Using Demographic Data as Predictor Variables: a Questionable Choice

Ryan S. Baker
University of Pennsylvania
ryanshaunbaker@gmail.com

Lief Esbenshade
Google
liefesbenshade@google.com

Jonathan Vitale
Google
jonvitale@google.com

Shamya Karumbaiah
University of Wisconsin
shamya16@gmail.com

Predictive analytics methods in education are seeing widespread use and are producing increasingly accurate predictions of students' outcomes. With the increased use of predictive analytics comes increasing concern about fairness for specific subgroups of the population. One approach that has been proposed to increase fairness is using demographic variables directly in models, as predictors. In this paper we explore issues of fairness in the use of demographic variables as predictors of long-term student outcomes, studying the arguments for and against this practice in the contexts where this literature has been published. We analyze arguments for the inclusion of demographic variables, specifically claims that this approach improves model performance and charges that excluding such variables amounts to a form of 'color-blind' racism. We also consider arguments against including demographic variables as predictors, including reduced actionability of predictions, risk of reinforcing bias, and limits of categorization. We then discuss how contextual factors of predictive models should influence case-specific decisions for the inclusion or exclusion of demographic variables and discuss the role of proxy variables. We conclude that, on balance, there are greater benefits to fairness if demographic variables are used to validate fairness rather than as predictors within models.

Keywords: predictive analytics, at-risk prediction, demographic variables, algorithmic bias, algorithmic fairness

1. Introduction

The methods of predictive analytics have emerged as a useful tool for predicting an individual's future decisions and outcomes, in a range of domains (Siegel, 2013). Predictive analytics has emerged in education as well, producing increasingly accurate predictions of future student performance during learning (Gasevic et al., 2016; Gervet et al., 2020) and enabling prediction of a range of longitudinal outcomes (Bowers et al., 2012; Kloft et al., 2014; Milliron et al., 2014; Almeda & Baker, 2020). With the increased use of predictive analytics in education has come increasing concern about whether they are unfair for specific subgroups of the population (Baker & Hawn, 2021; Kizilcec & Lee, 2022). Prediction models can be used for a range of applications. This paper explores issues of fairness in educational success prediction models' construction and addresses the question of whether demographic variables should – or should not – be included.

Models to predict whether a learner will drop out of high school or college, sometimes referred to as early warning systems, are perhaps the most widely-scaled use of prediction models in education (Bowers et al., 2012; Milliron et al., 2014). Early work in this area, such as the "Chicago Model" (developed using data from Chicago and first applied there) used a small set of variables from a student's first year of high school (ABC: attendance, behavior, and course performance), combined by a human analyst to predict whether a student was on track to graduation (Allensworth & Easton, 2007). This model was deployed in many districts across the United States, as part of the Diplomas Now comprehensive intervention (Corrin et al., 2016). Later work employed more complex machine learning algorithms and a broader range of variables describing student experience and performance, achieving significantly better predictive accuracy (Dekker et al., 2009; Christie et al., 2019). Work then explored questions such as which variables are most relevant for predicting dropout (Nagrecha et al., 2017; Gardner & Brooks, 2018), which algorithms are most effective (Dalipi et al., 2018), how to generate effective prediction for new districts (Coleman et al., 2019), and which metrics were most important for evaluating real-world use of models (Gardner & Brooks, 2018). The resultant models have scaled up and are now used to make predictions about millions of learners (Christie et al., 2019; Coleman et al., 2019). Additionally, school and district leaders now routinely use predictive analytics to guide decision making around allocation of support resources (Bowers, 2021) and for predicting staff turnover (Boyce & Bowers, 2016).

The same type of algorithm is now used in an expanding range of use cases in education. Predictive models have been used in adaptive learning system content selection to infer whether a student is likely to get a specific item right (Fancsali et al., 2013). Predictive models have been used to support research on the long-term impacts of student behavior, i.e., predicting whether learners will achieve educational objectives over a decade later (Almeda & Baker, 2020). The same types of algorithms have also been used to infer a range of constructs, from student affect (Hutt et al., 2019) to disengaged behaviors (Baker & Rossi, 2013) to conceptual understanding (Asbell-Clarke et al., 2015).

Beyond accuracy and applicability, a central concern in the development and use of predictive models is fairness - i.e., that models are not biased against specific groups of students, particularly those that have been historically marginalized. While there is general consensus on the importance of fairness for these models, a consensus on how best to achieve it has yet to emerge. In particular, there is currently considerable controversy around whether demographic variables should be included in models as predictors. This paper investigates this question, examining the role of student demographic variables in predictive models and the costs and benefits of using these variables in this fashion.

1.1. Increasing concern about algorithmic bias in education

The increased use of at-risk prediction models, and prediction models in general, has been accompanied by concerns that the algorithms being used are biased against individuals who are already disadvantaged or discriminated against societally. Increasing evidence is emerging that the choices made in developing algorithms often lead to algorithms replicating (or even amplifying) existing societal biases, termed *algorithmic bias*. Algorithmic bias has been documented in contexts as varied as criminal justice (Angwin et al., 2016), hiring decisions (Garcia, 2016), medicine (O'Reilly-Shah et al., 2020), and computer vision (Klare et al., 2012). Algorithmic bias has been documented in high-stakes decision-making such as predictions of recidivism (Angwin et al., 2016) or decisions of which patients to administer anesthesia to (O'Reilly-Shah et al., 2020).

This concern is increasingly found in education as well (and, indeed, concern about fairness and bias in testing has been a focus of educational measurement for decades – see review in Hutchinson & Mitchell, 2019). Although Paquette and colleagues (2020) note that most papers published in the *Journal of Educational Data Mining* and the *International Conference on Educational Data Mining* do not yet consider algorithmic bias, there has nonetheless been a recent expansion of work on algorithmic bias in education. Baker and Hawn (2021) review 23 papers documenting various forms of algorithmic bias in education, and note that algorithmic bias has been documented for a range of demographic variables, including race, ethnicity, gender, native language, disability, urbanicity, parental education, socioeconomic status, nationality and national origin, and military-connected status. They also note that algorithmic bias has been demonstrated for a range of uses of algorithms in education, including models of student affect, inferences of student knowledge, automated essay scoring, and at-risk prediction. In specific, several papers have documented that algorithms are less accurate at predicting whether a student will drop out of high school or college, if the student is a member of a historically underrepresented group (Kai et al., 2017; Anderson et al., 2019; Hu & Rangwala, 2020).

This concern has led to interest in creating fairer models, both in education and in the use of algorithms more broadly. Corbett-Davies & Goel (2018) provide a review of the history of algorithmic bias and definitions around algorithmic bias; a more extensive review of definitions can be found in Caton & Haas (2020), and the applications of these definitions in education is discussed by Kizilcec & Lee (in press). Suresh & Guttag (2019) provide a framework for understanding the sources of bias in machine learning; their sources of bias are applied to education in (Baker & Hawn, 2021) and alternate frameworks are offered for education by Holstein & Doroudi (2022) and Karumbaiah (2021). All of these reviews note the exploding interest in documenting algorithmic bias in recent years. There has been a corresponding expansion in the methods used to attempt to address and/or reduce algorithmic bias – see review in general in Caton & Haas (2020) and for education in Kizilcec & Lee (in press).

However, there is ongoing disagreement about not only which methods are best for reducing algorithmic bias, but also which definition of algorithmic bias should be optimized for. Different definitions of algorithmic bias have different practical and social implications, and several articles have demonstrated that it is infeasible to optimize for all definitions and metrics at once (Chouldechova, 2017; Kleinberg et al., 2017; Berk et al., 2018; Loukina et al., 2019; Lee & Kizilcec, 2020, Darlington, 1971). The practical differences between different definitions can be large – for example, attempts to achieve statistical/demographic parity (equal assignment to outcomes by demographic group) can produce very different results than attempts to achieve predictive equality (equal accuracy of predictions for demographic groups) (Corbett-Davies et al., 2017). These debates echo broader debates around the best ways to achieve fairness, equity, and justice in the use of algorithms, in education and beyond – small differences in goals can lead to major disagreements in whether a specific use of algorithms is acceptable or not (Holstein & Doroudi, 2022), and some have argued that questions of algorithmic fairness and justice are fundamentally political and ethical in nature rather than technical (Wong, 2020).

Specific methods of addressing algorithmic bias are particularly appropriate for specific definitions of algorithmic bias (Caton & Haas, 2020). But deciding which method of addressing algorithmic bias to choose depends on more than just how one defines algorithmic bias – it also depends on the application which the model will be used for.

1.2. Emerging debate: how should demographic data be used in modeling?

Although there is general consensus that predictive analytics algorithms should be fair for members of all groups, and in particular should not discriminate against individuals in historically disadvantaged and underrepresented groups, there is considerable disagreement around how indicators of group demographics should be used in the modeling process.

A popular viewpoint argues that demographic indicators should be used as predictors in models. According to some researchers and practitioners, using demographic data as predictors leads to more accurate prediction of whether a student is at risk (Wolff et al., 2013; Kleinberg et al., 2017), and therefore demographic data should be used as predictors because having the most accurate prediction of at-risk status is of key importance. However, in light of evidence that using demographic data as predictors sometimes leads to over-fitting and lower predictive accuracy on new data (Yu et al., 2020, 2021), some researchers have argued that using demographic data as predictors is appropriate even if it does not lead to substantially better performance:

> "In short, our results suggest limited effects of including protected attributes on the performance of college dropout prediction… This compensating effect, however, does not accumulate to statistically significant changes in predicted labels, possibly because the group differences in dropout rates were not sizeable in the past at the institution we study… Still, the existence of a weak compensating instead of segregating effect justifies the inclusion of these attributes. After all, a major argument for race-aware models, and more generally socio-demographic-aware models, is to capture structural inequalities in society that disproportionately expose members of minoritized groups to more adverse conditions. In addition, the deliberate exclusion of protected attributes from dropout prediction models can be construed as subscribing to a 'colorblind' ideology, which has been criticized as a racist approach that serves to maintain the status quo." (Yu et al., 2021, p. 98).

Currently, the viewpoint that demographic data should be used as predictors in models is probably the prevailing view in academic papers on at-risk prediction modeling, though generally more due to the belief that it leads to better prediction than other factors. Perhaps the first paper to clearly articulate this view in education is Wolff et al. (2013), but in succeeding years this view has become quite prominent. For instance, Paquette et al.'s (2020) review of articles published in educational data mining venues finds that approximately half of articles that use demographics in analysis do so by using demographic variables as predictors. It is also seen in the commercial at-risk prediction systems used at scale in K-12 such as the at-risk prediction incorporated into the PowerSchool student information system (Feathers, 2022).

A contrasting viewpoint, sometimes referred to in the AI community as “fairness through unawareness” (Kusner et al., 2017), is that demographic data should not be used as a predictor in models used in real-world decision-making, but should instead be used to validate those models. Researchers holding this view argue that using demographic data as predictors may instead reinforce bias (Paquette et al., 2020) and produce predictions based primarily on demographic variables rather than more actionable variables (Feathers, 2022). Instead, researchers in this paradigm typically recommend analyzing a model’s goodness of fit broken down by demographic group, to ensure that model quality is comparable across groups. Although this view is somewhat less prevalent in academic writing on at-risk prediction models, it is still seen in almost as many educational data mining papers as the practice of using demographic variables as predictors (Paquette et al., 2020). In addition, commercial at-risk prediction systems used at scale in K-12 such as BrightBytes (Coleman et al., 2019) and Infinite Campus (Christie et al., 2019) have adopted this view.

