Who's learning? Using demographics in EDM research

Luc Paquette
University of Illinois at Urbana-Champaign
lpaq@illinois.edu

Jaclyn Ocumpaugh
University of Pennsylvania
ojaclyn@upenn.edu

Ziyue Li
University of Illinois at Urbana-Champaign
ziyueli3@illinois.edu

Alexandra Andres
University of Pennsylvania
alexandraandres@gmail.com

Ryan Baker
University of Pennsylvania
rybaker@upenn.edu

The growing use of machine learning for the data-driven study of social issues and the implementation of data-driven decision processes has required researchers to re-examine the often implicit assumption that data-driven models are neutral and free of biases. The careful examination of machine-learned models has identified examples of how existing biases can inadvertently be perpetuated in fields such as criminal justice, where failing to account for racial prejudices in the prediction of recidivism can perpetuate or exasperate them, and natural language processing, where algorithms trained on human languages corpora have been shown to reproduce strong biases in gendered descriptions. These examples highlight the importance of thinking about how biases might impact the study of educational data and how data-driven models used in educational contexts may perpetuate inequalities. To understand this question, we ask whether and how demographic information, including age, educational level, gender, race/ethnicity, socioeconomic status (SES), and geographical location, is used in Educational Data Mining (EDM) research. Specifically, we conduct a systematic survey of the last five years of EDM publications that investigates whether and how demographic information about the students is reported in EDM research and how this information is used to 1) *investigate* issues related to demographics, 2) use the information as *input features for data-driven analyses*, or 3) to *test and validate models*. This survey shows that, although a majority of publications reported at least one category of demographic information, the frequency of reporting for different categories of demographic information is very uneven (ranging from 5% to 59%), and only 15% of publications used demographic information in their analyses.

Keywords: machine learning bias, equity, fairness, meta-analysis

1. INTRODUCTION

Machine learning is an increasingly popular method for analyzing important social questions (Barocas & Selbst, 2016), as it is more capable than traditional approaches of analyzing very

large data sets. For example, it has allowed researchers to study legislative and judicial behaviors of the U.S. Supreme Court (Lauderdale & Clark, 2014) and to more accurately identify potential recipients of food security and social safety net interventions (McBride & Nichols, 2016). In education research, this trend is reflected in the emergence of new academic societies, such as the *International Educational Data Mining Society* (*IEDMS*)[1] and the *Society for Learning Analytics Research* (*SoLAR*)[2], dedicated to the study of how advances in machine learning can be leveraged to support education (e.g., Ferguson, 2012; Siemens & Baker 2012; Verbert et al., 2012).

A powerful characteristic of machine learning is its ability to extract implicit information from data in order to reveal trends and associations that might not otherwise have been discovered through human inspection, enabling accurate prediction of real-world outcomes and labels. However, growing concerns have emerged about the fairness of models created through the usage of machine learning that is otherwise technically accurate (Feldman et al., 2015; Hajian & Domingo-Ferrer, 2013; Holstein et al., 2019; Romei & Ruggieri, 2014; Stoyanovich et al., 2018; Zafar et al., 2015, 2017; Žliobaite, 2017). As machine learning is designed to discover implicit trends in data, it sometimes models social biases that are implicitly encoded in the data it analyzes.

This question of fairness is especially relevant to social applications of machine learning, where uneven degrees of accuracy or systematically biased predictions could have important consequences. For example, "redlining" practices (where African Americans were not allowed to purchase real estate in certain neighborhoods) mean that zip-codes must be used cautiously, lest the models using it as an input variable inadvertently recreate racially discriminatory patterns. These concerns have already drawn legal attention in both real estate and banking/lending practices (cf. Barocas & Selbst, 2016), and would likely also create issues for any education researchers who were not vigilant in the use of this data as well.

As issues of social inequity are frequently found in educational contexts, the risk of producing an algorithm that replicates these social ills is high, particularly if researchers are not paying careful attention to the types of variables being used in their research and how they relate to these social trends. Sampling biases may also create issues that are not related to historical discriminatory processes or protected classes. For example, when creating automated detectors for the recognition of student emotions, Ocumpaugh et al. (2014) found that differences in student population (students from multiple urban, suburban and rural schools) impacted the detection of affect. Detectors trained on one population of students did not generalize well to other populations and could have substantially reduced the accuracy of predictions for some groups of students (in this case, those from rural schools).

Ensuring algorithmic fairness in education is a herculean task, but an important place to start is to look at the degree to which EDM researchers, as a community, are documenting and studying the kinds of information necessary to identify and address such social biases. For these reasons, we investigate how Educational Data Mining (EDM), as a community, approaches issues of equity and fairness in the data-driven study of education that examines the use of demographic information in EDM studies. Specifically, we survey the last five years of publications in the EDM community, coding for the categories of demographic information[3],

[1] http://www.educationaldatamining.org/
[2] https://solaresearch.org/
[3] The categories were chosen to reflect common categories included by the US Department of Education's student demographic statistics (see https://www2.ed.gov/rschstat/catalog/student-demographics.html)

including age, education level, gender, race/ethnicity, socioeconomic status (SES), and geographical location, that were reported (if any) in the publication and how those demographics are used (i.e., to *investigate* potential population differences, to *generate features for data-driven analyses,* or to *test and validate* models). The results show that, although around 72% of publications reported at least one demographic category, there is a wide range in the frequency of reporting for different categories – from 5% to 59%. Similarly, whereas 72% of the paper reported information on at least one demographic category, only 15% applied this information in any way during their analyses.

