses some challenges in terms of concerns around student privacy (Pardo & Siemens, 2014) and policies designed to protect student privacy (Baker, in press). If data on student identity and membership in key demographic groups was not collected initially, it can

be difficult to collect after the fact.

Once the data has been split into members of different groups, and the model has been applied to those learners, the results can be checked for differences in performance. There are a range of measures that can be used (Kizilcec & Lee, 2022), and ideally several will be used in concert. First, the same measures generally used to evaluate algorithm performance -- AUC ROC, Kappa, F1, precision, recall, and so on -- can also be used to evaluate performance for sub-groups. Second, some measures specific to algorithmic bias analysis -- ABROCA (Gardner et al., 2019), independence, separation, sufficiency, for instance -- can be applied.

After examining the metrics for the differences in algorithm performance between groups, it becomes possible to analyze the expected impacts, anticipating the ways that algorithmic bias might lead to a biased response or intervention. For example, if an algorithm for predicting high school drop-out achieves 20% poorer recall (the ability to identify all individuals at risk) for members of a historically disadvantaged group, then we know that many students in the group who are at risk and need an intervention will not receive it. By contrast, if the same algorithm were to achieve 20% poorer precision (the ability to avoid selecting an individual not at risk) for members of a historically disadvantaged group, then many students in the group will receive unnecessary interventions, at best wasting their time. Checking for expected impacts also gives a sense of what would be gained by fixing a specific bias identified, and ensures that work spent to address algorithmic biases, if successful, will increase the fairness and overall benefit of using the algorithm.

Bias against whom?

Researchers have considered a range of groups which have been, or could be, impacted by algorithmic bias. Many of these groups have been defined by characteristics protected by law. In the United Kingdom, for instance, the Equality Act of 2010 merged over a hundred disparate pieces of legislation into a single legal framework, unifying protections against discrimination on the basis of sex, race, ethnicity, disability, religion, age, national origin, sexual orientation, and gender identity. In the United States, the same categories are protected by a combination of different legislation, commission rulings, and court rulings, dating back to the Civil Rights Act of 1964. Similar laws afford protections in the European Union and most other countries around the world, though differing in which groups are protected and how they are defined.

While preserving fairness for these legally-defined groups is critical, looking for bias only under the lamppost of nationally protected classes (categories with their own complicated histories) may leave other, under-investigated, groups open to bias and harm. Other researchers have suggested additional characteristics which may be vulnerable to algorithmic bias in education: urbanicity (Ocumpaugh et al., 2014), military connected status (Baker et al., 2020), or speed of learning (Doroudi & Brunskill, 2019). Existing legal frameworks used to decide which classes of people merit protection from discrimination may be helpful in assessing the unknown risks that algorithmic bias poses to less studied or unidentified groups (Soundarajan & Clausen, 2018). Section 4 reviews the limited education research into algorithmic bias associated with other

groups.

2.2 Origins of bias and harm in the machine learning pipeline

In an effort to better catalog the origins of algorithmic bias, researchers have described the stages of the machine learning lifecycle alongside the kinds of bias and harm that can arise at each stage (Barocas et al., 2019; Friedman & Nissenbaum, 1996; Hellström et al., 2020; Mehrabi et al., 2021; Silva & Kenney, 2018; Suresh & Guttag, 2020). While Some authors collapse the machine learning process into broader stages (e.g., *measurement*, *model learning*, and *action)* (Barocas et al., 2019; Kizilcec & Lee, 2020), others delimit finer-grained stages, such as *data collection*, *data preparation*, *model development*, *model evaluation*, *model post-processing*, and *model deployment* (Suresh & Guttag, 2020). Industry researchers, in turn, have offered additional stages more common to applied contexts, such as *Task Definition*, *Dataset Construction*, *Testing Process*, *Deployment*, and ongoing *Feedback* from users (Cramer et al., 2019).

At each of these stages, particular forms of bias can arise. Examples include Historical bias, Representation bias, Measurement bias, Aggregation bias, Evaluation bias, and Deployment bias (Suresh & Guttag 2020). By grounding aspirational, goal-driven algorithms in data from an historically inequitable world, Historical bias is often perpetuated in education. The most common example, perhaps, is using student demographics as a feature to increase model performance, with the result of lowering the predicted grades for some students based on membership in a demographic group (i.e. Wolff et al., 2013). A recent survey of the role of demographics in educational data mining, finds that roughly half of papers incorporating demographics into models as features risk this form of bias, using at least one demographic attribute as a predictive feature without considering demographics during model testing or validation (Paquette et al., 2020).

Representational bias occurs when groups under-sampled in training data receive lower-performing predictions. Measurement bias occurs when the selected variables lack construct validity in a way that leads to unequal prediction across groups. A model predicting school violence, for example, might be biased if the labeling of which students engage in violence involves prejudice – e.g. the same violent behavior is documented for members of one race but not for members of another (Bireda, 2002).

Past the data collection stages of machine learning, the model learning phase is susceptible to *aggregation bias*, when training data from distinct populations are combined, with the resulting model working less well for some -- or all -- groups of learners (Suresh & Guttag, 2020). When detectors of student emotion, for instance, were trained on a combination of urban, rural, and suburban students, they functioned more poorly for all three groups than detectors trained on individual groups (Ocumpaugh et al., 2014). In the application phases of machine learning, *evaluation bias* occurs when the test sets used to evaluate a model fail to represent the populations with which the model will be applied, and *deployment bias* occurs when a model designed for one purpose is used for other tasks, such as applying a model designed to help

teachers identify student disengagement as a tool to assign summative participation grades to students.

Increasing research and journalism has exposed these forms of algorithmic bias in areas such as at-risk prediction for dropping out of high school or college (Anderson et al., 2019), at-risk prediction for failing a course (Hu & Rangwala, 2020; Lee & Kizilcec, 2020), automated essay scoring (Bridgeman et al., 2009, 2012), assessment of spoken language proficiency (Wang et al., 2018), and the detection of student emotion (Ocumpaugh et al., 2014). In these cases and others, reviewed below, algorithmic bias has impacted educational algorithms in terms of student race, ethnicity, nationality, gender, native language, urbanicity, parental educational background, socioeconomic status, and whether a student has a parent in the military. This evidence has prompted increasing academic and industry research into the ways that algorithmic bias can be more effectively identified, mitigated, and its harms reduced.