A third possibility, somewhat in between these possibilities, is the use of demographic data as constraints in models, to ensure that a model is fair to all groups. An example of this is cost-sensitive classification, where a model may be explicitly instructed to up-weight the cost of errors for some groups of students in order to have an even proportion of errors for all groups of students, or to meet another fairness constraint (Agarwal et al., 2018). Though uncommon in educational use of at-risk prediction, examples do exist (Vasquez Verdugo et al., 2022).

In this article, we present an argument in favor of using demographic variables for validation rather than as predictors, within the context of at-risk prediction models in education. We will consider some of the key arguments for both perspectives, discuss the relevant empirical evidence, and the implications of each of these positions to real-world practice. We will also consider the role of proxy variables and mediating variables in this debate. We will discuss alternate ways to use demographic variables, such as cost-sensitive classification, in section 5.2. First, however, we explore in a bit more detail one of the key underpinnings of any position on the use of demographic variables – exactly what a demographic variable is and is not.

2. DEFINING WHAT A DEMOGRAPHIC VARIABLE IS

The term “demographic variable” generally in education refers to information that provides context about students’ backgrounds and experiences that they bring with them to school. Commonly used demographic variables in student risk prediction include race/ethnicity, gender/sex, disabilities, whether a student’s family home language is the local language of schooling (i.e., English as a Second Language/ESL status in the United States and United Kingdom), Special Education status (a classification for students diagnosed with certain disabilities, Belcher & Hatley, 1994), free or reduced price meal eligibility (an imperfect proxy for poverty, Domina et al 2018). Less commonly, variables such as homelessness status (Hyman et al., 2011), mobility (Goldhaber et., 2022), and military-connected status (Baker et al., 2020) are also used.

Definitionally, demographic variables cannot be manipulated (Lee & Scheule, 2010), either due to ethical reasons or feasibility. Take homelessness, for example. A school can intervene to improve a student's attendance rate, but they generally cannot feasibly intervene to provide housing for a student experiencing homelessness. Non manipulability does not mean that schools do not engage in successful strategies to support students of specific demographic groups. Indeed, schools have provided services to support homeless students since the late 1980's (Yon, 1995), recent research has found that targeted supports can increase graduation rates for Black male students (Dee & Penner, 2021), and there is increasing awareness that student-teacher demographic match is an important factor in school success for underrepresented groups (Gottfried et al., 2022).

Some variables are often considered demographic but are potentially mutable by the school. Consider the distinction between English language learner (ELL) status and whether a student's family home language is the local language of schooling (English as a Second Language; ESL status). Both of these variables have been found to be significant predictors of graduation rates (Lee, 2012). Within modern ethical standards, a student's home language, and thereby their ESL status, cannot be changed by schools. However, schools explicitly work to change ELL designation through direct instruction in the English language. Therefore, using a continuous measure of English fluency rather than a binary designation would be more reactive to interventions in adjusting risk predictions, but even the binary status indicator can change, allowing the student's risk prediction to change.

Cohort-normed age - i.e., is a student older or younger than the average student in their class - is another relevant case to consider. It is sometimes used in risk predictions as retaining students in a grade has been found to be associated with increased risk of not-graduating (Bornsheuer et al 2011). Here the relevant variance in the demographic variable is a direct result of school policies to retain students in a grade. Though it may be difficult for a school to accelerate a student back onto grade level, this is a demographic variable over which the school has direct control. It should be noted, however, that in communities where red-shirting – deliberately delaying entry into Kindergarten so that the student is more developed than their peers (Bassok & Reardon 2013) – is common, this variable's meaning may be complicated.

In general, cases where a demographic variable may be mutable present complex questions, which need to be thought through on a variable-by-variable basis – there is a range of incentives for organizations (and the individual themself) to change an individual's label, some salutary and others less so. A change may be legitimate (i.e. a learner achieves language fluency), but the learner may still share other attributes with the other members of the original group.

3. Arguments for using demographic data as predictor variables: discussion and critiques

A number of researchers and practitioners have argued in favor of using demographic data as predictor variables (Kleinberg et al., 2018; Yu et al., 2021). Likewise, in practice, at-risk prediction models using demographic data have been employed for both research and educational applications (Wolff et al., 2013; Feathers, 2022). In this section, we present and offer a critique of some of the key arguments in favor of this practice.

One of the core arguments in favor of using demographic data as predictor variables is that it can lead to better model performance. If structural racism (or other forms of bias) produces different outcomes for otherwise identical students, in a way that cannot be captured by other predictor variables, then we should expect better predictive performance when we include demographic variables as predictors. In an early example of this, Wolff et al. (2013) find that course outcome prediction models perform better when an unnamed student demographic variable is included. In some cases, this improvement has been specifically documented for individuals in disadvantaged groups. For example, Kleinberg et al. (2018) investigates this question in the context of models predicting college GPA, comparing models that take race into account to models that do not take race into account. They find that models that take race into account perform substantially better.

However, other papers do not find this same improvement when demographic variables are used as predictors. For example, Yu and colleagues (2020) find that models predicting both course grade and GPA perform substantially worse when demographic information is included, and Yu and colleagues (2021) find no significant advantages by including a range of demographic variables in a model of college dropout, either in terms of overall model performance, performance for historically disadvantaged students, or fairness metrics. Deho and colleagues (2022) also find no performance benefit for including demographic variables in models. As such, it appears that it is unclear whether including demographic variables generally leads to better model performance, despite the strong intuitive appeal of such a notion. Some of the difference may also come down to how many non-demographic variables are available: Yu et al. (2020, 2021) have a much broader feature set than Kleinberg et al. (2018), for instance. Ultimately, the question of whether including demographic variables will usually lead to better performance or not, for specific problems in education, is an empirical question that will need to be settled by further research (and then meta-analyses, to summarize the evidence around benefit or lack of benefit for specific problems).

However, in response to the finding that adding demographic variables as predictors does not always lead to better model performance, Yu and colleagues (2021) argue for using demographic data as model predictors even when it has no benefit in terms of model performance or fairness metrics. They state that "the deliberate exclusion of protected attributes from dropout prediction models can be construed as subscribing to a 'colorblind' ideology, which has been criticized as a racist approach that serves to maintain the status quo." (Yu and colleagues, 2021, p. 98). The claim that exclusion of protected attributes is color-blind racism is also made in other papers such as (Andrus & Villaneueve, 2022; Schwobel & Remmers, 2022).

It is worth carefully considering this claim, and whether this analogy holds clearly.

According to Bonilla-Silva (2006), a key text on color-blind racism (and one of the two cited in Yu et al.'s discussion), color-blind racism is a means of implicitly justifying racial inequities without directly appealing to inherent racial characteristics. Color-blind racism, in this account, emerges from four conceptual frames:

- *abstract liberalism*: using ideas such as equal opportunity in the abstract rather than contextualizing them – for instance, opposing affirmative action on the basis of equal opportunity while ignoring the very unequal opportunity existing for members of different races (example taken from Bonilla-Silva, 2006, p. 28)
- *naturalization*: explaining away race-based choices as being natural to all groups – for instance, insisting that only having friends of one's own racial group is natural to all individuals (example taken from Bonilla-Silva, 2006, p. 28)
- *cultural racism*: explaining social inequities in terms of cultural differences between members of different groups – for instance, the argument that members of a given group are economically poorer because they do not value education (example taken from Bonilla-Silva, 2006, p. 28)
- *minimization of racism*: claiming that racism is no longer a major problem that significantly impacts peoples' lives; that inequality is due to other factors such as poverty or social class, not racism.

Importantly, beyond simply justifying inequality, when these frames are embedded into systems, they may obstruct opportunities to uncover and dismantle racist practices, and may even facilitate racist practices. In terms of abstract liberalism and minimization of racism, one can make the argument that excluding demographic variables, like race, from student risk prediction models creates a framework that reifies the idea that individuals are fully blameworthy for their own negative outcomes. If structural racism plays a role that goes beyond any other predictor or factor, then omitting race from prediction omits a key explanatory factor. Yet, even if such a system, focusing on attendance, grades, behavior, and test scores, could still capture some of the group disparities in graduation risk outcomes, it would not explain why these differences are occurring, and would unfairly implicate student behaviors as the source of these discrepancies.

In the literature on the "school to prison pipeline", for example, researchers have argued that structural racism leads to misdiagnosis of learning disabilities for black students and harsher punishments for the same action as other students (Darensbourg et al., 2010). In such cases it is unclear whether a predictive model would be more accurate if it included relevant demographic variables, because racist practices would likely influence already predictive behavioral variables (e.g., number of out-of-school suspensions; see discussion of mediating variables in section 6). However, when a teacher or school leader uses a model that does not include demographic variables to predict a student's outcomes and explain why these outcomes are likely, they may disregard structural factors in favor of individual explanations e.g., "students who misbehave are more likely to drop out." In such cases, school leaders may be inclined to ignore racial discrepancies in discipline, assuming that it is "natural" for such students to receive poor outcomes.

While the prior example might suggest that demographic variables should be used to avoid systemizing *abstract liberalism* in algorithms and their uses, it is worthwhile to consider unintended consequences. In particular, given the widespread and largely implicit nature of color-blind racism, including demographic variables as predictors may not primarily surface racist practices. Instead, demographic predictor variables may be interpreted in accordance with Bonilla-Silva's *cultural racism* and *minimization of racism* frames. For example, given racial discrepancies, a decision-maker might overlook school practices and instead interpret the use of the race as a predictor for a specific student as indicating that the cause is related to cultural values (e.g., devaluation of education) or other demographic features (e.g., social class). In these cases, including demographic indicators such as race communicates to decision-makers that the student is at-risk due to immutable factors beyond the control of the school.

In general, racism (including color-blind racism) manifests in complex ways and relies upon complex and intersecting conceptual frames. The practice of including demographic variables as predictors can interact with color-blind racism in complex ways, and depend in large part on how the models are intended to be used, how they are interpreted, and how they are actually used. To the extent that a model's interpretation and application attribute negative outcomes to immutable characteristics (i.e., demographic variables), actual racist practices that need redress will be overlooked. Therefore, while including demographic variables may seem to align with the practice of avoiding color-blind racism, such a practice can also lead to racist decision-making. The details of how models are developed, contextualized, and presented matter.

In this section, we have considered two of the major claims around why demographic variables should be used as predictors. The first, that it leads to better model performance, is intuitively persuasive but the evidence remains mixed and unclear at best. The second, that not doing so is akin to racial color-blindness, seems to apply in a more complicated fashion (and not entirely in favor of including demographic variables) once we consider in more detail what racial color-blindness consists of according to one of the key theoreticians and writers on the topic. It is important also to note that excluding demographic variables from graduation risk modeling does not mean ignoring differences by race – or other demographic factors – in educational outcomes. Indeed, checking predictive models for fairness and equal accuracy across different demographic groups is of great importance (see section 5.2 for further discussion). In the next section we argue that the cause of educational equity is best served by excluding demographic variables from use as predictors.