2. BIASES IN MACHINE LEARNING APPLICATIONS

The growing implementation of algorithmic-based decision processes was, in part, ushered in under assumptions that algorithms are, by default, impartial. This assumption, termed the *neutrality fallacy* (e.g., Sandvig, 2015), has been questioned by those who are concerned that these untested assumptions leave us vulnerable to inadvertently developing (prejudicially) *discriminatory classification rules* (e.g., Perdeschi et al., 2008), which are increasingly becoming a target of legal consequences (e.g., Gellert et al., 2013; Makkonen, 2007). As Veale and Binns (2017) point out, "'neutral' choices in machine learning systems do not exist – candidates for these, such as software defaults, are best thought of as arbitrary" (p. 3). Some have gone as far as to say models that do not explicitly test for the effects against protected classes "cannot be discrimination aware" because they cannot actively guard against prejudicial discrimination practices being replicated in the model (Žliobaite, 2017), or as Galhotra and colleagues (2017) summarize, "Knowing if there is discrimination can lead to better-informed decision making" (p. 499). In this section, we discuss some of the issues and findings surrounding algorithmic fairness.

2.1. ALGORITHMIC BIASES ACROSS DISCIPLINES

Researchers from a wide range of disciplines, from criminal justice and other areas of law (Simoiu et al., 2017), to natural language processing (Caliskan et al., 2017; Blodgett et al., 2016; Jurgens et al., 2017; Hovy, 2015; Kiritchenko & Mohammad, 2018; Tatman, 2017; Wiltz 2017), to facial recognition and other research related to vision (Alvi et al., 2018; Brosch et al., 2013; Klare et al., 2012; Hendrix et al., 2018; Misra et al., 2016; van Miltenburg, 2016; Nieva, 2015), to emotion recognition (Howard et al., 2017 Shaffer, 2018), to hiring practices (Chen et al., 2018), to medicine (Ashen et al., 2017; Gianfrancesco et al., 2018; Goodman et al., 2018; Maddox et al., 2018; Rajkomar et al., 2018; Verghese et al., 2018), to the testing of software for consumer services and advertising (Galhotra et al., 2017; Sweeney, 2013), have raised concerns about the effects of subpopulations—whether or not they are part of a protected class—and reminded their colleagues that algorithms are only as good as the data they contain since algorithms' biases are likely to reflect their training data (Jaton, 2017). While opaque black-box algorithms that maximize goodness of fit have become increasingly popular, many have cautioned that the opacity of some algorithms makes them more susceptible to misinterpretation when their output is presented to human decision makers (e.g., Ziewitz, 2016).

In some cases, historically problematic practices may influence the data in ways that are understood generally, but not addressed by machine learning. As d'Alessandro and colleagues (2017) discuss, an algorithm that finds that men are more likely to achieve a longer tenure in employment may be detecting a poor working environment for women, which could inadvertently be reinforced if it were used as part of the hiring criteria. This kind of bias in an

algorithm is also seen in algorithms that make use of U.S. zip codes, where there is a history of "redlining" African Americans into and out of certain neighborhoods and subsequently not providing people with loans based on the perceived risk of those neighborhoods (see discussion in Barocas & Selbst, 2016). It is also seen in criminal justice data (e.g., Chouldechova, 2017).

Cases like these are particularly problematic when the results of machine learning algorithms are used to train future algorithms, primarily because they are driven by *precision feedback* (the ratio of good to bad candidates who passed through the initial algorithm) with no feedback on the initial algorithm's recall (how many good candidates were excluded). As d'Allesandro and colleagues (2017) explain, "Once the model has learned to shut out a particular group, it has little opportunity to unlearn the said trend" (p. 132), a process that continues to perpetuate inequality, whether or not the variable triggering the algorithmic decision is an actual demographic category (e.g., sex or race) or just a proxy variable associated with that group.

Moreover, just because a system is treating two groups as the same does not mean that the treatment results in identical outcomes. As Mitchell et al. (2018) point out, the consequences for a low-income applicant who is denied a loan may be far more severe than that of a high-income applicant. Given these concerns, it is important to recognize that a high performing algorithm (by some statistical performance metrics) might still fail at value-sensitive design criteria, which evaluate algorithms by their ability to reflect moral or legal values like fairness of opportunity (e.g., Friedman, 1996).

Increasingly, researchers are working towards developing techniques for non-discriminatory algorithms (e.g., Calders & Verwer, 2010; Kamisha et al., 2012; Kamiran & Calders, 2012; Ruggieri et al., 2010a; Yao et al., 2017), including explorations of bias within historical data (e.g., Ruggieri et al., 2010b) and metrics that explicitly investigate differences in predictive goodness between groups (e.g., Gardner et al., 2019). Some suggest that the only way to ensure equity is to collect the sensitive information associated with discriminatory practices so that discriminatory biases in the algorithms can be explicitly evaluated (e.g., Žliobaite & Custers, 2016), while others have worked to develop new methods for statistically inferring such potentially sensitive information when it is either unavailable or undesirable to collect (e.g., Chakraborthy et al. 2017). Holstein and colleagues (2019) suggest a need for fairness-aware data collection to "support practitioners in collecting and curating high-quality datasets in the first place, with an eye towards fairness in downstream ML models" (p. 12) Such work could help to solve the *value alignment problem* suggested by Hadfield-Menell and colleagues (2016), that seeks to generate performance criteria which align with the cultural values of its users (see also Bostrom, 2016), in some cases to meet legal criteria (Goodman & Flaxman, 2017; Zliobaite, 2017).