2.3 Mitigating bias by formalizing fairness

Much current work addressing algorithmic bias has focused on mitigation at the model evaluation and postprocessing stages of the machine learning process. Recent surveys present several taxonomies and definitions of fairness with related metrics (Barocas et al., 2019; Caton & Haas, 2020; Kizilcec & Lee, 2020; Mehrabi et al., 2021; Mitchell et al., 2021; Verma & Rubin, 2018). While these formalized metrics make a clear contribution to clarifying algorithmic bias, their application has revealed obstacles. Specifically, technical challenges to the use of fairness metrics manifest in several “impossibility” results (Chouldechova, 2017; Kleinberg et al., 2017; Berk et al., 2018; Loukina et al., 2019; Lee & Kizilcec, 2020, Darlington, 1971), where satisfaction of one statistical criterion of fairness makes it impossible to satisfy another. For instance, Kleinberg et al. (2017) demonstrate that it is mathematically impossible under normal conditions for a risk estimate model to avoid all three of the following undesirable properties: 1) systematically skewing upwards or downwards for one demographic group; 2) assigning a higher average risk estimate to individuals not at risk for one group than the other; 3) assigning a lower average risk estimate to individuals who are at risk in one group than the other.

Other challenges for this pathway to mitigating bias include the difficulty in describing optimal tradeoffs in fairness for domain-specific problems (Lee & Kizilcec, 2020; Makhlouf et al., 2020; Suresh & Guttag, 2020), as well as the sociotechnical critique that an overemphasis on seemingly objective, statistical criteria for fairness may provide an excuse for developers and users of algorithms to avoid grappling with the full range of potential bias and harms from employing algorithms for high-stakes decisions (Green, 2020; Green & Hu, 2018; Green & Viljoen, 2020). In order to address the fuller picture of algorithmic bias, it is critical to identify and mitigate bias, not only during the later stages of the process, but also during the earlier stages of data collection and data preparation.

2.4 Representational and measurement biases: the key role for data collection

Attempts to address algorithmic bias solely by adjusting algorithms may be ineffective if we have not collected the right data. Specifically, representational and measurement bias (Suresh & Guttag 2020) can prevent methods further down the pipeline from resolving, or even detecting, bias.
As a key example, If we collect training data only from suburban upper middle-class children, we should not expect our model to work for urban lower-income students. More broadly, if we do not collect data from the right sample of learners, we risk representational bias and cannot expect our models to work for all learners.

Measurement bias is another significant challenge that improved metrics or algorithms cannot overcome on their own. While measurement bias can occur in both predictor variables and training labels (Suresh & Guttag, 2020), the most concerning cases involve the latter. If, for instance, Black students behave similarly to students from other groups, but are still more likely to be *labeled* in a dataset as engaging in school violence, then it is difficult to determine whether an algorithm works equally well for both groups, or to be at all confident that the algorithm's functioning is not biased. Surprisingly, this bias in training labels may even come from students themselves if the label depends on students' responses and can be impacted by confidence, cultural interpretation, or stereotype threat (Tempelaar et al., 2020). In these cases, finding an alternate variable to predict -- one not as impacted by bias -- may be the best alternative. Other cases of Measurement bias may be easier to mitigate, such as when human coders, impacted by their own bias (Kraiger & Ford, 1985; Okur et al., 2018), label some aspect of previously collected data. In the situation where predictor variables are biased, they may be substituting for other variables that would explicitly define group membership, in which case it may be best to discard the biased predictors from consideration.

Ultimately, the best path to addressing both representational and measurement bias is to collect better data -- data that includes sufficient proportions of relevant groups, and in which key variables are not themselves biased (Holstein et al., 2019). Completing this task, however, depends on knowing what groups are critical to represent in the data sets used to develop models, the focus of our next section.

3. Algorithmic Bias: Impact on Students in Common Demographic Categories

A great majority of research has focused on a limited number of groups within the diverse student population, focusing on variables involving race and ethnicity, nationality, and gender (Baker and Hawn, 2022). Race and ethnicity, nationality and gender, unsurprisingly, represent

the most common demographic categories or variables collected by or made available to the researchers, whether by convention or for convenience, especially as most research was conducted in the United States.

Within these broad categories, there is some variance in how the variables are considered. At times, specific racial groups are considered and other times they are considered in terms of whether a student is an URM (Under-Represented Minority) or not. Although a minority in most studies, Asians are typically treated as non-URM in US educational research. Even when racial groups are separated in analysis, heterogeneity within these groups is typically ignored (i.e. people whose ancestors have lived in their current country for generations versus recent immigrants; individuals with different national origins with very different histories and cultures; Baker et al., 2019).

In this section, we will examine the evidence on algorithmic bias in education by addressing which groups of students have been systematically impacted, in terms of these most common categories. The overview will be organized into the different locations in the world in which each study was conducted, in order to illustrate the uneven amount of research on algorithmic bias in education that has occurred in different regions. We will discuss the implications of that unevenness, and how to address it, later in this chapter.

Algorithmic Bias in Education in the USA (Widely-Studied Categories)

The majority of research on algorithmic bias in education thus far has been conducted in the USA. The strong interest in documenting and addressing algorithmic bias in the USA maps to broader societal concern in the United States about algorithmic bias (Corbett-Davies & Goel, 2018) and discrimination in general (Barocas et al., 2019; O'Neil, 2017). It also may reflect the relatively high availability of educational data for research purposes in the United States. Even most of the research on how learners from different nationalities are impacted by research on algorithmic bias has often been conducted in the United States (Bridgeman et al., 2009, 2012; Li et al., 2021; Ogan et al., 2015; Wang et al., 2018).

Within the United States, a considerable amount of research has investigated the impact of algorithmic bias in education on different racial groups. A recent review by Baker and Hawn (2022) identifies ten cases where this was investigated, across algorithms for purposes ranging from predicting dropout (Anderson et al., 2019; Christie et al., 2019; Kai et al., 2017; Yu et al., 2021), predicting course failure (Lee & Kizilcec, 2020; Yu et al., 2020), and automated essay scoring (Bridgeman et al., 2009, 2012; Ramineni & Williamson, 2018). Typically, across studies, algorithms were less effective for Black and Hispanic/Latino students in general (Anderson et al., 2019; Bridgeman et al., 2012; Lee & Kizilcec, 2020; Ramineni & Williamson, 2018; Yu et al., 2021), and often also had different profiles of false positive and negative results for students in different racial groups (Anderson et al., 2019). More recently, the Penn Center for Learning

Analytics (PCLA) wiki (PCLA, 2022) has identified an additional six studies (published since Baker & Hawn, 2022 was finalized) on this topic. Curiously, smaller effects were seen in many of these more recent studies than in earlier studies, which may suggest either that there were some “file drawer” problems with earlier work (that is, results with small effects were not published in the past), or that a broader range of possible contexts are being investigated. Though there has been considerable attention to race in general, less quantity of research has been paid to indigenous learners, often due to issues of sample size (Anderson et al., 2019), though notable counter-examples exist (e.g. Christie et al., 2019).