4. Arguments against using demographic data as predictor variables

4.1. Actionability

One concern about the use of demographic data as predictor variables is that models that incorporate demographic data may be less actionable than models that do not incorporate demographic data. This concern comes down to how models are used.

While prediction models are used to drive automated decisions by adaptive learning software (Corbett, 2001; D'Mello et al., 2010) and by designers of learning environments to iterate their designs (Koedinger et al., 2013; Li et al., 2022), the most common use of prediction models in education is to provide information to educators through dashboards. The two most common types of reports provided to instructors are reports on student risk and reports on student knowledge (which typically base their assessments solely on student correctness and time taken).

Dashboards on student risk typically provide information on which students are at risk and what factors are associated with that risk (Verbert et al., 2013). For example, the BrightBytes Student Success risk prediction system (Coleman et al., 2019) for K-12 provides educators with information on students' overall predicted risk of not graduating as well as their specific degree of absenteeism and lateness, course performance, formative assessments performance, and disciplinary incidents. Civitas, at the higher education level, provides educators with information such as declining GPA and standard deviation in GPA, procrastination in enrolling for courses, and courses withdrawn from (Milliron et al., 2014).

Dashboards on student risk are used by instructors, academic advisors, and school leaders to identify students at risk of dropping out, and design interventions for those students (Verbert et al., 2013). Designing effective interventions depends not solely on knowing which students will drop out, but also knowing why they will drop out (Herodotou et al., 2019). As such, the information provided to these instructors, academic advisors, and school leaders should be actionable – it should be possible to use it to select an intervention. While there are certainly examples of effective interventions that are targeted to all members of a specific demographic group in aggregate (e.g. Schwartz et al., 2018), it is more difficult to intervene for an individual student based on a demographic factor compared to more direct factors. For example, knowing that a student is frequently absent from school suggests immediate interventions that can be taken (e.g. DeSocio et al., 2007), as does knowing that a student has performed poorly or started their work on assignments later than other students (Hafner et al., 2014). A student's propensity to be absent, to perform poorly on assignments, or to procrastinate can be influenced.

Furthermore, demographic variables are often correlated with more direct predictors. Due to historic inequities, students in historically underrepresented demographic groups are more likely to experience poverty or lack of prior preparation that can in turn lead to some of the more direct signs of student risk. Yet, in such cases, the manifestation of historic inequities may differ from student to student. For example, while two students may face similar disadvantages, one student may be chronically absent due to caregiving needs at home, while another student may disengage or "act out" in class. While both students may be "at risk", the appropriate intervention may be occluded by focusing on demographic factors. By their design, machine learning algorithms figure out which variables are the best predictors, and de-emphasize variables that do not contribute additional predictive power once those variables are included. As such, including demographic variables in models can cause the model to de-emphasize correlated risk factors that are more directly aligned with the particular needs of the student – even if the demographic variable achieves only slightly better fit than the correlated variables or achieves worse fit than the correlated variables combined. Real-world models that incorporate demographic variables often emphasize those variables over other factors, in cases with single demographic variables accounting for over a quarter of the predictive power of the models (see tables in Feathers, 2022).

Demographic variables may also be directly associated with student risk – for instance, through overt racism on the part of school administrators pushing students out, stresses associated with homelessness, or challenges inherent to poverty. On the left side of Figure 1, we present a path diagram to demonstrate a pattern showing this hypothetical relationship, following recommendations to communicate causal relationships using diagrams of this nature (Hicks et al., 2022; Weidlich et al., 2022). This diagram shows possible relationships between a demographic variable – in this case homelessness – an academic variable (attendance), and graduation outcomes. In this diagram, homelessness has a direct impact on graduation outcomes (due to factors that can't be assessed at the school level), and homelessness also has an impact on attendance rates, which in turn impact graduation outcomes. With correlation-based machine learning techniques, it is not possible to separate the causal impact of each individual factor on the outcome, even if the relative additional predictive utility each factor provides can be estimated (Archer & Kimes, 2008). In this Figure, the effect of homelessness and attendance on graduation rates are confounded. Including homelessness in the model may make it possible to predict graduation more accurately, but the model's ability to assess the role played by the more actionable variable – attendance – is reduced due to the confounding of attendance and homelessness status. This can hypothetically play out in even more complex ways, as the right side of Figure 1 shows, where homelessness correlates with graduation but also with multiple predictors.

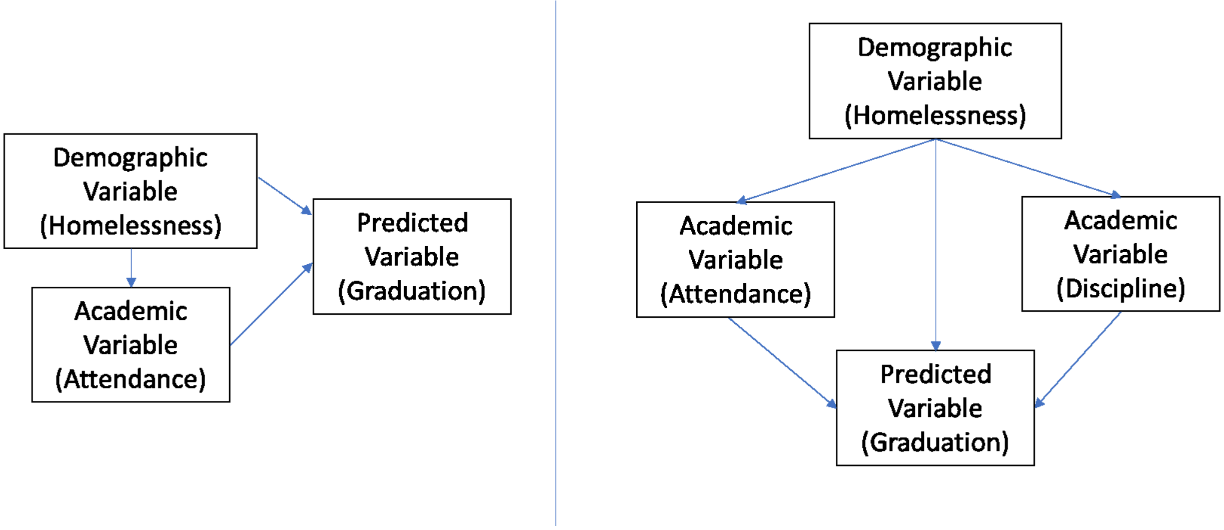
Demographic Variable (Homelessness)
Predicted Variable (Graduation)
Academic Variable (Attendance)
Demographic Variable (Homelessness)
Academic Variable (Attendance)
Academic Variable (Discipline)
Predicted Variable (Graduation)

Figure 1: Diagrams of relationships between variables where including a demographic variable can obscure more actionable variables.

While demographic variables such as homelessness and race are correlated with student graduation outcomes – and schools can take action to support students in those groups – the variables themselves are not malleable by the school. Actions that schools take to support Black male students, for example, may change the students' attendance rates, disciplinary interactions, and grades, but won't change the students' race. Therefore, not only does prediction based on demographic variables make it more difficult to target an intervention directly to the student's immediate risk factors, a risk prediction model based on demographic variables risks discounting the impact of the intervention on the student's graduation risk as it continues to include and make predictions based on the static racial factor. Including these immutable demographic characteristics within a risk prediction system amounts to embedding a form of demographic destiny – it assumes that no intervention will change the negative outcomes initially associated with a specific demographic group of students. If we think that a school can change the trajectory of members of a group of students – which many schools have successfully done – then a model which includes demographic variables in at-risk prediction will actively ignore the progress being made. It may also obscure the needs of students within that group who still need support, grouping them with students in that group now on track to successful outcomes.

This is not to say that it is not valuable to know that members of a specific demographic group are generally at higher risk of dropout (or some other negative outcome) than members of other groups. Knowing these relationships can help in the design of group-wide interventions that are applicable to many learners. For example, evidence on lower retention of female students in computer science programs has led to comprehensive programs that address the inequities leading to these poorer outcomes (Pantic & Clarke-Midura, 2019). However, this type of problem is not best found through a model that makes predictions about individual students – to discover these group-level problems, it is better to use statistical models that were designed from the start to represent and communicate relationships at the group level (see, for instance, Robison et al., 2017). Group-level interventions can be developed based on statistical modeling and the effects of these interventions can be assessed using the same statistical models. Doing so avoids presenting an individual decision-maker intervening to immediately support an individual student with information that is not immediately actionable, and avoids the assumption that no intervention will change the negative outcomes initially associated with a specific demographic group of students.

Overall, then, there is some reason for concern that incorporating demographic variables into predictive models in education may reduce their usefulness for one of their key applications – providing information on which specific students are at risk and why, so that educators can use that information to benefit those students.

4.2. Reinforcing the bias in training labels

A related possible problem with using demographic variables as predictors comes from the risk of reinforcing the biases in training labels, where training labels partially represent the discrimination occurring in society rather than what they were intended to represent. Training labels (and by extension, the real-world situations that they represent) incorporate the biases already present in the world. The bias occurring in training data can occur in either direct or indirect fashions. Indirect incorporation of biases comes about when the conditions in the world (or even in the individual we are making predictions about) are such that their outcomes are poorer than might be expected from other variables. For example, the experience of systematic inequities in a student's life may make them more likely to drop out of high school than other students whose in-school variables are seemingly identical. Direct incorporation of biases comes about when decisions are made in the predictive context by stakeholders that are biased, leading to a worse outcome. For example, a high school assistant principal may be more likely to suspend students of one demographic group than a student from another demographic group, even when all other factors are identical (Shollenberger, 2015).

Situations such as these – already problematic – can be worsened by the use of demographic variables as predictors. Take, for instance, the case where one is attempting to predict student course grades, either by a combination of assignment grades and participation, or by those factors along with student demographics. One paper, published in the early years of the International Conference on Learning Analytics and Knowledge, reported that student course grades could be predicted better if a demographic variable (unnamed in the paper) was included in models (Wolff et al., 2013). In discussion of that paper during the question period, the authors revealed that the demographic variable was student race, and that when two students had identical assignment grades and participation, including race in the model could separate out a student who achieved a lower course grade versus a higher course grade, with students of a historically disadvantaged racial group achieving lower course grades. The authors viewed this as evidence that better prediction of student outcomes – and therefore risk – could be achieved by including race. While true, this situation also represents a clear case of the model encoding biased decisions made directly by stakeholders (the students' instructors). If a student in a historically disadvantaged group is likely to be assigned a worse course grade by an instructor than their performance merits, this is perhaps a problem that cannot be resolved by informing the instructor that these students are at greater risk.