Ultimately, therefore, the fairness may result more from the application of the algorithm (and the subsequent consequences to the groups it categorizes) than how it categorizes a particular group. We have seen this problem in medical research, where biological differences mean that men and women often need different doses of medicine or different types of treatment altogether, but when working with behavioral data, there are reasons to believe that both biological and non-biological differences may be important to address. For example, Gosse and Arnocky (2012) describe a concern where all children are treated the same, but there appears to be a disparate effect on some sub-population(s). Specifically, they suggest that the repeated reduction of playground time and other opportunities for physical education may be exasperating symptoms of autism and attention deficit hyperactivity disorder, requiring more medication than would be otherwise necessary to treat these conditions. Moreover, they worry that this may be responsible for the over-diagnoses of young boys with these symptoms

(although readers are encouraged to explore the literature on the potential under diagnoses of young girls with these conditions, e.g., Bierdman et al., 2002; Gaub & Carlson, 1997).

2.2. Examples from Natural Language Processing

As some have pointed out, hidden biases are relatively common in data (Rosenbaum, 2001; Hacker and Wiedemann, 2017), and replicating these biases may be useful for some applications, but dangerous in others (i.e., the value alignment problem). One area where biases have become of particular concern is in the field of Natural Language Processing (NLP), where algorithms trained on large corpora of human languages have been shown to (re)produce strong biases in how men and women are described (e.g., Bolukbasi et al., 2016). These biases could have dire consequences when applied to hiring practices, but as Zadronzy and colleagues (2000) point out, gendered biases can be quite beneficial to the development of a dialogue system for certain conversational topics (e.g., clothing).

Likewise, if those in education wanted to use NLP tools to evaluate letters of recommendation for students, they should seek to use tools that eliminate gendered biases. On the other hand, if designers of a learning system were trying to recommend examples to increase student engagement and reinforce learning, the use of an algorithm that reflected gendered "roles, norms, and expectations" (e.g., Eckert's 1989 discussion of gendered language variation), could help to provide more appropriate recommendations for conversational strategies. These would need to be used cautiously, however, since the kinds of gender preferences discussed in some research (e.g., Pinkard 2005; Kinzie & Joseph, 2008) are not universally generalizable. If the cultural climate reflected in the training data is not valid for the students who are being educated with that algorithm, the use would obviously not align with the values.

2.3. Bias and education data

In order to reach goals of developing truly personalized education, traditional demographic information may not result in the best algorithmic recommendations, as demographics are often a proxy for a number of biological, cultural, and other environmental conditions. In addition to demographic segmentation, behavioral and psychometric segmentation (e.g., Burr et al., 2018) might be more appropriate. However, it seems particularly important that researchers recognize the different ways in which demographic variables may influence this behavior, while also taking into consideration the different kinds of statistical underrepresentation that may occur (e.g., Yao & Huang, 2017).

There is, after all, no shortage of literature on demographic differences in education (e.g., Gutiérrez & Rogoff, 2003; Nasir & Hand, 2006), and there are at least a half a dozen major theories on education that explicitly reference sociocultural contexts (Ladson-Billings, 1998; Ladson-Billings, 1995; Bakhtin, 2010; Paris & Byrnes, 1989; Bandura, 2001; Roth & Lee, 2007; Lave, 1991; Siemens, 2005; Wadsworth, 1996; Ültanir, 2012). In addition to the language-based practices that may cause the kind of biases we see in NLP algorithms, there are other important cultural practices that may impact learning in unexpected ways. For instance, Karumbaiah et al.'s (2019) findings on the relationship between school-level demographics and help-seeking practices suggest that more subtle socio-linguistic differences may influence students' interactions with online learning systems. This finding may help to explain the myriad of conflicting evidence the EDM community has seen in the efficacy of help systems in online learning (e.g., Koedinger & Aleven, 2007; Karumbaiah et al., 2019). Similarly, Agranovich et

al.'s (2011) analyses show that cultural differences inform American and Russian students' reactions to timed tests, a finding that could also inform the design of online learning systems. If the EDM community is not even reporting demographics, it can be difficult to determine which results will generalize to new populations, let alone to ensure that we are not exasperating the opportunity gaps (Childs, 2017; Milner IV, 2012; Welner & Carter, 2013) that have historically prevented some demographics from achieving their full potential.

3. METHODS

In order to determine the extent to which the current practices in the EDM community are adequately addressing the potential effects of demographic differences, we conducted a survey of every paper published at an official EDM venue in the last five years (2015 to 2019). As described below, these papers were first coded to determine which demographic categories (if any) are reported, and then if/how those categories were used in the analyses.