Within the United States, considerable research has also investigated the impact of algorithmic bias in education in terms of learners with different genders. Baker and Hawn (2022) identified nine cases, and three additional papers have been identified since then by the PCLA wiki. Across these papers, gender effects were highly inconsistent, with significant biases against females in some cases (Gardner et al., 2019; Yu et al., 2020) and significant biases against males in other cases (Hu & Rangwala, 2020; Lee & Kizilcec, 2020; Kai et al., 2017).

Algorithmic Bias in Education in Europe (Widely-Studied Categories)

While race has been used as a predictor variable in Europe (Wolff et al., 2013), it has not been the subject of systematic investigation into algorithmic bias in education. Research into how algorithmic bias impacts learners from different nationalities has also not been carried out in Europe, to the best of our knowledge, although Bridgeman and colleagues (2009, 2012) investigated algorithmic bias in automated essay scoring on learners from around the world, including several European countries, finding that learners from European countries were less impacted than learners in Asia. However, Wang et al. (2018) found substantial inaccuracies in speech evaluation for learners from Germany, and Li et al. (2021) found that academic achievement prediction was less effective for learners from Moldova than learners in wealthier countries.

However, research on algorithmic bias in terms of gender has occurred in Europe: Riazy et al. (2020) investigated the impacts of gender on course outcome prediction and Rzepka et al. (2022) investigated the impacts of gender on prediction conducted during a spelling learning activity. Only small effects were found.

Overall, then, there is not yet evidence for major impacts of algorithmic bias in Europe, in terms of race, nationality, or gender, but there also have been few studies, and these studies do not cover the range of applications which research in the United States has investigated.

Algorithmic Bias in Education in the Rest of the World (Widely-Studied Categories)

Multiple studies of algorithmic bias in terms of nationality have been conducted on learners from around the world, though primarily involving researchers based in the United States. Baker & Hawn (2022) identified four such studies. These studies have involved a range of applications, from predicting academic achievement (Li et al., 2021), to automated essay scoring (Bridgeman et al., 2009, 2012), to speech evaluation (Wang et al., 2018), to models of help-seeking (Ogan et al., 2015). The studies have documented biases impacting learners from China, Korea, India, Vietnam, the Philippines, Costa Rica, and individuals living in countries where the primary language is Arabic. These studies have been fairly different from each other (except for the two Bridgeman et al. studies) and have documented a range of patterns, clearly indicating that considerably more research is needed.

Three studies on algorithmic bias in education in terms of gender have been conducted outside of the United States and Europe. Verdugo and colleagues (2022) found bias in algorithms predicting university dropout in Chile, negatively impacting female students. Sha and colleagues (2021, 2022) investigate algorithms for four different applications in Australia, finding substantial gender biases but not always in the same direction.

Examples of Algorithmic Bias in Education
(Bridgeman et al., 2012) Automated essay scoring used in a high-stakes examination (the Test of English as a Foreign Language) was found to systematically rate essays differently than human graders. Specifically, the algorithm rated native speakers of Arabic, Hindi, and Spanish lower than students from other countries, relative to human graders. The algorithm had been used to replace one of two human coders. In response to this evidence, the test developer instituted a new practice: First a single human grader and the machine rate the essay. If the human and machine give substantially different ratings, a second human rates the essay. If the two humans agree, the automated score is discarded.
(Verdugo et al., 2022) A model predicting first-year dropout from a Chilean university was found to perform more poorly for female students and students who attended private high schools. A range of fairness techniques were applied, improving the equity of model performance, and in turn the equity of the provision of dropout supports to students.
(Ocumpaugh et al., 2014) Models detecting student affect (whether a student was bored, frustrated, confused, or engaged) within an online learning platform were found to perform more poorly for students in rural communities than for students in urban or suburban communities. By creating a model tailored to rural students, model performance was improved for this group of students. The models are being used to conduct learning engineering research on how to improve the design of learning content; reducing inequities in the models reduces the risk that incorrect design decisions are made.

4. Algorithmic Bias: Impact on Students in Other Categories

While the majority of research on algorithmic bias in education has investigated race and ethnicity, nationality and gender, other categories of identity have also been investigated. In this section, we will examine the evidence for algorithmic bias in education impacting learners in these categories. Across studies, researchers have investigated algorithmic bias in terms of learners' urbanicity (city or rural area), socioeconomic status, type of school attended (public or private), native language, disabilities, parental educational background, military-connected status. These variables have generally not been investigated in sufficient detail to draw solid conclusions. As with race and ethnicity, nationality and gender, the majority of studies occurred in the USA (15 studies), compared to 3 in Europe and 2 in the rest of the world.

According to the PCLA wiki, four studies have thus far investigated native language in terms of algorithmic bias in education: two in the USA (Naismith et al., 2018; Loukina et al., 2019), one in Europe (Rzepka et al., 2022), and one in Australia (Sha et al., 2021). Three of four studies found evidence for algorithmic biases negatively impacting non-native speakers, but one (Rzepka et al., 2022) found slightly better model accuracy for non-native speakers. All four studies involved educational tasks where the use of language was central (essay-writing, speaking, spelling, and posting to discussion forums).

The PCLA wiki identified five studies on parental educational background, four in the USA and one in Europe. All five show differences in model performance and prediction in terms of this variable, but the ways that bias manifest are inconsistent across studies, with some studies finding better performance for students with more educated parents but other studies finding better performance for students with less educated parents.

The PCLA wiki also identifies five studies on socioeconomic status, all five conducted in the USA. Four of the five papers (predicting dropout, grade point average [GPA], and learning) find that algorithms are less effective for students with poorer socioeconomic backgrounds, but the fifth (on automated essay scoring) finds no evidence of difference. A sixth study conducted in Chile, on whether students are attending public or private schools (highly associated with socioeconomic status), finds that models predicting university dropout are more accurate for learners from public schools.

There has been relatively little research on how algorithmic bias impacts learners with disabilities. In the USA, Baker & Hawn (2022) document a single study, by Loukina & Buzick (2017), which found that a system for assessing proficiency in spoken English was less accurate for students who were identified by test administrators to have a speech impairment.

In Europe, Baker & Hawn (2022) also document a single study, by Riazy et al. (2020), who found that a system for predicting course outcomes had systematic inaccuracies for learners with self-declared disabilities. These two studies clearly do not cover the range of disabilities that may lead to algorithmic bias in education, and no studies have been documented outside the USA and Europe.