There are related challenges in using demographic variables in reporting on why a student is at risk. Typically, predictive models used in education today report why they are making a prediction. Telling an instructor or other stakeholder that a student is likely to achieve a negative outcome *because they are a member of an underrepresented group* is problematic – especially so if that prediction is due to bias in the training data. Telling instructors and school leaders (who already make biased decisions, for instance by disciplining students differently for the same behavior – Shollenberger, 2015) that students in the disadvantaged group are higher risk may even reinforce the same beliefs that had already led to students in that group being assigned lower grades or provide a justification for that action. For example, if an instructor believes that differences in outcomes are due to natural differences between groups and are not worth acting on – Bonilla-Silva's (2006) naturalization frame of racism – telling them that the student will perform worse because of their group membership risks reinforcing this belief. An analogy can be seen in (Corbett-Davies et al., 2017), where replicating the biases in the training data led to biased recommendations and biased outcomes. There is also a risk of creating a self-fulfilling prophecy (Rosenthal & Jacobson, 1968), where the instructor believes race causes worse outcomes and therefore acts in ways that bring about that outcome.

In other words, if there is direct bias in the decisions that the model is trying to predict, including these direct biases in prediction may reinforce the beliefs that produce those biased decisions. As Paquette et al. (2020) note, "...if instructors use more discretion to lift grades for students of one group than another group, including group status in the model may capture these differences in risk without understanding or treating the systematic causes of the differences."

In extreme cases (e.g. Feathers, 2022), when the training data is biased, a predictive model may make a prediction for a specific student based entirely or almost entirely on demographic variables. This may result in prediction that is not very informative about individual students and says that all members of one group are high-risk while all members of another group are low-risk. Presenting this type of information to decision-makers is likely to have one of two outcomes: reinforcing their existing biases or causing the decision-maker not to use the model (the latter choice is reported in Feathers, 2022).

Thus, when training data is biased, instead of simply including demographic variables in prediction, it may be better to use statistical analysis to demonstrate that the demographic variable is being used as a basis for decision and intervene at the decision-maker or system level rather than at the student level. This logic applies in terms of indirect biases in training data as well. Take, for instance, an undergraduate program where students from historically underrepresented groups are more likely to drop out (in this case due perhaps to experiences such as the feeling of non-representation or microaggressions impacting the decision to drop out, or greater economic pressure). While it is important to identify that students in a specific group are at higher risk, there may be better ways to address that risk than incorporating it into individual predictions. Instead of incorporating the demographic variable into each student's prediction, it may be more appropriate to conduct systemic interventions in the program to identify and address these factors (see, for instance, Schwartz et al., 2018; Pantic & Clarke-Midura, 2019).

As a final note, using demographic variables when training sets are already biased is problematic not just for learners in historically disadvantaged groups. Even though there are several concerns for these learners, as this section indicates, there are also potential risks for learners in historically advantaged groups. Take a student who is a member of a historically advantaged group. Including their demographic variables in predictions about them may reduce the assessment of their risk when they are genuinely at high risk. In other words, if an algorithm relies heavily upon demographic variables (e.g. Feathers, 2022) then a white male student from a wealthy background may be reported as low-risk despite failing classes and obtaining low grades, high absenteeism and lateness, and a high rate of disciplinary violations. This student will receive an unfairly optimistic prediction, making it less likely they will receive the support they need.

4.3. LIMITATIONS IN CATEGORIZATION

A third major area of challenge in using demographic categories as predictor variables is the limits in the categorization schemas used to identify those demographic categories. The demographic variables we typically have access to have been created through political processes (Lee, 1993). The impulse to categorize people into groups has been influenced by a range of voices throughout the last century, with varying motives and goals – ranging from the desire to limit immigration to the desire to equalize political representation and funding between groups (Snipp, 2003); our current system of categorization is highly influenced from the start by racial theories that assigned characteristics to different racial categories (Lee, 1993). The demographic categories today used to describe race and ethnicity in most formal records within the United States are overseen by a budgetary office, while some U.S. states maintain different schemas (such as the consideration of Brazilian ancestry in Massachusetts) and some U.S. states collect additional demographic variables not seen in other states (such as military-connected status for children in Texas schools). Around the world, the demographic categories typically collected and considered depend on political circumstances, from the *bumiputra* categorization in Malaysia to the *Scheduled Castes* and *Scheduled Tribes* of India, to the categorization of Roma and Irish Travellers as sub-categories of White in the United Kingdom and the different ways the Roma people are categorized around the world. Indeed, in some countries, legislation makes it infeasible to systematically collect specific demographic variables, such as race (Blech, 2001), making it challenging to study these forms of algorithmic bias at all. Debates continue to this day about how different groups should be categorized, seen for instance in the recent debates in the United States about whether White Hispanics should be considered differently than other Hispanics in terms of university admissions decisions. And until recently, non-binary and transgender individuals were classified for administrative purposes in most countries and contexts as either male or female, with no ability to choose the category that best fit their own view of their gender (National Academies of Science, Engineering, and Medicine, 2022).

One immediately notices the degree to which almost any schema for categorization must by nature oversimplify and be arbitrary. Two of the first author's children, who are of one quarter Spanish ancestry, are told they are not Hispanic because their mother was born in Brazil. The British categorization of Asian incorporates individuals from China, India, and Pakistan, groups that have had very different histories in Britain. In the United States, some have critiqued the situation where affirmative action policies designed to address inequities impacting the descendants of families affected by slavery and Jim Crow instead support more recent immigrants.

The use of demographic variables as predictors assumes that there is commonality between individuals labeled by that demographic variable. The degree to which this is true will depend on the true degree of common history and experience relevant to the prediction among individuals with that demographic variable. If there is significant heterogeneity within a group, then the use of the group variable will disadvantage atypical members. The use of the variable "White" in the United Kingdom will probably not represent members of the Roma and Irish Traveller subgroups well. The use of the variable "African-American" in the United States may not represent the children of recent immigrants from Ghana or Somalia very well. In Seattle, for example, nearly 40% of "Black / African-American" students speak a language other than English at home (Cooley et. al 2021). A variable as broad as "Asian" in the US or UK is likely to represent relatively few individuals well, given the considerable diversity in culture and economic circumstances within this group, and many have argued for differentiating within this group for purposes such as affirmative action (Brest & Oshige, 1995). A model that relies heavily upon these variables may produce models with better prediction for typical members, but worse prediction for atypical members – giving less emphasis to individual factors and biasing towards a group mean.

The heavy reliance on demographic variables also creates challenges for individuals whose groups are less well-represented in data. In many cases, a given demographic group may not have sufficient membership in the training data for their identity to be predictive – Baker & Hawn (2021) note that Native Americans are often insufficiently represented in training sets in education in the US. Other demographic groups may be classified as part of broad "Other" categories, such as Native South Americans who do not classify as Hispanic. The use of demographic variables as predictors can create worse predictions for these individuals, even if they are not represented in those demographic variables. For example, if there is an important (and actionable) mediating variable between the demographic variable and the outcome (see diagram in section 4.1), then the inclusion of the demographic variable may eliminate the mediating variable from the model. This may not be directly problematic (aside from the loss of actionability) for the included group, but if the same mediational pathway applies to other underrepresented groups, then prediction may be worsened for members of these groups. For example, if students who are African American, Native (North) American, Native Canadians/First Nation, and Native South American all receive disproportionate discipline for minor infractions, but there is only sufficient data to include African American as a variable in the model, then including that variable may eliminate disciplinary infractions from the model, reducing model effectiveness for Native (North) Americans, Native Canadians/First Nations individuals, and Native South Americans while only mildly improving model performance for African American students.

5. Contextual factors influencing whether and how demographic variables are used

5.1. Cases where demographic factors produce different levels of risk

As the previous sections have attempted to show, there are a range of reasons why one may or may not choose to use demographic variables as predictors in a prediction model in education. Taking for the moment a reader who is neither convinced that demographic variables should always be avoided as predictors, nor that demographic variables should always be utilized in this fashion. For this reader, what factors should influence their decision to use demographic variables or not use demographic variables as predictors in a specific situation?

One key question is the degree to which using demographic data as predictors increases the likelihood of discriminatory decision making. In some cases, as discussed above, it seems plausible that using demographic data as predictors could increase the risk of this. For example, take a Masters program that uses these variables as predictors in an algorithm, prior to admissions, to determine whether a student is more likely to succeed in the program (for instance, grades in the first year or successful completion). If that algorithm is used to select candidates – only selecting students for admission if they are likely to succeed – then we have cause for concern. For example, if some demographic group is less likely to succeed in the program due to economic or historical reasons, then a success prediction algorithm using that information will systematically exclude those students from participation in the program: effectively reinforcing historic inequities. Or take, for instance, an algorithm predicting the future probability of a student becoming involved in school violence. If this algorithm uses demographic data as predictors, and members of some demographic groups are more frequently reported as being involved in school violence – again perhaps due to economic or historical reasons – then the algorithm may systematically overstate some students' risk and understate other students' risk due to factors outside of their control. Depending on what the intervention is, innocent members of both of these groups may be unfairly disadvantaged: for more punitive interventions (such as greater surveillance or stronger punishment for the same behavior) students in a higher-risk demographic group may be unfairly punished, whereas for more positive interventions (such as access to after-school activities or support programs) students in a lower-risk demographic group may be unfairly excluded.

By contrast, using demographic data as predictors in a prediction model may be less risky when it is less likely to drive discriminatory decision-making. For example, it may be less risky to use this type of data when informing a student of their own level of risk or potential future outcome. For instance, in adaptive learning systems that offer students a choice of activities (for instance, mathematics problems involving different cover stories), it may not be problematic for an algorithm to rank those activities in a recommendation to students using demographic information, since it is still the student's choice which activity they select and the cost of choosing poorly is relatively low (the student works on a single activity that is less interesting than it could be). To give another example, using demographic information as predictors in college application recommendations may not be problematic, since the student still chooses where to apply. Take, for example, a college that offers programs designed to support first-generation students. If this college is unusually successful in improving graduation rates for first-generation students, it may be beneficial for an algorithm to take this information into account when recommending which colleges a first-generation student should apply to. In general, then, using demographic data as predictors may be less harmful if the person being impacted ultimately still has the final choice about what happens, or if the stakes involved in the decision are low.

However, even these examples suggest some possible cause for concern. Again consider activity ranking within a recommendation system. Recommending activities based on a student's declared sex may be undesirable for students who have gender-atypical interests (Haldeman, 2000); in general, if handled poorly, one can easily imagine a system recommending content in a stereotypical way, based solely on very slight average preferences by group members being amplified over time by the algorithm's behavior (i.e., slight initial differences will be magnified by the system's emphasis in rankings – students will not choose the options ranked lower for them, solely because of those rankings). Take also college recommendations – over time, and in aggregate, recommendations explicitly based on a demographic factor will tend to drive students in specific groups towards and away from specific colleges. Small shifts in which students apply to a college may over time lead to major disparities in who attends that college. It is also worth noting that the use of race in college recommendation already has some history of prejudice – such as recommendation systems where specific colleges have set flags to request they not be advertised to specific groups of historically underrepresented students (Feathers, 2022). This history of prejudice raises concerns about whether the use of demographic data as predictors in this type of model will be conducted in appropriate fashions.