3.1. PUBLICATIONS SURVEYED

Specifically, we surveyed the 385 papers published in the Journal of Educational Data Mining and the International Conference on Educational Data Mining. Publications in related venues (e.g., the Journal of Learning Analytics) were not surveyed, as the goal was to survey current practices in the EDM community. Despite the considerable thematic overlap between EDM and learning analytics, both the community of researchers (Labarthe et al., 2018) and their practices (Baker & Siemens, 2014) are different. Since the goal of the survey is to identify current trends in EDM research, the survey was limited to the last five years of EDM publications (2015 to 2019 inclusively), as earlier papers may not be representative of the rapidly changing practices of this community.

For publications at the EDM conference, both full papers and short papers were surveyed. Poster papers, however, were excluded, as their limited format often lends itself toward early work in progress rather than more detailed analyses. Furthermore, in practice, they often include papers that were cut for length.

3.2. CODING CATEGORIES

Several characteristics were considered when coding these papers. As outlined in Table 1, this included whether or not demographic information was potentially relevant to the analyses. Among those papers where coders did deem demographic information to be *applicable*, papers were further categorized by whether or not demographic information was (1) *analyzed*, (2) merely *reported*, or (3) *not reported*. Among those papers that analyzed demographic information, we subdivided them by *how* the demographic information was used, namely, whether it was (i) *investigated*, (ii) included as *features in models*, or (iii) used to *test and validate* a machine-learned model. Finally, for any paper where demographic information was analyzed or reported, we recorded which demographic information was considered. This nested coding scheme is described in greater detail below.

Table 1: Nested coding scheme, based on how demographic information is used, if at all. Note that a publication could potentially use the information in more than one type of analysis (e.g., both *investigated* and *features in models*). Categories where the type of demographic information was also coded are marked with an asterisk.

Code	Description
I. Applicable	Demographic data was deemed potentially relevant to the analyses
1. Analyzed*	Paper makes use of at least one of the following categories in its analysis: *age, educational level, gender, race, SES, or geographic location*
i. Investigated*	Analysis specifically investigates the relationship between demographic information and construct being examined
ii. Features in Models*	Demographic information is used as input features during model training
iii. Testing & Validation*	The model is validated across different sub-groups of participants, based on their demographic information
2. Reported*	Demographic information of any kind is reported *only* (it is not used in any analyses)
3. Not Reported*	No demographic information is reported (and, as such, it is not used in the analyses in any way)
II. Not Applicable	Demographic data was deemed NOT relevant to the analyses

3.2.1. Demographic categories considered

Six demographic categories were considered in this study: age, educational level, gender, race/ethnicity (referred to as "race" in the paper for conciseness), socioeconomic status (SES) – sometimes reported using the proxy variable of free and reduced lunches – and geographic location – including the participant's location in the world (country, region of a country, etc.) or the urbanicity of the location. These were chosen to reflect common categories included by the U.S. Department of Education's student demographics statistics[4]. The U.S. Department of Education usually organizes statistics by state (geographic location) and grade (education-level). They collect information about age, sex/gender, race/ethnicity, English proficiency, disability, and income to relate them to various educational variables such as assessment results, school suspensions, bullying, etc.

Out of the categories commonly reported, data related to English proficiency (or proficiency in a primary language for studies conducted in countries where English is not the primary language) and data about student disabilities were not coded in our survey. This decision was made because very few of the surveyed papers reported on these categories. Both of those categories were almost exclusively reported when the study presented in the paper was on a topic specifically related to the category (see discussion at the end of section 5.1 for examples).

3.2.2. Applicability of demographic categories

As EDM research sometimes addresses methodological issues or involves data that would not contain demographic variation (e.g., the analysis of instructional content), we first categorized publications in regard to whether the reporting of demographic information was applicable or not. For example, papers comparing the properties of different knowledge tracing algorithms (Khajah et al., 2016; Doroudi & Brunskill, 2017), using document segmentation to automatically label a document with learning objectives (Bhartiya et al., 2016) or automatically identifying learning paths through web learning resources (Labutov & Lipson. 2016) would all be considered as paper for which the study of demographic information is *not applicable*. On the other hand, most any paper that uses student data to develop a model of learning outcomes

[4] https://www2.ed.gov/rschstat/catalog/student-demographics.html

(or associated constructs) could potentially be using data that shows demographic-based differences. These papers were therefore coded as *applicable.*

3.2.3. Subdivisions of papers where demographics were applicable

Among papers where demographics were found to be *applicable,* coders determined whether the demographics were (1) *analyzed.* If the demographics were not part of the analysis, but the publication provided an overview of the distribution of the participants across any demographics (or it provided more detailed data about this category for each participant), the publication was coded as (2) *reported.* Publications that did not provide any demographic information about their research subjects, even though it may have been relevant to the analyses, were coded as (3) *not reported*.

3.2.4. Subdivisions of papers where demographics were analyzed

Finally, among papers where the data was *analyzed,* we coded for *how* that information was incorporated into the analysis. Specifically, we coded whether the paper (i) *investigated* how demographic categories interacted with the outcome variable being examined, (ii) used the demographic variables as a *feature in the model* development process, and/or (iii) *tested and validated* the model to see if it generalizes across these demographics. Papers could potentially be coded in multiple categories of use.