According to the PCLA wiki, two studies have investigated algorithmic bias in terms of student urbanicity (urban versus rural), both in the USA. Ocumpaugh and colleagues (2014) find that models predicting student emotion are less effective if they are developed using data from urban learners and then tested on data from rural learners, compared to if they are tested on data from unseen urban learners. The same is true if the model is developed using data from rural learners and tested on data from urban learners. However, Samei and colleagues (2015) find that models on classroom discourse do not differ between urban and rural settings. More research is needed to determine which types of prediction are impacted when going between urban and rural settings.

Finally, one study conducted in the USA finds that models predicting graduation and standardized examination scores are less accurate for students with family members in the military (Baker et al., 2020).

Across these studies, a variety of variables are investigated and there is on the whole evidence that algorithmic bias has impacts that go beyond race/ethnicity, gender, and nationality. A range of variables have not yet been investigated at all, including religion, age, children of migrant workers, specific disabilities beyond speech impairment, transgender status, and sexual orientation.

5. From unknown bias to known bias, from fairness to equity

The previous sections of this chapter outline what is currently known to the field about how algorithmic bias is manifesting in education. Our review indicates that there is clear evidence that algorithmic bias is manifesting in many ways, but also indicates how limited our current knowledge is. Many potential areas of algorithmic bias are documented in just a single article, and many potential areas where algorithmic bias could be occurring have not yet been studied at all. There also is no clear sense as to the magnitude of the problem for different use cases and groups of students.

As Baker and Hawn (2022) note, we are at the very beginning of a progression in fixing the problem of algorithmic bias. At first, there is *unknown bias* -- a problem exists, but developers and researchers do not know that the problem exists. Perhaps it is known that there is a

problem in general, but not exactly who is being affected or exactly how. Descriptive research can move a specific educational algorithm from unknown bias to *known bias.*

In known bias, it is now known that there is a problem, where it is occurring, and who is impacted. Our knowledge may still be incomplete, but it is sufficient to potentially take action. Once we know what the bias is, it becomes possible to move towards *fairness.* There is increasing understanding of the steps that can be taken to increase the fairness of algorithms, within the broader machine learning community (Mehrabi et al., 2021; Narayanan, 2018). Although this work remains far from perfect, and debate is ongoing about the best methods (Kleinberg et al., 2017; Berk et al., 2021), there is enough knowhow about addressing algorithmic bias that once we know a bias exists, we can take steps to fix it. Finally, increasing algorithmic fairness can be a step towards creating a world with *equity*, with equal opportunity for all learners (see Holstein & Doroudi, 2022).

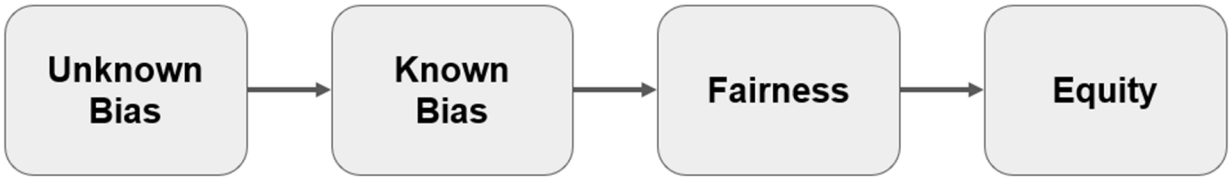

Figure X. The progression from the current situation to equity

Working towards equity necessarily implies determining where current technologies and pedagogies are today most unfair, and working to fix these problems first. Many of these places of greatest unfairness involve inequities that are already widely known. But some may also be less well-known to the educational and policy communities. We may miss key inequities due to our own biases and assumptions. In other words, more research is needed, because the world of education today is mostly in a state of *unknown bias.*

5a. Obstacles to fairness

There are currently many obstacles to achieving fairness and equity with educational technology. The biggest, as the previous section notes, is how much we do not know about the biases that exist in the world, in general but also between countries. As Baker and Hawn (2022) note, unknown biases can be split into two categories. The first is when we do not know that algorithmic bias exists for a specific group of learners. The second is when we know that there is bias impacting a specific group, but we do not know how this bias manifests. Both types of bias appear to exist in our current understanding of algorithmic bias in education. The research thus far is limited, both in terms of what groups have been studied, and how thoroughly we have studied algorithmic bias in education for the groups it is known to impact. Even for relatively thoroughly studied problems such as racism and sexism, we do not know all the ways that racism and sexism impact the effectiveness of educational algorithms. The biases of educational algorithms for indigenous populations has been much less studied than the biases

for other groups, for instance; transgendered learners have been much less studied; and the experience of racial minorities with educational algorithms has been much more thoroughly studied in the United States than in other countries.

One of the key barriers to conducting this type of research is the lack of high-quality and easily-accessible data on group membership, identity, perception, or status. As Belitz et al. (2022) notes, even when identity data is collected, it is in terms of a small number of categories. And most studies do not obtain even this limited level of identity or group membership data.

Barriers to collecting data on group membership come for many reasons, including convenience, regulatory barriers, and concerns around student privacy. Often, compliance organizations such as privacy officers and institutional review boards consider demographic data to be high-risk and create incentives (not always consciously) to avoid collecting this type of data. If -- to give an example commonly seen in the United States -- a researcher is required to collect parental consent if they collect demographic data, but is not required to collect any consent at all if they avoid demographic data, then there is a strong incentive to avoid collecting demographic data, and in turn to ignore issues of algorithmic bias (and other forms of bias as well). There are current efforts in many countries to create stricter data privacy laws for education -- laws that have the goal of protecting children, but as currently designed may make it impossible to identify or address algorithmic bias (see discussion in Baker, in press).

Another key incentive that reduces investigation into algorithmic bias is the risk involved to any commercial organization that is open about the flaws in their product. Any openness about a product's flaws -- or even openness about a product's design -- can be an opportunity for their competitors. An environment where commercial companies can choose whether or not to analyze their product's flaws, and where there is significant competition, is an environment where companies will have a strong reason not to look into (and fix) biases in their product. Being too open about bias may not just lead to sales competition -- it may lead to criticism by journalists, community members, and academics. At the extreme, an organization that is public about bias in their content risks lawsuits or action by regulators.

While there is currently some incentive to learning systems to demonstrate educational effectiveness (see platforms such as the What Works Clearinghouse and Evidence for ESSA – Clearinghouse, 2012; Slavin, 2020), currently these initiatives treat a curriculum as either effective or ineffective, rather than being effective or ineffective for specific groups of learners.