5.2. OTHER USES OF DEMOGRAPHIC VARIABLES IN PREDICTION MODELS

Much of the discussion of the use of demographic variables in prediction models (including within this paper) has focused on their use as predictors. However, it is also worth pointing out that there are other uses for these variables, that may not share some of the concerns and limitations that using them as predictors can cause.

For example, demographic variables can be used to apply fairness constraints in modeling without explicitly using them as predictors (Zafar et al., 2019). Fairness constraints specify that only models with certain attributes should be considered. There are a range of fairness constraints that can be considered, including *disparate treatment* (a model cannot make different decisions for two cases where all variables are identical except for a demographic variable), *disparate impact* (a model should not assign a disproportionately large fraction of positive or negative outcomes to a specific demographic group), and *disparate mistreatment* (a model should not have different accuracy/error rate for different groups) (Zafar et al., 2019). It is worth noting that the disparate treatment constraint actively contradicts the use of demographic variables as predictors. Although it is mathematically impossible to meet all standardly considered fairness constraints at once (see review in Lee & Kizilcec, 2020), attempts can be made to balance across multiple constraints, and algorithmic approaches such as cost-sensitive classification have been successful in some cases at increasing model compliance with fairness constraints (Agarwal et al., 2018; for predicting student dropout, see Vasquez Verdugo et al., 2022).

There are a variety of ways that demographic variables can be used to apply fairness constraints, and depending on how they are used, the concerns raised in section 4 of our paper may or may not be relevant. While the goal of fairness constraints is overall to avoid the types of problems that using demographic variables as predictors can increase, emphasizing some fairness constraints can worsen performance for others (see Lee & Kizilcec, 2020). While ensuring that a model is accurate for all groups (disparate mistreatment) is unlikely to produce consequences for actionability or encourage prejudiced beliefs within decision-makers, setting a constraint of disparate impact may result in inaccurate beliefs about how serious a problem is that is impacting specific groups of students. Thus, while promising, further study is needed to fully understand the impacts of the many ways that demographic variables can be used to apply fairness constraints.

Furthermore, demographic variables can and should be used for fairness audits even when they are not used to produce a prediction. Fairness audits, where a model is inspected for its fairness towards different demographic groups of students, are starting to become part of educational practice, with Digital Promise offering a product certification in Prioritizing Racial Equity in AI Design (Digital Promise, 2021). Fairness audits also are becoming an increasing part of scientific publishing in learning analytics, with increasing publication in this area (see Baker & Hawn, 2021) and increasing calls for papers to report on algorithmic bias and fairness (Baker & Hawn, 2021; Holstein & Doroudi, 2022). Fairness audits typically rely on checking whether fairness constraints are met for demographic variables, with the majority of fairness audits in education focusing on measuring for disparate mistreatment (this practice is also sometimes referred to as slicing analysis – Gardner et al., 2019). Using demographic variables in fairness audits generally does not invoke the concerns raised in section 4 of our paper, as the goal of fairness audits is to check for the types of problems that using demographic variables as predictors can increase. Instead, using fairness audits can produce important knowledge that can be used both summatively (to demonstrate that an algorithm does not have certain forms of bias) and formatively (to identify bias in order to address it).

6. PROXY VARIABLES AND MEDIATING VARIABLES

Another case worth considering – one that is often discussed in the literature on algorithmic bias within other fields – is the case of biased proxy variables. A biased proxy variable (Johnson, 2021) is when a seemingly non-problematic variable (the proxy variable) correlates sufficiently highly with a demographic variable that using the proxy variable produces results (and bias) similar to using the demographic variable (perhaps to a lesser degree). In addition, a demographic variable may be captured by multiple proxy variables working in concert, rather than just a single proxy variable.

To those who find it unproblematic to predict using demographic variables, proxy variables are similarly unproblematic. To those with concerns about predicting using demographic variables, proxy variables raise similar concerns (e.g. Matloff & Zhang, 2022), and add the possible issue of being hard to identify a priori (Johnson, 2021).

One challenge in thinking about proxy variables is distinguishing them from mediating variables. A mediating variable is one where some demographic factor impacts the mediating variable, but which in turn impacts the outcome being predicted. The difference between a proxy variable and a mediating variable is that the proxy variable primarily correlates to the variable being predicted *because of the correlation the demographic variable has to the outcome.* In other words, a proxy variable's relationship to the predicted variable is actually due entirely to the demographic variable. By contrast, a mediating variable is highly correlated to the predicted variable on its own, even when the demographic variable is not present. The difference between these relationships is shown in Figure 2.

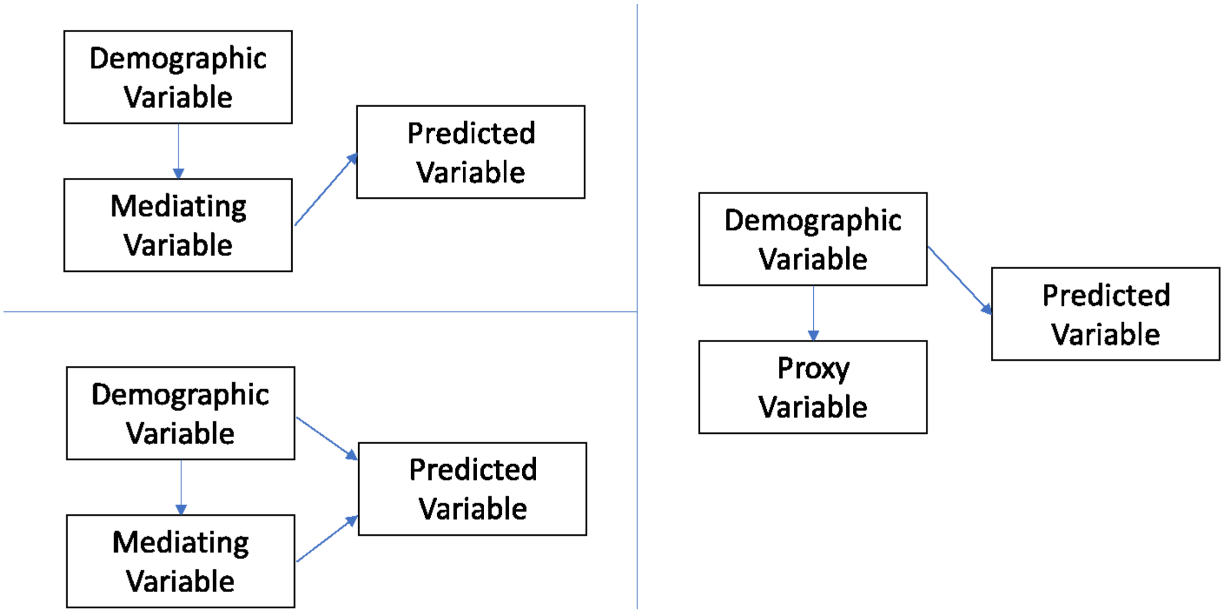

Figure 2: Diagram differentiating mediating variables from proxy variables.

To give an example of how these relationships could play out, take first the case of a mediating variable. Perhaps a demographic variable (poverty) causes school absences, but school absences cause school dropout regardless of whether the student is experiencing poverty. In this case, school dropout is a mediator variable, and its inclusion is not simply including poverty in the model without realizing it. Take instead the case of a proxy variable. Perhaps students experiencing poverty receive a special "government cheese" from the government (a common situation in past decades in the United States). Only students in poverty receive government cheese, and no students not in poverty receive government cheese (in this example). In that case, the correlation between poverty and dropout will be identical to the correlation between government cheese and dropout. Excluding poverty from the model but including government cheese would not resolve any biases in the model caused by including poverty.

There are many examples of proxy variables in algorithmic bias. The use of zip codes can serve as a proxy for race and socioeconomic status (Jackson, 2018), and for other factors based on those original demographic variables. Possessing an uncommon last name for a data set can serve as a proxy for race or national origin (Lowry & Macpherson, 1988). Having been repeatedly stopped and frisked by police without a known cause can serve as a proxy for race (Babuta & Oswald, 2019).

Mediating variables, by contrast, can be more actionable than the demographic variables themselves. Take the example of school disciplinary incidents. Disciplinary incidents are frequently included in models predicting high-school dropout (e.g. Coleman et al., 2019; Christie et al., 2019). However, it is known that some school administrators punish Black students more often than White students for the same behavior (Shollenberger, 1995). As such, disciplinary incidents are related to race (and ethnicity, and gender). However, disciplinary incidents still typically reflect some behavior, and that behavior is plausibly related to dropout. Using disciplinary incidents as a dropout predictor can lead to then finding ways to reduce disciplinary incidents, whether at a school-wide level or at an individual level – both of which have been shown to reduce dropout in cases where implementation fidelity is high and interventions are sustained (Freeman et al., 2015; Ecker-Lyster & Niileksela, 2016). As such, disciplinary incidents are a mediating variable and they are actionable. Telling a school administrator that a specific student is at risk of dropout because they have many disciplinary incidents has the potential to lead to positive impacts. In cases where the mediating variable is a punishment rather than a behavior – such as school suspension – using this variable to predict and inform action may even moderate existing biased behaviors. For instance, an administrator who would be inclined to (unfairly) suspend a Black student for behavior that would lead to a lesser punishment for a White student may be influenced to avoid the suspension, given its association to drop-out. It may even be possible to note the relationship between the demographic and mediating variable in communicating the problem to the decision-maker – for instance, by noting that Black students may receive suspensions in situations where White students do not, or by (to return to an example from earlier) noting that a student experiencing homelessness may be absent for reasons outside of their direct control.
In some cases, it can be difficult to distinguish mediating variables from proxy variables. Take the variable of having had experience of being stopped and frisked by a police officer without having committed any crime (Babuta & Oswald, 2019). While this is primarily a proxy variable for race/ethnicity when used to predict committing violent crime, it is possible that the experience of being stopped and frisked may itself impact the probability of committing a crime regardless of race/ethnicity. In this case, the variable has a mediating component – but the extremely strong association of this variable with race/ethnicity (Gelman et al., 2007) makes the proxy aspect of the variable much more prominent.

Overall, then, proxy variables share the same concerns as using demographic variables themselves. Mediating variables, by contrast – when they can be carefully assessed to be mediating rather than proxy – may have a place even for cases where an algorithm developer decides not to use demographic variables.

7. Conclusions

Table 1. The potential pros and cons of using demographic variables as predictors.

Pros (Reasons to use demographic variables as predictors)	Cons (Reasons not to use demographic variables as predictors)

May lead to better prediction (evidence mixed – see section 3)	
May avoid systematizing *abstract liberalism,* associated with colorblind racism (section 3)	May create risk of interpreting model outputs in terms of *cultural racism,* associated with colorblind racism (section 3)
	May suppress more actionable variables (section 4.1)
	May amplify biases in training labels (section 5.1) and reinforce prejudiced beliefs that led to biased training labels (section 4.2)
	May result in overly optimistic predictions for at-risk members of groups with overall positive outcomes (section 4.2)
	May reduce model effectiveness for members of small groups or members of sub-groups that do not resemble their overall groups (section 4.3)

In this article, we have discussed the practice of putting demographics in a model as predictor variables. Many researchers and practitioners have argued that this practice is not only beneficial, but a moral imperative, and that not following this practice is a racist decision (e.g. Yu et al., 2021).