Papers coded as *investigated* use EDM approaches to investigate and answer at least one question related to the experiences of different student demographics. For example, issues of cultural differences can be investigated through the use of information related to the geographic location of participants. This approach was used by Liu and colleagues (2016), who used hierarchical clustering to construct online course activity profiles and then studied whether students in different countries showed different profiles. Other examples are seen in Saarela and Kärkkäinen (2015), who used clustering to investigate the existence of country stereotypes in PISA, and in Feng and colleagues (2016), who used hierarchical linear regression models (HLM) to investigate gendered differences in how students benefited from online homework intervention. Investigation studies are important, as they attempt to measure potential biases, improving our understanding of how demographic factors may interact with the constructs being studied. Their results provide researchers with valuable information about the factors that need to be accounted for in order to ensure minimal biases.

Papers in the *features in models* category used demographic variables as input features to improve the quality of the models that were created as part of a study. This could be achieved, for example, by including information about gender and race when training a predictive model. Papers were coded as *features in models* if demographic variables were reported to have been used as input variables, regardless of whether or not they were selected as features in the final models. Within the EDM community, this approach has been used to identify students who are most likely to enroll at a university (Slim et al., 2018) and students who are at risk of dropping out of high school (Knowles, 2015). One of the main benefits of using this approach is that including demographic variables as inputs can lead to increased performance for a model (see Wolff et al.'s (2013) comparison of models with and without demographic variables). In particular, incorporating demographic variables in this fashion can make it possible for a model to differentiate between populations by applying different criteria to students from different groups when generating predictions. However, simply using demographic information as an input feature when developing a model does not validate that the model will be equitable for every population. In addition, doing so creates some risk. For one thing, if there are base rate

differences between groups due to societal biases, incorporating demographics as a variable may simply replicate these biases (saying, for instance, that a student is more at risk because of their group). This use of demographic variables may actually hide problematic forms of bias. For example, 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. Treating two students as different solely because of their race may also risk violating regulations around equal treatment for members of different groups. As such, even when including demographic variables as features in models, it might be desirable to also validate that inclusion of those variables results in a more equitable performance for the models across diverse populations and to investigate *why* including these features improves performance.

Papers using demographics for *testing and validation* do not necessarily investigate specific questions related to fairness and equity, nor do they necessarily directly include demographic variables in their analyses. Rather, the values of the demographic variables are used to define different testing sets that can be used to validate that a created model performs similarly across multiple populations of participants. For example, Kai et al. (2017) created models to predict whether students were likely to continue their participation in an online degree. While the models were trained using only behavioral variables, they were repeatedly applied to diverse testing sets to evaluate how well the models performed for students of different genders and different races. Samei et al. (2015) used a similar approach to evaluate how well models developed for the classification of classroom discourse generalizes across datasets collected from different schools in different geographical settings (urban vs. non-urban school). The result of using demographic information for testing and validation is not an increase in our understanding of possible biases and other issues of fairness and equity nor an increase in the performance of models created using EDM methods. Rather, it is a way for researchers to identify potential limitations of their models and to facilitate fair and equitable use of EDM models, based on a better understanding of the context in which a model will or will not perform as expected. While we may not explicitly learn of bias from this approach, at minimum we avoid replicating bias or even magnifying it.

3.3. Coding Process

The publications in this survey were coded by the third and fourth authors of this paper. First, the third author surveyed the 2016 and 2017 conference proceedings in order to validate our survey categories. Next, a subset of these publications was selected by the first author to include examples from each of the categories of demographic information use in analyses (Table 1), and these were re-coded by the fourth author in order to check their reliability. As few papers made use of demographic information beyond reporting (only 8 across both years), the number of papers selected for this subset was limited to 14. In order to prevent biases, the fourth author was not aware of this selection process. Both authors had near-perfect agreement, with only one disagreement across all categories (regarding whether demographic information was applicable). This disagreement was resolved through discussion, and it was decided that the definition used by the first coder (third author) was preferable. Once this disagreement was resolved, the fourth author completed the survey for the 2015-2018 publications of the *International Conference on Educational Data Mining* and of the *Journal of Educational Data Mining*.

Following this coding session, a second set of papers, coded by the fourth author, was selected to formally evaluate inter-rater agreement. Selection of the paper was semi-random and

designed to ensure that multiple examples of rare codes were included in the selected set. The two coders were not informed about how papers were selected to avoid coding bias. The set was composed of all the papers coded by the fourth author, which included any one of the *investigated*, *features in models*, and *testing & validation* codes. For papers coded as *reported*, *not reported*, or *not applicable*, a random sample of 15% of the papers in each category was selected. In total, this second set included 61 papers.

Inter-rater agreement was computed across all papers that were coded by both coders (75 papers in total). Agreement across all the six demographic data categories and the six codes related to the use of demographic data in the papers' analyses was evaluated using Cohen's Kappa and ranged from 0.721 to 0.946 (see Table 2).

Table 2: Inter-rater agreement for each of the 12 codes included in the coding scheme.

Code	Kappa
Age	0.902
Education Level	0.916
Gender	0.946
Race	0.744
Socio-economic status	0.801
Geographical location	0.765
Investigated	0.721
Features in model	0.888
Testing & validation	0.850
Reported Only	0.803
Not Reported	0.868
Not applicable	0.825

Finally, the fourth author coded publications from the year 2019, including both conference proceedings and journal articles.

4. RESULTS

Results demonstrate 1) the percentage of publications for which demographic information was applicable, 2) how many publications reported demographic information from at least one category, 3) which categories of demographic information are most commonly reported, 4) what percentage of publications made use of the demographic information, and 5) what type of use was most frequent. Each coding category was aggregated by year for each venue (conference vs. journal) and by publication format (journal article vs. full and short conference papers).