Another important obstacle to addressing algorithmic bias in education is the lack of toolkits for assessing and fixing algorithmic bias that are specifically tailored to education. Educational data has been known to be different than other types of data commonly used in machine learning, possessing a complex multi-level nature (actions within students within classrooms within teachers within schools within districts; and identity factors that are confounded with those levels) that must be accounted for in order for an analysis to be valid (O'Connell & McCoach, 2008). While existing toolkits are applicable up to a point, more work is needed to make it easy for them to adapt and use in education (see Kizilcec & Lee, 2022; Holstein & Doroudi, 2022).

Existing toolkits for identifying algorithmic bias offer generally useful metrics (discussed above), but often ignore the unique aspects of educational data, making them less relevant. The reason is that existing toolkits for addressing algorithmic bias are designed to treat data points as interchangeable, and therefore are not compatible with educational algorithms that explicitly consider the multi-level nature of educational data. The lack of toolkits currently increases the cost of testing for and resolving algorithmic bias for organizations without expertise in this area.

All in all, then, while the importance of addressing algorithmic bias in education is clear, without concerted efforts there are also manifestly several challenges and obstacles which will slow efforts in this area. Fortunately, there are several steps which policy-makers can take.

5b. Recommendations for policy-makers

In this section, we present six recommendations for policy-makers that can help to address algorithmic bias, resolving or working around the challenges currently present in the environment, and building on existing work by academics, NGOs, and industry (Box X.2).

Box X.2: Six policy pointers for policy makers

1. Consider algorithmic bias when considering privacy policy and mandates so that privacy requirements do not prevent researchers from identifying and addressing algorithmic bias
Require algorithmic bias analyses, including requiring necessary data collection
Guide algorithmic bias analysis based on local context and local equity concerns
Fund research into unknown biases around the world
Fund development of toolkits for algorithmic bias in education
Re-design effectiveness clearinghouses to consider learner diversity

1. Consider algorithmic bias when considering privacy policy and mandates
The first recommendation is simply to not make it impossible to address algorithmic bias. As mentioned above, many countries are currently considering legislation around data privacy in education that would make it impossible to collect (or retain for sufficient time to conduct analysis) the data on student identity, interaction, and outcomes which is necessary to identify and address algorithmic bias. If educational technology providers cannot collect or cannot use data on learner identity, they cannot determine who is negatively impacted by algorithmic bias, and almost certainly cannot produce algorithms less impacted by algorithmic bias. If educational technology providers cannot retain data on student usage long enough to measure relevant

outcomes, they cannot know if students in different groups are being differentially impacted. Student privacy is important but so is fairness.

2. Require algorithmic bias analyses, including requiring necessary data collection
Ideally, rather than create policy preventing the collection of data necessary to check for and address algorithmic bias, policy-makers would require the collection of data needed for these purposes, under best-practices safeguards. Ideally, this data collection mandate would be combined with some degree of protection or release from liability for companies that fully followed required security practices (especially in today's environment, where maintaining perfect data security is challenging even when following best practices).

This would be the first step towards requiring educational algorithms used beyond a certain scale (perhaps 1000 active users) to explicitly document and publish checks for algorithmic bias, at minimum providing evidence on whether the models have substantial difference in their quality of performance for different populations (if present in their user base). The requirement to publicly release evidence on algorithmic bias would probably be sufficient to create strong pressure to fix biases found in the algorithms.

3. Guide algorithmic bias analysis based on local context and local equity concerns
One current challenge faced by organizations making good-faith attempts to collect data to investigate algorithmic bias is deciding which identity variables to collect data on (Belitz et al., 2022). Policy-makers can assist with this. While census categories provide one source of possible variables, census categories simultaneously miss key categories shown to be associated with algorithmic bias (as discussed above) and also can include groups not present in a specific data set due to uneven distribution of that group across the general population. Policy standardizing a minimum set of identity markers to collect and report on in each policy region would provide consistency and comparability between different reports of algorithmic bias. It would also help to ensure that groups currently most disadvantaged in each region are supported rather than further disadvantaged by educational algorithms. Finally, standardizing on a minimum set of identity categories would also prevent organizations from reporting only the groups for whom their tool is unbiased. The actual process of selecting which categories are relevant within a specific policy environment should not be arbitrary; ideally, selection would be made by a representative combination of stakeholders in the local community, including researchers who can evaluate the data available.

4. Fund research into unknown biases around the world
As the discussion above illustrates, it is difficult to fix a problem if we do not know if it is there; it is difficult to fix *unknown biases.* Thus far, the super-majority of research on algorithmic bias has involved race/ethnicity and gender in the United States -- and even in the United States, key racial or ethnic groups more represented in specific geographical areas (such as Native Americans, and members of the Portuguese and Brazilian diasporas in New England) have been under-studied, as have other categories connected to algorithmic bias.

Outside the United States, there has been much less research into algorithmic bias. There is a clear need for further research on algorithmic bias in education in other OECD countries, investigating which groups are impacted and how they are impacted. Without this research, developers around the world will be limited to addressing the inequity problems known to exist in the United States, which are different from the problems in other countries (Wimmer, 2017), or will be guided by intuition rather than data in which problems they attempt to address.

Policy-makers can address this current limitation by creating grant-making programs which make funds available for research into who is impacted by algorithmic bias in education in their region.

5. Fund development of toolkits for algorithmic bias in education
As discussed above, the current lack of good toolkits for identifying and addressing algorithmic bias in education raises the cost of doing so; an organization must either hire an expert in this area or develop their own expertise over time. The development of high-quality, usability-engineered toolkits supporting the use of best practices will increase the feasibility of conducting this type of analysis and improvement, for a wide range of educational technology providers and researchers. Policy-makers can address this limitation by creating grant-making programs which make funds available for the development of toolkits of this nature. Even one such toolkit would make a substantial difference to the field.

6. Re-design effectiveness clearinghouses to consider learner diversity
Currently, effectiveness clearinghouses such as the What Works Clearinghouse and Evidence for ESSA -- created (respectively) directly by a governmental agency and with foundation grant funding -- summarize the evidence for the effectiveness of different curricula, including computer-delivered curricula. However, they treat effectiveness as a single dimension -- either a curriculum is effective for all or for none. Curricula and educational technologies may, however, be effective for specific groups of learners and not for others (Cheung & Slavin, 2013). An educational technology that is algorithmically biased is unlikely to be equally effective for all learners; if its algorithms function less effectively for specific groups of learners, the technology is very likely to function less effectively at supporting those learners in achieving better outcomes. As new clearinghouses are developed, or existing clearinghouses seek future funding, it may be possible for policy-makers to influence their directions towards considering differences between groups of learners in effectiveness. Doing so will provide greater incentive for educational technology providers (and curriculum developers in general) to document (and attend to) the effectiveness of their products for the full diversity of learners.