In this paper, we have attempted to investigate this methodological decision, within the context of at-risk prediction models. Across the literature, it remains unclear whether one of the primary goals of including these variables as predictors – more accurate prediction – is upheld. On the other hand, several potential negatives of this practice are found. Including demographic variables as predictors can reduce actionability, can reinforce racist beliefs held by stakeholders, can reduce model effectiveness for members of historically underrepresented groups who are not well represented in demographic category labels, and can lead to biased decisions for individuals who are atypical members of the demographic groups. These types of impacts are often not considered in papers discussing the use of predictive analytics in education, though ultimately actionability is often more important than moderate improvements in model goodness metrics, and high overall model goodness can mask poor performance for specific subgroups of students. More explicit consideration and measurement of these possible impacts will help to understand whether and the degree to which this modeling choice produces unintended consequences. For example, actionability could be measured by collecting and studying the actual decisions made by decision-makers based on these models, including their self-reports on why they made a specific decision, and changes in belief could be measured by measuring racist beliefs (e.g. Morrison & Kiss, 2017) before and after use of a system for several months. It also may be valuable to use ethnographic methods to understand how decision-makers interpret and use a system's predictions and interpretations – i.e. when a model indicates that a student is at risk because of a demographic factor, how do decision-makers use this information, and does it lead to or reinforce the decision-makers prejudiced beliefs?

Our discussion also suggests that the choice of whether or not to use demographic variables as predictors can be shaped by how the model is being used. Cases where a model is intended to be used by the learner themself may present lower risk than cases where a model is intended to be used to drive decision-making by an outside stakeholder. In addition, as noted in the discussion around mediating variables, demographic data may provide useful context to decision-makers around why a more actionable predictor is present. Finding ways to communicate this context to decision-makers (perhaps through a causal graph) could lead to enhanced actionability, rather than reduced actionability. Thus, future work should more explicitly consider use cases and investigate the impacts of the use of demographic variables in these different situations and designs.

In considering use cases and situations, future treatments of the issues discussed here should consider whether the use of models and the context they operate in differs in important ways in different cultural and political contexts. Most of the literature on algorithmic bias in education has occurred thus far in the United States and other English-speaking countries (see review in Baker & Hawn, 2022). Our positionality as researchers currently living and working in the United States means that we may have missed key issues impacting the way algorithmic bias plays out in the rest of the world. We sincerely hope that research will emerge to a greater degree worldwide on the effects and manifestations of algorithmic bias. When this occurs, it will be important to re-consider the issues discussed here in those broader contexts.

Our recommendation not to use demographic data as predictors does not, however, mean that demographic data should simply be ignored in at-risk prediction modeling. Other uses of demographic variables pose potential benefits while avoiding the risks involved in directly using demographic variables as predictors. For example, the use of demographic variables in slicing analyses or fairness audits – checking for violation of fairness constraints by a model – poses none of the risks identified in this paper, while helping to catch cases where a model is unfair. In addition, approaches such as cost-sensitive classification can increase a model's degree of compliance with fairness constraints, reducing the degree of algorithmic bias in a model.

Overall, our recommendation is that the practice of using demographic variables as predictors is risky. Rather than treating its use as a necessity and tarring those who choose not to use it as engaging in color-blind racism, this practice should be used sparingly and carefully and clearly justified when it is used. A paper using this practice should identify why the specific risks identified here are not problematic for their specific context of use, and check for negative impacts of the nature discussed here. There are situations where the practice is warranted, but those cases must be clearly identified, and this practice should be treated as a risky practice (akin, perhaps to the decision not to use a held-out test set) that occasionally can be justified.

ACKNOWLEDGEMENTS

We would like to thank Chelsea Porter for document editing assistance.

REFERENCES

AGARWAL, A., BEYGELZIMER, A., DUDIK, M., LANGFORD, J., AND WALLACH, H. (2018) A reductions approach to fair classification. In *Proceedings of the 35th International Conference on Machine Learning*, N. Lawrence, Ed. Proceedings of Machine Learning Research, 60-69.

ALLENSWORTH, E. M., AND EASTON, J. Q. (2007). What matters for staying on-track and graduating in Chicago public high schools: A close look at course grades, failures, and attendance in the freshman year. Research Report. Consortium on Chicago School Research.

ALMEDA, M. V., AND BAKER, R. S. (2020) Predicting student participation in STEM careers: The role of affect and engagement during middle school. *Journal of Educational Data Mining 12,* 2, 33-47.

ANDERSON, H., BOODHWANI, A., AND BAKER, R. (2019) Assessing the fairness of graduation predictions. Poster paper. In *Proceedings of the 12th International Conference on Educational Data Mining (EDM 2019)*, C. F. Lynch, A. Merceron, M. Desmarais, and R. Nkambou, Eds. International Educational Data Mining Society, 488-491.

ANDRUS, M., AND VILLENEUVE, S. (2022). Demographic-reliant algorithmic fairness: Characterizing the risks of demographic data collection in the pursuit of fairness. In *Proceedings of the 5th ACM Conference on Fairness, Accountability, and Transparency (FAccT '22)*, Association for Computing Machinery, 1709-1721. https://doi.org/10.1145/3531146.3533226

ANGWIN, J., LARSON, J., MATTU, S., AND KIRCHNER, L. (2016). Machine bias. In *Ethics of Data and Analytics,* Auerbach Publications, 254-264. http://dx.doi.org/10.1201/9781003278290-37

ARCHER, K. J., AND KIMES, R. V. (2008). Empirical characterization of random forest variable importance measures. *Computational Statistics & Data Analysis 52,* 4, 2249-2260. https://doi.org/10.1016/j.csda.2007.08.015

ASBELL-CLARKE, J., ROWE, E., BARDAR, E., EAGLE, M., BROWN, R., BAKER, R., BARNES, T., AND EDWARDS, T. (2015). Leveling Up: Measuring and leveraging implicit STEM learning in games. In *Proceedings of the 11th Annual Conference on Games+Learning+Society (GLS 2015)*, K. E. H. Caldwell, Ed. Carnegie Mellon University, 306-313.

BABUTA, A., AND OSWALD, M. (2019). Data analytics and algorithmic bias in policing. The Royal United Services Institute for Defence and Security Studies.

BAKER, R. S., BERNING, A. W., GOWDA, S. M., ZHANG, S., AND HAWN, A. (2020). Predicting K-12 dropout. *Journal of Education for Students Placed at Risk (JESPAR) 25*, 1, 28-54. https://doi.org/10.1080/10824669.2019.1670065

BAKER, R. S., AND HAWN, A. (2021). Algorithmic bias in education. *International Journal of Artificial Intelligence in Education 32,* 1, 1052-1092. https://doi.org/10.1007/s40593-021-00285-9

BAKER, R. S. J. D., AND ROSSI, L. M. (2013). Assessing the disengaged behavior of learners. In R. Sottilare, A. Graesser, X. Hu, and H. Holden, (Eds.), *Design Recommendations for*

Intelligent Tutoring Systems – Volume 1 – Learner Modeling, U.S. Army Research Lab, 155-166.

BAROCAS, S., AND SELBST, A. D. (2016). Big Data's Disparate Impact. *104 California Law Review 671*, 671-732. http://dx.doi.org/10.2139/ssrn.2477899

BASSOK, D., AND REARDON, S. F. (2013). "Academic redshirting" in kindergarten: Prevalence, patterns, and implications. *Educational Evaluation and Policy Analysis 35,* 3, 283-297. https://doi.org/10.3102/0162373713482764

BELCHER, D. C., AND HATLEY, R. V. (1994). A dropout prediction model that highlights middle level variables. *Research in Middle Level Education 17,* 2, 67-78. https://doi.org/10.1080/10825541.1994.11670032

BERK, R., HEIDARI, H., JABBARI, S., KEARNS, M., AND ROTH, A. (2018). Fairness in criminal justice risk assessments: The state of the art. *Sociological Methods & Research 50,* 1, 3–44. https://doi.org/10.1177/0049124118782533

BLECH, E. (2001) Race Policy in France. Washington, DC: Brookings. Retrieved January 26, 2023 from https://www.brookings.edu/articles/race-policy-in-france/

BONILLA-SILVA, E. (2006). *Racism without racists: Color-blind racism and the persistence of racial inequality in the United States.* Rowman & Littlefield Publishers.

BORNSHEUER, J. N., POLONYI, M. A., ANDREWS, M., FORE, B., AND ONWUEGBUZIE, A. J. (2011). The relationship between ninth-grade retention and on-time graduation in a southeast Texas high school. *Journal of At-Risk Issues 16,* 2, 9-16.

BOWERS, A. J. (2021). Early warning systems and indicators of dropping out of upper secondary school: The emerging role of digital technologies. *OECD Digital Education Outlook 2021: Pushing the Frontiers with AI, Blockchain and Robots*, 173-194.

BOWERS, A. J., SPROTT, R., AND TAFF, S. A. (2012). Do we know who will drop out? A review of the predictors of dropping out of high school: Precision, sensitivity, and specificity. *The High School Journal 96,* 2, 77-100.

BOYCE, J., AND BOWERS, A. J. (2016). Principal turnover: Are there different types of principals who move from or leave their schools? A latent class analysis of the 2007–2008 schools and staffing survey and the 2008–2009 principal follow-up survey. *Leadership and Policy in Schools 15,* 3, 237-272. https://doi.org/10.1080/15700763.2015.1047033

BREST, P., AND OSHIGE, M. (1995). Affirmative action for whom?. *Stanford Law Review 47,* 5, 855-900. https://doi.org/10.2307/1229177

CATON, S., AND HAAS, C. (2020). Fairness in machine learning: A survey. *arXiv preprint arXiv:2010.04053*. https://doi.org/10.48550/arXiv.2010.04053

CHOULDECHOVA, A. (2017). Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *Big data 5*, 2, 153-163. https://doi.org/10.1089/big.2016.0047

CHRISTIE, S. T., JARRATT, D. C., OLSON, L. A., AND TAIJALA, T. T. (2019). Machine-learned school dropout early warning at scale. In *Proceedings of the 12th International Conference on Educational Data Mining (EDM 2019)*, C. F. Lynch, A. Merceron, M. Desmarais, and R. Nkambou, Eds. International Educational Data Mining Society, 726-731.

COLEMAN, C., BAKER, R., AND STEPHENSON, S. (2019) A better cold-start for early prediction of student at-risk status in new school districts. In *Proceedings of the 12th International*

Conference on Educational Data Mining (EDM 2019), C. F. Lynch, A. Merceron, M. Desmarais, and R. Nkambou, Eds. International Educational Data Mining Society, 732-737.

COOLEY, S., BYRDO, N., WILLIAMS, M., KING, W., BAIRS, S., HAIZLIP, A., AND LOYAL, K. (2021). Our Voice Our Vision. Seattle Public Schools, Seattle WA.