4.1. APPLICABLE DEMOGRAPHIC INFORMATION

First, we examine how frequently demographic information could potentially have influenced the data being examined in EDM publications. The reader may recall that this coding category, *applicable*, differentiates analyses of data that do not contain demographic variation (e.g., analyses of instructional content that does not make use of student data) from investigations

where demographic differences could potentially influence the outcomes of the research (e.g., analyses of student behaviors).

As Table 3 shows, nearly 75% of EDM publications in this study contained analyses where demographic information was *applicable*. This finding holds across all five years of publications that we coded. While journal articles show more variation than conference proceedings, the majority of EDM publications use data that could potentially have been influenced by demographic variation.

Table 3: Total number of publications in the survey and number of publications where demographic information was considered applicable.

	Total Publications	**Demographic Info is applicable**
	N	**N (%)**
Conference by year		
2015	91	59 (64.83%)
2016	82	64 (78.05%)
2017	51	47 (92.16%)
2018	60	44 (73.33%)
2019	64	47 (73.43%)
Journal by year		
2015	13	8 (61.54%)
2016	6	6 (100.00%)
2017	6	5 (83.33%)
2018	7	4 (57.14%)
2019	5	4 (80.00%)
Publication type		
Conf (Full)	136	101 (74.26%)
Conf (Short)	212	160 (75.47%)
Journal	37	27 (72.97%)
All papers	385	288 (74.81%)

4.2. Demographic Information, Reported and Analyzed

Next, we look at what demographic categories were reported among the publications that were coded as *applicable.* Within this category are papers that *analyzed* demographic data; these included papers that *investigated* demographics, that used demographics as input *features in the model*, and that used the demographics for *testing and validation.* There are also papers where demographic information was merely *reported* and also papers where no demographic information is reported at all.

4.2.1. Which demographic categories were analyzed or reported?

Table 4 shows which of the six demographic categories (age, gender, race, geographic location, educational level, and SES) were *reported* or *analyzed* across publication type. It also calculates the percentage of papers that did so as a function of the number of papers where demographic information was deemed *applicable.* Note that some of the papers that *analyzed* demographic

data reported on more demographic categories than were analyzed. That is, a study could report the age and grade level of the students, but only incorporate gender in the analyses.

Table 4: Frequency that each of the demographic categories was reported across all publications for which demographic information was applicable.

	At least one		Age		Ed. Level		Gender		Race		SES		Geographic Loc	
	N	**%**	**N**	**%**	**N**	**%**	**N**	**%**	**N**	**%**	**N**	**%**	**N**	**%**
Conference by year														
2015	42	71%	10	17%	33	56%	17	29%	3	5%	2	3%	22	37%
2016	38	59%	15	23%	32	50%	13	20%	2	3%	0	0%	27	42%
2017	31	66%	5	11%	29	62%	7	15%	3	6%	2	4%	17	36%
2018	34	77%	8	18%	28	64%	12	27%	6	14%	4	9%	25	57%
2019	38	81%	8	17%	29	62%	12	26%	6	13%	4	9%	19	40%
Journal by year														
2015	8	100%	3	38%	7	88%	5	63%	1	13%	1	13%	3	38%
2016	6	100%	5	83%	6	100%	5	83%	3	50%	0	0%	4	67%
2017	4	80%	1	20%	3	60%	0	0%	0	0%	0	0%	3	60%
2018	2	50%	1	25%	0	0%	1	25%	0	0%	0	0%	2	50%
2019	4	100%	0	0%	4	100%	1	25%	1	25%	1	25%	3	75%
Publication type														
Conf (Full)	71	70%	25	25%	60	59%	32	32%	9	9%	7	7%	39	39%
Conf (Short)	112	70%	21	13%	91	57%	29	18%	11	7%	5	3%	71	44%
Journal	24	89%	10	37%	20	74%	12	44%	5	19%	2	7%	15	56%
Total	207	72%	56	19%	171	59%	73	25%	25	9%	14	5%	125	43%

As these results show, journal articles are more likely than conference proceedings to report at least one demographic variable (89% for journals vs. 70% for both full and short conference papers). However, it is important to note that the smaller number of papers published in the journal each year results in larger variations in reporting frequency when comparing each publication year for the journal. For example, the publication year 2018 only contained four journal papers for which demographic information was considered applicable.

Moreover, the reporting of different demographic categories was uneven, with some categories reported much more frequently than others (from 5% to 59%). However, general trends can be observed in regard to which categories are more often reported than others. For instance, SES is almost always the least reported category, while education level is almost always the most reported. (The only exception is for 2018 journal articles, where only two papers reported demographic information).

4.2.2. Which kinds of analyses were most common?

We looked in more detail at the papers that *analyzed* demographics. In total, 44 of the 288 publications where demographic data was deemed *applicable* were found to have *analyzed* at least one demographic category. As Table 5 shows, 20 investigated potential differences that demographic data might help to explain in the construct being studied. Another half of these papers (22) used the demographic data as an input feature when developing a model, and only 4 explicitly used demographics in the testing/validation process.

Table 5: List of the 44 papers that used demographics in at least one of their analyses, arranged alphabetically by year. Type of analyses is given, and papers marked with * are repeated in two levels of analyses. Analyses did not always include every category of Demographics Reported.