6. Conclusion

In this chapter, we have reviewed the current evidence for algorithmic bias in education: who is impacted, how they are impacted, and the (large) gaps in the field's understanding of this area. We review some of the factors slowing progress in this area, and conclude with recommendations for what policy-makers can do to support the field in understanding and reducing algorithmic biases in education.

The potential of algorithms for education is high. The best adaptive learning systems and at-risk prediction systems have made large positive impacts on student outcomes (Ma et al., 2014; Mojarad et al., 2018; VanLehn, 2011; Milliron et al., 2014). However, this potential cannot be fully reached if algorithms replicate or even magnify the biases occurring in societies around the world. It is only by researching and resolving algorithmic bias that we can develop educational technologies that reach their full potential, and in turn support every student in achieving their own full potential.

Policy makers around the world are at a key moment in the progress towards resolving algorithmic bias and developing educational technologies that are fair and equitable for all learners. There is increased understanding that algorithmic bias exists, including in education. There are the beginnings of progress in understanding who is impacted and how. However, this progress is limited in scope -- specific dimensions of student identity (particularly race/ethnicity and gender) have been much more heavily studied than other dimensions which also appear to be affected by algorithmic bias. Furthermore, research on algorithmic bias in education has been heavily concentrated in the United States, creating a lack of clarity on who is being negatively impacted in the rest of the world, and how to support them. Finally, this progress is put at risk by the possibility of imbalanced privacy laws, which may prevent future work to investigate and fix algorithmic biases and, ultimately, enhance equity.

References

Anderson, H., Boodhwani, A., & Baker, R. S. (2019). Assessing the fairness of graduation predictions. *Proceedings of the 12th International Conference on Educational Data Mining*, 488–491.

Angwin, J., Larson, J., Mattu, S., & Kirchner, L. (2016). Machine bias. In *Ethics of Data and Analytics* (pp. 254-264). Auerbach Publications.

Baker, R.S. (in press) The Current Trade-off Between Privacy and Equity in Educational Technology. To appear in G. Brown III, C. Makridis (Eds.) *The Economics of Equity in K-12 Education: Necessary Programming, Policy, and Systemic Changes to Improve the Economic Life Chances of American Students.* Lanham, MD: Rowman & Littlefield.

Baker, R. S., Berning, A., & Gowda, S. M. (2020). Differentiating military-connected and non-military-connected students: Predictors of graduation and SAT score. EdArXiv.

Baker, R. S., & Hawn, A. (2022). Algorithmic bias in education. *International Journal of Artificial Intelligence in Education*, *32*(4), 1052-1092.

Baker, R.S., Ogan, A.E., Madaio, M., Walker, E. (2019) Culture in Computer-Based Learning Systems: Challenges and Opportunities. *Computer-Based Learning in Context, 1*(1),

1-13.
Barocas, S., Hardt, M., & Narayanan, A. (2019). Fairness and machine learning. Online Book, fairmlbook.org. http://www.fairmlbook.org
Belitz, C., Ocumpaugh, J., Ritter, S., Baker, R. S., Fancsali, S. E., & Bosch, N. (2022). Constructing categories: Moving beyond protected classes in algorithmic fairness. *Journal of the Association for Information Science and Technology*.
Berk, R., Heidari, H., Jabbari, S., Kearns, M., & Roth, A. (2018). Fairness in criminal justice risk assessments: The state of the art. *Sociological Methods & Research*, 50(1), 3–44.
Berk, R., Heidari, H., Jabbari, S., Kearns, M., & Roth, A. (2021). Fairness in criminal justice risk assessments: The state of the art. *Sociological Methods & Research*, 50(1), 3–44
Bireda, M. R. (2002). Eliminating racial profiling in school discipline: Cultures in conflict. Scarecrow Press.
Blodgett, S. L., Barocas, S., III, H. D., & Wallach, H. (2020). Language (Technology) is Power: A Critical Survey of "Bias" in NLP. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 5454–5476.
Bridgeman, B., Trapani, C., & Attali, Y. (2009). Considering fairness and validity in evaluating automated scoring [Paper presentation]. *Annual Meeting of the National Council on Measurement in Education (NCME)*, United States.
Bridgeman, B., Trapani, C., & Attali, Y. (2012). Comparison of human and machine scoring of essays: Differences by gender, ethnicity, and country. *Applied Measurement in Education*, 25(1), 27–40.
Caton, S., & Haas, C. (2020). Fairness in machine learning: A survey. arXiv preprint arXiv:2010.04053.
Cheung, A. C., & Slavin, R. E. (2013). The effectiveness of educational technology applications for enhancing mathematics achievement in K-12 classrooms: A meta-analysis. *Educational research review*, 9, 88-113.
Christie, S. T., Jarratt, D. C., Olson, L. A., & Taijala, T. T. (2019). Machine-Learned School Dropout Early Warning at Scale. P*roceedings of the International Conference on Educational Data Mining.*
Chouldechova, A. (2017). Fair prediction with disparate impact: A study of Bias in recidivism prediction instruments. *Big Data*, 5(2), 153–163.
Clearinghouse, W. W. (2012). What works clearinghouse. Internet site: http://ies. ed. gov/ncee/wwc.
Corbett-Davies, S., & Goel, S. (2018). The measure and mismeasure of fairness: A critical review of fair machine learning. *arXiv preprint arXiv:1808.00023*.
Cramer, H., Holstein, K., Vaughan, J. W., Daumé, H., Dudik, M., Wallach, H., Reddy, S., & Jean,
G.-G. [The Conference on Fairness, Accountability, and Transparency (FAT*)]. (2019). *FAT* 2019 translation tutorial: Challenges of incorporating algorithmic fairness* [video]. YouTube. https://youtu.be/UicKZv93SOY
Crawford, K. [The Artificial Intelligence Channel]. (2017). The Trouble with Bias - NIPS 2017 Keynote - Kate Crawford [Video]. YouTube. https://youtu.be/fMym_BKWQzk
Darlington, R. B. (1971). Another look at "cultural fairness.". *Journal of Educational Measurement*, 8(2), 71–82.

Digital Promise. (n.d.). 1. Prioritizing Racial Equity in AI Design. Digital Promise. Retrieved December 23, 2022, from https://productcertifications.microcredentials.digitalpromise.org/explore/1-prioritizing-racial-equity-in-ai-design-4

Dorans, N. J. (2010). Misrepresentations in unfair treatment by Santelices and Wilson. *Harvard Educational Review*, 80(3), 404–413.

Doroudi, S., & Brunskill, E. (2019). Fairer but not fair enough on the equitability of knowledge tracing. *Proceedings of the 9th International Conference on Learning Analytics & Knowledge*, 335339.