CORBETT, A. (2001). Cognitive computer tutors: Solving the two-sigma problem. In *Proceedings of the 8th International Conference on User Modeling,* M. Bauer, P. J. Gmytrasiewicz, and J. Vassileva, Eds. Springer, 137-147.

https://doi.org/10.1007/3-540-44566-8

CORBETT-DAVIES, S., AND GOEL, S. (2018). The measure and mismeasure of fairness: A critical review of fair machine learning. arXiv preprint arXiv:1808.00023. https://doi.org/10.48550/arXiv.1808.00023

CORBETT-DAVIES, S., PIERSON, E., FELLER, A., GOEL, S., AND HUQ, A. (2017, August). Algorithmic decision making and the cost of fairness. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '17)*, S. Matwin, S. Yu, and F. Farooq, Eds. Association for Computing Machinery, 797-806. https://doi.org/10.1145/3097983.3098095

CORRIN, W., SEPANIK, S., ROSEN, R., AND SHANE, A. (2016). Addressing early warning indicators: Interim impact findings from the Investing in Innovation (i3) evaluation of Diplomas Now. New York, NY: *MDRC*.

DALIPI, F., IMRAN, A. S., AND KASTRATI, Z. (2018, April). MOOC dropout prediction using machine learning techniques: Review and research challenges. In *Proceedings of the 2018 IEEE Global Engineering Education Conference (EDUCON)*, IEEE, 1007-1014. https://doi.org/10.1109/EDUCON.2018.8363340

DARENSBOURG, A., PEREZ, E., AND BLAKE, J. J. (2010). Overrepresentation of African American males in exclusionary discipline: The role of school-based mental health professionals in dismantling the school to prison pipeline. *Journal of African American Males in Education (JAAME) 1,* 3, 196-211.

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

DEE, T. S., AND PENNER, E. K. (2021). My brother's keeper? The impact of targeted educational supports. *Journal of Policy Analysis and Management 40,* 4, 1171-1196. https://doi.org/10.1002/pam.22328

DEHO, O. B., JOKSIMOVIC, S., LI, J., ZHAN, C., LIU, J., AND LIU, L. (2022). Should learning analytics models include sensitive attributes? Explaining the why. *IEEE Transactions on Learning Technologies 15,* 6, 1-13. https://doi.org/10.1109/TLT.2022.3226474

DEKKER, G. W., PECHENIZKIY, M., AND VLEESHOUWERS, J. M. (2009). Predicting students drop out: A case study. *Proceedings of the Second International Conference on Educational Data Mining (EDM 2009)*, T. Barnes, M. Desmarais, C. Romero, and S. Ventura, Eds. International Educational Data Mining Society, 41-50.

DESOCIO, J., VANCURA, M., NELSON, L. A., HEWITT, G., KITZMAN, H., AND COLE, R. (2007). Engaging truant adolescents: Results from a multifaceted intervention pilot. *Preventing School Failure: Alternative Education for Children and Youth 51,* 3, 3-9. https://doi.org/10.3200/PSFL.51.3.3-11

DIGITAL PROMISE (2021). *What Do Edtech and AI Have to Do With Racial Bias?* Retrieved June 14, 2022 from

https://digitalpromise.org/2021/10/28/what-do-edtech-and-ai-have-to-do-with-race-bias/

D'MELLO, S., LEHMAN, B., SULLINS, J., DAIGLE, R., COMBS, R., VOGT, K., PERKINS, L., AND GRAESSER, A. (2010, June). A time for emoting: When affect-sensitivity is and isn't effective at promoting deep learning. In *Proceedings of the 10th International Conference on Intelligent Tutoring Systems (ITS 2010)*, V. Aleven, J. Kay, and J. Mostow, Eds. Springer, 245-254. https://doi.org/10.1007/978-3-642-13388-6_29

DOMINA, T., PHARRIS-CIUREJ, N., PENNER, A. M., PENNER, E. K., BRUMMET, Q., PORTER, S. R., AND SANABRIA, T. (2018). Is free and reduced-price lunch a valid measure of educational disadvantage? *Educational Researcher 47,* 9, 539-555.

https://doi.org/10.3102/0013189X18797609

ECKER-LYSTER, M., AND NIILEKSELA, C. (2016). Keeping students on track to graduate: A synthesis of school dropout trends, prevention, and intervention initiatives. *Journal of At-Risk Issues 19,* 2, 24-31

FANCSALI, S., NIXON, T., AND RITTER, S. (2013, July). Optimal and worst-case performance of mastery learning assessment with Bayesian knowledge tracing. In *Proceedings of the International Conference on Educational Data Mining (EDM 2013)*, S. K. D'Mello, R. A. Calvo, and A. Olney, Eds. International Educational Data Mining Society, 35-42.

FEATHERS, T. (2022) College prep software Naviance is selling advertising access to millions of students. *The Markup*, Jan. 13, 2022. Retrieved June 14, 2022 from https://themarkup.org/machine-learning/2022/01/13/college-prep-software-naviance-is-selling-advertising-access-to-millions-of-students

FELDMAN, M., FRIEDLER, S. A., MOELLER, J., SCHEIDEGGER, C., AND S. VENKATASUBRAMANIAN, S. (2015, August). Certifying and removing disparate impact. In *Proceedings of the 21st ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '15)*, L. Cao, and C. Zhang, Eds. Association for Computing Machinery, 259-268. https://doi.org/10.1145/2783258.2783311

FREEMAN, J., SIMONSEN, B., MCCOACH, D. B., SUGAI, G., LOMBARDI, A., AND HORNER, R. (2015). An analysis of the relationship between implementation of school-wide positive behavior interventions and supports and high school dropout rates. *The High School Journal 98,* 4, 290-315. http://dx.doi.org/10.1353/hsj.2015.0009

GARCIA, M. (2016). Racist in the machine. *World Policy Journal 33,* 4, 111-117. http://dx.doi.org/10.1215/07402775-3813015

GARDNER, J., AND BROOKS, C. (2018). Dropout model evaluation in MOOCs. In *Proceedings of the 32nd AAAI Conference on Artificial Intelligence*, AAAI Press, 32(1), 7906-7912. https://doi.org/10.1609/aaai.v32i1.11392

GARDNER, J., BROOKS, C., AND 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 (LAK '19)*, S. Hsiao, J. Cunnigham, K. McCarthy, and G. Lynch, Eds. Association for Computing Machinery, 225-234.

https://doi.org/10.1145/3303772.3303791

GAŠEVIĆ, D., DAWSON, S., ROGERS, T., AND GASEVIC, D. (2016). Learning analytics should not promote one size fits all: The effects of instructional conditions in predicting academic success. *The Internet and Higher Education 28,* 1, 68-84. https://doi.org/10.1016/j.iheduc.2015.10.002

GELMAN, A., FAGAN, J., AND KISS, A. (2007). An analysis of the New York City police department's "stop-and-frisk" policy in the context of claims of racial bias. *Journal of the American Statistical Association 102,* 479, 813-823.

https://doi.org/10.1198/016214506000001040

GERVET, T., KOEDINGER, K., SCHNEIDER, J., AND MITCHELL, T. (2020). When is deep learning the best approach to knowledge tracing?. *Journal of Educational Data Mining 12,* 3, 31-54. https://doi.org/10.5281/zenodo.4143614

GOLDHABER, D., KOEDEL, C., ÖZEK, U., AND PARSONS, E. (2022). Using longitudinal student mobility to identify at-risk students. *AERA Open 8,* 1, 1-13. http://dx.doi.org/10.1177/23328584211071090

GOTTFRIED, M., KIRKSEY, J. J., AND FLETCHER, T. L. (2022). Do high school students with a same-race teacher attend class more often?. *Educational Evaluation and Policy Analysis 44, 1*, 149-169. https://doi.org/10.3102/01623737211032241

HÄFNER, A., OBERST, V., AND STOCK, A. (2014). Avoiding procrastination through time management: An experimental intervention study. *Educational Studies 40,* 3, 352-360. https://doi.org/10.1080/03055698.2014.899487

HALDEMAN, D. C. (2000). Gender atypical youth: Clinical and social issues. *School Psychology Review 29,* 2, 192-200. https://doi.org/10.1080/02796015.2000.12086007

HERODOTOU, C., RIENTIES, B., BOROOWA, A., ZDRAHAL, Z., AND HLOSTA, M. (2019). A large-scale implementation of predictive learning analytics in higher education: the teachers' role and perspective. *Educational Technology Research and Development 67,* 5, 1273-1306. https://doi.org/10.1007/s11423-019-09685-0

HICKS, B., KITTO, K., PAYNE, L., AND BUCKINGHAM SHUM, S. (2022, March). Thinking with causal models: A visual formalism for collaboratively crafting assumptions. In *Proceedings of the 12th International Learning Analytics and Knowledge Conference (LAK '22)*, A. F. Wise, R. Martinez-Maldonado, I. Hilliger, Eds. Association of Computing Machinery, 250-259. https://doi.org/10.1145/3506860.3506899

HOLSTEIN, K., AND DOROUDI, S. (2022). Equity and artificial intelligence in education: Will "AIEd" Amplify or Alleviate Inequities in Education? In *The Ethics of Artificial Intelligence in Education: Practices, Challenges, and Debates*, K. Porayska-Pomsta, and W. Holmes, Eds. Routledge Press. https://doi.org/10.4324/9780429329067

HU, Q., AND RANGWALA, H. (2020). Towards fair educational data mining: A case study on detecting at-risk students. In *Proceedings of the 13th International Conference on Educational Data Mining (EDM 2020),* A. N. Rafferty, J. Whitehill, C. Romero, and V. Cavalli-Sforza, Eds. International Educational Data Mining Society, 431-437.

HUTCHINSON, B., AND MITCHELL, M. (2019). 50 years of test (un) fairness: Lessons for machine learning. In *Proceedings of the Conference on Fairness, Accountability, and Transparency (FAT* '19)*, Association for Computing Machinery, 49-58. https://doi.org/10.1145/3287560.3287600

HUTT, S., GRAFSGAARD, J. F., AND D'MELLO, S. K. (2019, May). Time to scale: Generalizable affect detection for tens of thousands of students across an entire school year. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems (CHI*

'19), S. Brewster, G. Fitzpatrick, A. Cox, and V. Kostakos, Eds. Association for Computing Machinery, 1-14. https://doi.org/10.1145/3290605.3300726

Hyman, S., Aubry, T., and Klodawsky, F. (2011). Resilient educational outcomes: Participation in school by youth with histories of homelessness. *Youth & Society 43,* 1, 253-273. https://doi.org/10.1177/0044118X10365354

Jackson, J. R. (2018). Algorithmic bias. *Journal of Leadership, Accountability and Ethics 15,* 4, 55-65.

Johnson, G. M. (2021). Algorithmic bias: on the implicit biases of social technology. *Synthese 198,* 10, 9941-9961. https://doi.org/10.1007/s11229-020-02696-y

Karumbaiah, S. (2021) *The Upstream Sources of Bias: Investigating Theory, Design, and Methods Shaping Adaptive Learning Systems*. Doctoral dissertation, University of Pennsylvania.

Kai, S., Andres, J. M. L., Paquette, L., Baker, R. S., Molnar, K., Watkins, H., and Moore, M. (2017) Predicting student retention from behavior in an online orientation course. In *Proceedings of the 10th International Conference on Educational Data Mining (EDM 2017)*, X. Hu, T. Barnes, A. Hershkovitz, and L. Paquette, Eds. International Educational Data Mining Society, 250-255.