		Analyses Type				Demographics Reported				
Authors	**Pub. Venue**	**Inv**	**FiM**	**T&V**	**Age**	**Ed Lev**	**Sex**	**Race**	**SES**	**Geo Loc**
Bhatnagar et al., 2015	Conf-S	X	--	--	--	X	--	--	--	--
Bravo et al., 2015	Conf-S	X	--	--	X	X	X	--	--	X
Dee Miller et al., 2015	Journal	--	X	--	--	X	X	--	--	--
Ezen-Can & Boyer, 2015	Conf-F	--	X	--	X	X	X	--	--	--
Knowles, 2015	Journal	--	X	--	--	X	X	X	X	X
Luo et al., 2015	Conf-S	--	X	--	--	X	--	--	--	X
Nižnan et al., 2015	Conf-F	--	--	X	--	--	--	--	--	X
Riddle et al., 2015	Conf-F	--	X	--	--	X	X	X	--	--
Rowe et al., 2015	Conf-S	X	--	--	--	X	X	--	--	--
Saarela & Kärkkäinen, 2015*	Conf-F	X	X	--	X	--	X	--	X	X
Samei et al., 2015	Conf-S	--	--	X	--	X	--	--	--	X
Schneider & Blikstein, 2015	Journal	X	--	--	X	X	X	--	--	--
Strecht et al., 2015	Conf-S	--	X	--	X	X	X	X	X	X
Warner et al., 2015	Conf-S	X	--	--	--	X	--	--	--	X
Zheng et al., 2015	Conf-S	--	X	--	--	--	X	--	--	--
Zimmerman et al., 2015	Journal	--	X	--	X	X	X	--	--	X
Bydžovská, 2016	Conf-S	--	X	--	X	X	X	--	--	--
Crossley et al., 2016	Journal	X	--	--	X	X	X	X	--	X
Feng et al., 2016	Conf-S	X	--	--	X	X	X	--	--	--
Labarthe et al., 2016	Conf-S	--	--	X	X	X	X	--	--	X
Liu et al., 2016	Conf-F	X	--	--	--	--	--	--	--	X
Rowe et al., 2016	Conf-S	X	--	--	--	X	X	--	--	--
Spoon et al., 2016*	Journal	X	X	--	X	X	X	X	--	X
Sweeney et al., 2016	Journal	--	X	--	X	X	X	X	--	X
Yang et al., 2016	Journal	X	--	--	X	X	X	--	--	--
Kai et al., 2017	Conf-S	--	--	X	X	X	X	X	X	--
Liu et al., 2017	Conf-F	--	X	--	--	X	X	X	--	--
Madaio et al., 2017	Conf-S	--	X	--	X	--	X	--	--	--
Backenköhler et al., 2018	Conf-S	--	X	--	X	X	X	X	--	X
Chopra et al., 2018	Conf-F	--	X	--	--	--	X	--	--	X
Crues et al., 2018	Conf-F	--	X	--	X	--	X	--	--	--
Du et al. 2018	Conf-S	X	--	--	--	X	X	--	--	X
Naismith et al., 2018	Conf-S	X	--	--	--	--	--	--	X	--
Nguyen & Liew, 2018	Conf-S	--	--	X	--	X	--	--	--	X
Park et al., 2018	Conf-F	X	--	--	--	X	X	X	X	--
Pytlarz et al., 2018	Conf-S	--	X	--	--	X	X	X	X	X
Slim et al., 2018	Conf-S	--	X	--	--	X	X	X	X	X
Aulck et al., 2019	Conf-F	--	X	--	X	X	X	X	X	X
Hutt et al., 2019	Conf-F	X	--	--	--	X	X	X	X	--
Jensen et al., 2019	Conf-S	--	X	--	--	X	X	X	X	X
Karumbaiah et al., 2019	Conf-F	X	--	--	--	X	--	X	X	X
Palaez et al., 2019	Journal	X	--	--	--	X	X	X	X	X
Ren et al., 2019	Conf-F	X	--	--	--	X	--	--	--	--
Toda et al., 2019	Conf-S	X	--	--	X	--	X	--	--	--

These results (summarized in Table 5) suggest that, when researchers use demographic information, it is mostly to take advantage of demographic information to improve their models (i.e., as *features in models*) or to answer research questions related to demographic information (i.e., *investigated*). Few publications used demographic information to validate that the results of the analyses were not biased (i.e., *testing & validation*).

4.2.3. Which types of analyses appeared in which kinds of publications?

Table 6 summarizes this information according to the same publication categories reported above, comparing the frequency in which at least one demographic category was included in the various kinds of analyses to the frequency in which demographics were only reported (without analyzing any demographics) and the frequency in which demographics were not reported at all. Because each sub-category of analyzed (i.e., investigated, features in models, and testing & validation) is independent, a publication could use demographic information in multiple ways. However, the reported category is treated as exclusive; a publication is considered as reported information only if the information was reported but not explicitly used. We also counted the number of publications that used demographic information for at least one of the three identified uses.

Table 6: Frequency for each category of use for demographic information across all publications for which demographics were *applicable*. The *analyzed* category includes any publication coded as *investigated, features in models,* or *testing and validation*.