Executive Office of the President. (2014). Big Data: Seizing Opportunities, Preserving Values. https://obamawhitehouse.archives.gov/sites/default/files/docs/big_data_privacy_report_may_1_2014.pdf

Finkelstein, S., Yarzebinski, E., Vaughn, C., Ogan, A., & Cassell, J. (2013). The effects of culturally congruent educational technologies on student achievement. In H. C. Lane, K. Yacef, J. Mostow, & P. Pavlik (Eds.), *Proceedings of the 16th International Conference on Artificial Intelligence in Education* (pp. 493–502). Springer Berlin Heidelberg.

Friedman, B., & Nissenbaum, H. (1996). Bias in computer systems. *ACM Transactions on Information Systems*, 14(3), 330–347.

Garcia, M. (2016). Racist in the Machine: The Disturbing Implications of Algorithmic Bias. *World Policy Journal*, 33(4), 111–117.

Gardner, J., Brooks, C., & Baker, R. (2019, March). Evaluating the fairness of predictive student models through slicing analysis. In *Proceedings of the 9th international conference on learning analytics & knowledge* (pp. 225-234).

Green, B. (2020). The false promise of risk assessments: Epistemic reform and the limits of fairness. *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, 594–606.

Green, B., & Hu, L. (2018, July 10–15). The Myth in the Methodology: Towards a Recontextualization of Fairness in Machine Learning [Conference presentation]. The Debates Workshop at the 35th International Conference on Machine Learning, Stockholm, Sweden.

Green, B., & Viljoen, S. (2020). Algorithmic realism: Expanding the boundaries of algorithmic thought. *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, 19–31.

Hellström, T., Dignum, V., & Bensch, S. (2020). Bias in Machine Learning -- What is it Good for? In A. Saffiotti, L. Serafini, & P. Lukowicz (Eds.), *Proceedings of the First International Workshop on New Foundations for Human-Centered AI (NeHuAI) co-located with 24th European Conference on Artificial Intelligence* (ECAI 2020) (pp. 3–10). RWTH Aachen University.

Holstein, K., & Doroudi, S. (2022). Equity and Artificial Intelligence in education. In *The Ethics of Artificial Intelligence in Education* (pp. 151-173). Routledge.

Hu, Q., & Rangwala, H. (2020). Towards Fair Educational Data Mining: A Case Study on Detecting At-risk Students. *Proceedings of The 13th International Conference on Educational Data Mining* (EDM 2020), 431–437.

Kai, S., Andres, J. M. L. ., Paquette, L., Baker, R. S. ., Molnar, K., Watkins, H., & Moore, M.

(2017). Predicting Student Retention from Behavior in an Online Orientation Course. *Proceedings of the 10th International Conference on Educational Data Mining*, 250–255.

Kizilcec, R. F., & Lee, H. (2022). Algorithmic fairness in education. In *The Ethics of Artificial Intelligence in Education* (pp. 174-202). Routledge.

Klare, B. F., Burge, M. J., Klontz, J. C., Bruegge, R. W. V., & Jain, A. K. (2012). Face recognition performance: Role of demographic information. *IEEE Transactions on Information Forensics and Security, 7(6), 1789–1801*.

Kleinberg, J., Mullainathan, S., & Raghavan, M. (2017). Inherent trade-offs in the fair determination of risk scores. In C. H. Papadimitriou (Ed.), *Proceedings of the 8th Innovations in Theoretical Computer Science Conference* (ITCS 2017) (Vol. 67, pp. 43:1–43:23)

Knight, W. (2019). The Apple Card Didn't "See" Gender—and That's the Problem. *Wired*.

Kraiger, K., & Ford, J. K. (1985). A meta-analysis of ratee race effects in performance ratings. *Journal of Applied Psychology*, 70(1), 56–65.

Kung, C., & Yu, R. (2020, August). Interpretable models do not compromise accuracy or fairness in predicting college success. In *Proceedings of the seventh acm conference on learning@ scale* (pp. 413-416).

Lee, H., & Kizilcec, R. F. (2020). Evaluation of fairness trade-offs in predicting student success. arXiv preprint arXiv:2007.00088.

Li, X., Song, D., Han, M., Zhang, Y., & Kizilcec, R. F. (2021). On the limits of algorithmic prediction across the globe. arXiv preprint arXiv:2103.15212.

Loukina, A., & Buzick, H. (2017). Use of automated scoring in spoken language assessments for test takers with speech impairments. *ETS Research Report Series*, 2017(1), 1–10. H

Loukina, A., Madnani, N., & Zechner, K. (2019). The many dimensions of algorithmic fairness in educational applications. *Proceedings of the Fourteenth Workshop on Innovative Use of NLP for Building Educational Applications*, 1–10.

Obermeyer, Z., Powers, B., Vogeli, C., & Mullainathan, S. (2019). Dissecting racial bias in an algorithm used to manage the health of populations. *Science*, 366(6464), 447–453.

Ma, W., Adesope, O. O., Nesbit, J. C., & Liu, Q. (2014). Intelligent tutoring systems and learning outcomes: A meta-analysis. *Journal of educational psychology*, 106(4), 901.

Makhlouf, K., Zhioua, S., & Palamidessi, C. (2020). On the applicability of ML fairness notions. ArXiv E-Prints, arXiv:2006.16745. https://arxiv.org/abs/2006.16745.

Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., & Galstyan, A. (2021). A survey on bias and fairness in machine learning. *ACM Computing Surveys* (CSUR), 54(6), 1-35.

Mojarad, S., Essa, A., Mojarad, S., & Baker, R. S. (2018, March). Studying adaptive learning efficacy using propensity score matching. In *Companion Proceedings of the 8th International Conference on Learning Analytics and Knowledge (LAK'18)* (pp. 5-9).

Milliron, M. D., Malcolm, L., & Kil, D. (2014). Insight and Action Analytics: Three Case Studies to Consider. *Research & Practice in Assessment*, 9, 70-89.

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.

Naismith, B., Han, N.-R., Juffs, A., Hill, B., & Zheng, D. (2018). Accurate Measurement of

Lexical Sophistication with Reference to ESL Learner Data. *Proceedings of 11th International Conference on Educational Data Mining*, 259–265.

Narayanan, A. (2018, February). Translation tutorial: 21 fairness definitions and their politics. In *Proc. Conf. Fairness Accountability Transp.*, New York, USA (Vol. 1170, p. 3).

O'Connell, A. A., & McCoach, D. B. (Eds.). (2008). Multilevel modeling of educational data. IAP.

Ocumpaugh, J., Baker, R., Gowda, S., Heffernan, N., & Heffernan, C. (2014). Population validity for educational data mining models: A case study in affect detection. *British Journal of Educational Technology*, 45(3), 487–501.