Kizilcec, R. F., and Lee, H. (2022). Algorithmic fairness in education. In W. Holmes & K. Porayska-Pomsta (Eds.), *The Ethics of Artificial Intelligence in Education: Practices, Challenges, and Debates.* Routledge Press. https://doi.org/10.4324/9780429329067

Klare, B. F., Burge, M. J., Klontz, J. C., Bruegge, R. W. V., and Jain, A. K. (2012). Face recognition performance: Role of demographic information. *IEEE Transactions on Information Forensics and Security 7*, 6, 1789-1801. https://doi.org/10.1109/TIFS.2012.2214212

Kleinberg, J., Mullainathan, S., and Raghavan, M. (2017). Inherent trade-offs in the fair determination of risk scores. In *Proceedings of the 8th Innovations in Theoretical Computer Science Conference (ITCS 2017)*, C. H. Papadimitriou, Ed. Schloss Dagstuhl–Leibniz-Zentrum fuer Informatik, 67, 43:1–43:23. https://doi.org/10.4320/LIPIcs.ITCS.2017.43

Kloft, M., Stiehler, F., Zheng, Z., and Pinkwart, N. (2014, October). Predicting MOOC dropout over weeks using machine learning methods. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP 2014),* A. Moschitti, B. Pang, and W. Daelmanns, Eds. Association for Computational Linguistics, 60-65. http://dx.doi.org/10.3115/v1/W14-4111

Koedinger, K. R., Stamper, J. C., McLaughlin, E. A., and Nixon, T. (2013, July). Using data-driven discovery of better student models to improve student learning. In *Proceedings of the 16th International Conference on Artificial Intelligence in Education (AIED 2013)*, H. C. Lane, K. Yacef, J. Mostow, and P. Pavlik, Eds. Springer, 421-430. https://doi.org/10.1007/978-3-642-39112-5_43

Kusner, M. J., Loftus, J., Russell, C., and Silva, R. (2017). Counterfactual fairness. In *Advances in Neural Information Processing Systems 30 (NIPS 2017)*, I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, Eds. Curran Associates, Inc.

LEE, H., AND KIZILCEC, R. F. (2020). Evaluation of fairness trade-offs in predicting student success. ArXiv E-Prints, arXiv:2007.00088. https://arxiv.org/abs/2007.00088. Accessed 1 Oct 2021. https://doi.org/10.48550/arXiv.2007.00088

LEE, S. M. (1993). Racial classifications in the US Census: 1890–1990. *Ethnic and racial studies*, *16*(1), 75-94. https://doi.org/10.1080/01419870.1993.9993773

LEE, M., AND SCHUELE, C. M. (2010). Demographics. *Encyclopedia of research design*, 347-348.

LEE, S. J. (2012). New talk about ELL students. *Phi Delta Kappan 93,* 8, 66-69. https://doi.org/10.1177/003172171209300816

LI, Y., ZOU, X., MA, Z., AND BAKER, R. S. (2022) A multi-pronged redesign to reduce gaming the system. In *Proceedings of the 23rd International Conference on Artificial Intelligence in Education (AIED 2022)*, M. M. Rodrigo, N. Matsuda, A. I. Cristea, and V. Dimitrova, Eds. Springer, 334-337. https://doi.org/10.1007/978-3-031-11647-6_64

LOUKINA, A., MADNANI, N., AND ZECHNER, K. (2019). The many dimensions of algorithmic fairness in educational applications. In *Proceedings of the Fourteenth Workshop on Innovative Use of NLP for Building Educational Applications*, H. Yannakoudakis, E. Kochmar, C. Leacock, N. Madnani, I. Pilan, and T. Zesch, Eds. Association for Computational Linguistics, 1–10. http://dx.doi.org/10.18653/v1/W19-4401

LOWRY, S., AND MACPHERSON, G. (1988). A blot on the profession. *British medical journal (Clinical research ed.) 296,* 6623, 657. https://doi.org/10.1136%2Fbmj.296.6623.657

MATLOFF, N., AND ZHANG, W. (2022). A novel regularization approach to fair ML. *arXiv preprint arXiv:2208.06557*. https://doi.org/10.48550/arXiv.2208.06557

MILLIRON, M. D., MALCOLM, L., AND KIL, D. (2014). Insight and action analytics: Three case studies to consider. *Research & Practice in Assessment 9*, 70-89.

MORRISON, T. G., AND KISS, M. (2017). Modern racism scale. *Encyclopedia of personality and individual differences*, 1-3. https://doi.org/10.1007/978-3-319-28099-8_1251-1

NAGRECHA, S., DILLON, J. Z., AND CHAWLA, N. V. (2017, April). MOOC dropout prediction: lessons learned from making pipelines interpretable. In *Proceedings of the 26th International Conference on World Wide Web Companion (WWW '17 Companion),* R. Barrett, R. Cummings, E. Agichtein, and E. Gabrilovich, Eds. International World Wide Web Steering Committee, 351-359. https://doi.org/10.1145/3041021.3054162

NATIONAL ACADEMIES OF SCIENCES, ENGINEERING, AND MEDICINE. 2022. *Measuring Sex, Gender Identity, and Sexual Orientation.* Washington, DC: The National Academies Press. https://doi.org/10.17226/26424

O'REILLY-SHAH, V. N., GENTRY, K. R., VAN CLEVE, W., KENDALE, S. M., JABALEY, C. S., AND LONG, D. R. (2020). The COVID-19 pandemic highlights shortcomings in US health care informatics infrastructure: A call to action. *Anesthesia and analgesia 131*, 2, 340-344. http://dx.doi.org/10.1213/ANE.0000000000004945

PANTIC, K., AND CLARKE-MIDURA, J. (2019). Factors that influence retention of women in the computer science major: A systematic literature review. *Journal of Women and Minorities in Science and Engineering 25,* 2, 119-145.

http://dx.doi.org/10.1615/JWomenMinorScienEng.2019024384

PAQUETTE, L., OCUMPAUGH, J., LI, Z., ANDRES, A., AND BAKER, R. (2020). Who's learning? Using demographics in EDM research. *Journal of Educational Data Mining 12,* 3, 1–30. https://doi.org/10.5281/zenodo.4143612

PEARL, J., AND MACKENZIE, D. (2018). *The Book of Why: The New Science of Cause and Effect*. Basic Books.

ROBISON, S., JAGGERS, J., RHODES, J., BLACKMON, B. J., AND CHURCH, W. (2017). Correlates of educational success: Predictors of school dropout and graduation for urban students in the Deep South. *Children and Youth Services Review 73,* 1, 37-46. https://doi.org/10.1016/j.childyouth.2016.11.031

ROSENTHAL, R., AND JACOBSON, L. (1968). Pygmalion in the classroom. *The Urban Review*, *3*(1), 16-20. http://dx.doi.org/10.1007/BF02322211

SCHWARTZ, S. E., KANCHEWA, S. S., RHODES, J. E., GOWDY, G., STARK, A. M., HORN, J. P., PARNES, M., AND SPENCER, R. (2018). “I'm having a little struggle with this, can you help me out?”: Examining impacts and processes of a social capital intervention for first-generation college students. *American Journal of Community Psychology 61,* 1-2, 166-178. https://doi.org/10.1002/ajcp.12206

SCHWÖBEL, P., AND REMMERS, P. (2022, June). The long arc of fairness: Formalisations and ethical discourse. In *Proceedings of the 5th ACM Conference on Fairness, Accountability, and Transparency (FAccT '22)*, Association for Computing Machinery, 2179-2188. https://doi.org/10.1145/3531146.3534635

SHOLLENBERGER, T. L. (2015). Racial disparities in school suspension and subsequent outcomes. In D. Losen (Ed.) *Closing the School Discipline Gap: Equitable Remedies for Excessive Exclusion*, Teachers College Press, 31-43.

SIEGEL, E. (2013). *Predictive Analytics: The Power to Predict Who Will Click, Buy, Lie, or Die*. John Wiley & Sons.

SNIPP, C. M. (2003). Racial measurement in the American census: Past practices and implications for the future. *Annual Review of Sociology 29,* 1, 563-588. https://doi.org/10.1146/annurev.soc.29.010202.100006

SURESH, H., AND GUTTAG, J. V. (2019). A framework for understanding unintended consequences of machine learning. arXiv preprint arXiv:1901.10002, 2, 8.

VASQUEZ VERDUGO, J., GITIAUX, X., ORTEGA, C., AND RANGWALA, H. (2022, March). FairEd: A systematic fairness analysis approach applied in a higher educational context. In *Proceedings of the 12th International Learning Analytics and Knowledge Conference (LAK '22)*, A. F. Wise, R. Martinez-Maldonado, I. Hilliger, Eds. Association of Computing Machinery, 271-281. https://doi.org/10.1145/3506860.3506902

VERBERT, K., DUVAL, E., KLERKX, J., GOVAERTS, S., AND SANTOS, J. L. (2013). Learning analytics dashboard applications. *American Behavioral Scientist 57,* 10, 1500-1509. https://doi.org/10.1177/0002764213479363

WEIDLICH, J., GAŠEVIĆ, D., AND DRACHSLER, H. (2022). Causal inference and bias in learning analytics: A primer on pitfalls using directed acyclic graphs. *Journal of Learning Analytics 9,* 3, 183-199. https://doi.org/10.18608/jla.2022.7577

WOLFF, A., ZDRAHAL, Z., NIKOLOV, A., AND 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 (LAK '13)*, D. Suthers, K. Verbert, E. Duval, and X. Ochoa, Eds. Association for Computing Machinery, 145-149. https://doi.org/10.1145/2460296.2460324

WONG, P. H. (2020). Democratizing algorithmic fairness. *Philosophy & Technology 33*, 225-244. https://doi.org/10.1007/s13347-019-00355-w

YON, M. (1995). Educating homeless children in the United States. *Equity and Excellence in Education 28,* 1, 58-62. https://doi.org/10.1080/1066568950280110

YU, R., LEE, H., AND KIZILCEC, R. F. (2021). Should college dropout prediction models include protected attributes?. In *Proceedings of the Eighth ACM Conference on Learning @ Scale (L@S '21)*, C. Meinel, M. Perez-Sanagustin, M. Specht, and A. Ogan, Eds. Association for Computing Machinery, 91–100. https://doi.org/10.1145/3430895.3460139

YU, R., LI, Q., FISCHER, C., DOROUDI, S., AND XU, D. (2020). Towards accurate and fair prediction of college success: evaluating different sources of student data. In *Proceedings of the 13th International Conference on Educational Data Mining (EDM 2020),* A. N. Rafferty, J. Whitehill, C. Romero, and V. Cavalli-Sforza, Eds. International Educational Data Mining Society*,* 292-301.

ZAFAR, M. B., VALERA, I., GOMEZ-RODRIGUEZ, M., AND GUMMADI, K. P. (2019). Fairness constraints: A flexible approach for fair classification. *The Journal of Machine Learning Research 20,* 1, 2737-2778.