Ogan, A., Walker, E., Baker, R., Rodrigo, M. M. T., Soriano, J. C., & Castro, M. J. (2015). Towards understanding how to assess help-seeking behavior across cultures. *International Journal of Artificial Intelligence in Education*, 25(2), 229–248

Okur, E., Aslan, S., Alyuz, N., Arslan Esme, A., & Baker, R. S. (2018). Role of socio-cultural differences in labeling students' affective states. In C. Penstein Rosé, R. Martínez-Maldonado, H. U. Hoppe, R. Luckin, M. Mavrikis, K. Porayska-Pomsta, B. McLaren, & B. du Boulay (Eds.), *Proceedings of the 19th International Conference on Artificial Intelligence in Education* (pp. 367–380). Springer International Publishing.

O'Neil, C. (2017). Weapons of math destruction: How big data increases inequality and threatens democracy. Crown.

Paquette, L., Ocumpaugh, J., Li, Z., Andres, A., & Baker, R. (2020). Who's learning? Using demographics in EDM research. *Journal of Educational Data Mining*, 12(3), 1–30.

Pardo, A., & Siemens, G. (2014). Ethical and privacy principles for learning analytics. *British journal of educational technology*, 45(3), 438-450.

Penn Center for Learning Analytics Wiki. (n.d.). Retrieved from https://www.pcla.wiki/index.php/Algorithmic_Bias_in_Education

Rzepka, N., Simbeck, K., Müller, H. G., & Pinkwart, N. (2022). Fairness of In-session Dropout Prediction. In *CSEDU (2)* (pp. 316-326).

Ramineni, C., & Williamson, D. (2018). Understanding mean score differences between the e-rater® automated scoring engine and humans for demographically based groups in the GRE® general test. *ETS Research Report Series*, 2018(1), 1–31

Riazy, S., Simbeck, K., & Schreck, V. (2020). Fairness in Learning Analytics: Student At-risk Prediction in Virtual Learning Environments. *Proceedings of the 12th International Conference on Computer Supported Education* (CSEDU 2020), 1, 15–25.

Samei, B., Olney, A. M., Kelly, S., Nystrand, M., D'Mello, S., Blanchard, N., & Graesser, A. (2015). Modeling Classroom Discourse: Do Models That Predict Dialogic Instruction Properties Generalize across Populations?. *International Educational Data Mining Society*.

Santelices, M. V., & Wilson, M. (2010). Unfair treatment? The case of Freedle, the SAT, and the standardization approach to differential item functioning. *Harvard Educational Review*, 80(1), 106–134.

Sha, L., Raković, M., Das, A., Gašević, D., & Chen, G. (2022). Leveraging class balancing techniques to alleviate algorithmic bias for predictive tasks in education. *IEEE Transactions on Learning Technologies*, *15*(4), 481-492.

Sha, L., Rakovic, M., Whitelock-Wainwright, A., Carroll, D., Yew, V. M., Gasevic, D., & Chen, G.

(2021, June). Assessing algorithmic fairness in automatic classifiers of educational forum posts. In *International conference on artificial intelligence in education* (pp. 381-394). Springer, Cham.

Silva, S., & Kenney, M. (2018). Algorithms, platforms, and ethnic Bias: An integrative essay. *Phylon* (1960-), 55(1&2), 9–37.

Slavin, R. E. (2020). How evidence-based reform will transform research and practice in education. *Educational Psychologist*, 55(1), 21-31.

Soundarajan, S., & Clausen, D. L. (2018). Equal Protection Under the Algorithm: A Legal-Inspired Framework for Identifying Discrimination in Machine Learning. *Proceedings of the 35th International Conference on Machine Learning.*

Suresh, H., & Guttag, J. V. (2020). A framework for understanding unintended consequences of machine learning. ArXiv E-Prints, arXiv:1901.10002.

Sweeney, L. (2013). Discrimination in online ad delivery. *Communications of the ACM*, 56(5), 44–54.

Telford, T. (2019). Apple Card algorithm sparks gender bias allegations against Goldman Sachs.
Washington Post. https://www.washingtonpost.com/business/2019/11/11/apple-card-algorithmsparks-gender-bias-allegations-against-goldman-sachs/.

Tempelaar, D., Rienties, B., & Nguyen, Q. (2020). Subjective data, objective data and the role of bias in predictive modelling: Lessons from a dispositional learning analytics application. *PLoS One*, 15(6), e0233977.

VanLehn, K. (2011). The relative effectiveness of human tutoring, intelligent tutoring systems, and other tutoring systems. *Educational psychologist*, 46(4), 197-221.

Vasquez Verdugo, J., Gitiaux, X., Ortega, C., & Rangwala, H. (2022, March). FairEd: A Systematic Fairness Analysis Approach Applied in a Higher Educational Context. In *LAK22: 12th International Learning Analytics and Knowledge Conference* (pp. 271-281).

Verma, S., & Rubin, J. (2018, May). Fairness definitions explained. In 2018 ieee/acm *international workshop on software fairness* (fairware) (pp. 1-7). IEEE.

Wang, Z., Zechner, K., & Sun, Y. (2018). Monitoring the performance of human and automated scores for spoken responses. *Language Testing*, *35*(1), 101-120.

Wimmer, A. (2017). Power and pride: national identity and ethnopolitical inequality around the world. *World Politics*, 69(4), 605-639.

Wolff, A., Zdrahal, Z., Nikolov, A., & Pantucek, M. (2013, April). Improving retention: predicting at-risk students by analysing clicking behaviour in a virtual learning environment. In *Proceedings of the third international conference on learning analytics and knowledge* (pp. 145-149).

Yu, R., Lee, H., & Kizilcec, R. F. (2021, June). Should college dropout prediction models include protected attributes?. In *Proceedings of the eighth ACM conference on learning@ scale* (pp. 91-100).

Yu, R., Li, Q., Fischer, C., Doroudi, S., & Xu, D. (2020). Towards Accurate and Fair Prediction of College Success: Evaluating Diferent Sources of Student Data. *Proceedings of The 13th International Conference on Educational Data Mining* (EDM 2020), 292–301.

Zhang, J., Andres, J.M.A.L., Hutt, S., Baker, R.S., Ocumpaugh, J., Mills, C., Brooks, J.,

Sethuraman, S., Young, T. (2022) Detecting SMART Model Cognitive Operations in Mathematical Problem-Solving Process. *Proceedings of the International Conference on Educational Data Mining*.

Zuiderveen Borgesius, F. (2018). Discrimination, artificial intelligence, and algorithmic decision-making. Council of Europe, Directorate General of Democracy.