	Investigated		**Feature in Model**		**Testing & validation**		**Analyzed (Subtotal)**		**Reported**		**Not Reported**	
	N	**%**	**N**	**%**	**N**	**%**	**N**	**%**	**N**	**%**	**N**	**%**
Conference												
2015	5	8%	6	10%	2	3%	12	20%	29	49%	18	31%
2016	3	5%	1	2%	1	2%	5	8%	34	53%	25	39%
2017	0	0%	2	4%	1	2%	3	6%	28	60%	16	34%
2018	3	7%	5	11%	1	2%	9	20%	25	57%	10	23%
2019	4	9%	2	4%	0	0%	6	13%	32	68%	9	19%
Journal												
2015	1	13%	3	38%	0	0%	4	50%	4	50%	0	0%
2016	3	50%	2	33%	0	0%	4	67%	2	33%	0	0%
2017	0	0%	0	0%	0	0%	0	0%	4	80%	1	20%
2018	0	0%	0	0%	0	0%	0	0%	2	50%	2	50%
2019	1	25%	0	0%	0	0%	1	25%	3	75%	0	0%
Paper Type												
Conf (Full)	6	6%	7	7%	1	1%	13	13%	56	55%	32	32%
Conf (Short)	9	6%	9	6%	4	3%	22	14%	92	58%	46	29%
Journal	5	19%	5	19%	0	0%	9	33%	15	56%	3	11%
Total	20	7%	21	7%	5	2%	44	15%	163	57%	81	28%

5. DISCUSSION

Results from the survey of publications for which demographic information is applicable shows that, despite the important technological focus of EDM research, EDM research generally involves an important human component as research project often involves the study of human behaviors or include data from key actors (e.g., teachers, students, administrators, etc.). Overall, demographic information was considered *applicable* for 75% of the inventoried publications. Despite observed variations in this percentage across venues and years, no obvious trend is observed based on the year of publications. When considering the different types of publications, the percentage of papers for which demographic information is applicable is fairly stable, ranging from 73% for journal papers to 75% for short conference papers.

5.1. REPORTING DEMOGRAPHIC CATEGORIES

While a majority of publications for which demographic information is *applicable* described at least one category (72%), they varied greatly in which categories they included. Some categories, such as the education level of participants in the study, are often included (59% of the time), whereas others such as SES are much less common (5%). Additionally, while there are no clear trends in reporting practices when comparing full and short conference proceedings, journal articles consistently report more categories of demographic information (see Figure 1). We hypothesize that this might be due to the nature of journal articles, which tend to have more space to provide demographic information and often present more mature work.

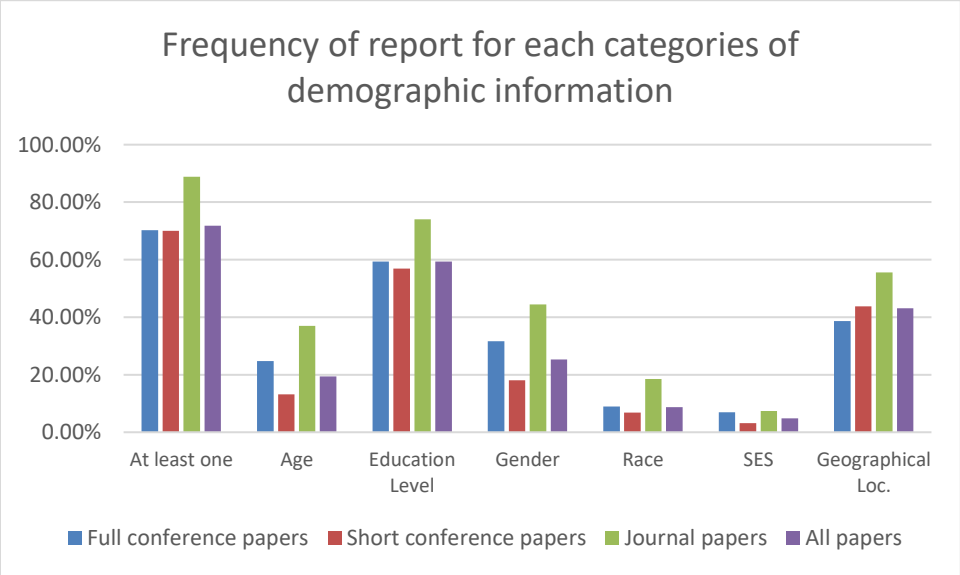

Figure 1: Percentage of time each category of demographic information was *reported* (among publications for which demographic information was considered *applicable*) for all papers and for each type of publication.

Unsurprisingly, the two demographic categories that are most often included in EDM publications are geographic location and education level (respectively 43% and 59%). Researchers regularly report the geographic location of the school where the data was collected (with varying levels of specificity), and the education level is often described alongside the learning domain under investigation (e.g., "8th-grade mathematics class" or an "undergraduate

science course"). This kind of demographic information is easy to document, as it does not require the collection of additional data from each unique participant in the study. However, if other important demographic categories are not included, these descriptions may over-represent the homogeneity of the student population. That is, when we describe students with terms like "9th graders from a small, suburban school district," without capturing other categories known to affect educational behaviors, we exaggerate the uniformity of the data.

Categories that are more closely tied to individual students—as opposed to the entire population of the study—were less frequently reported. Among the other four demographic categories, we considered, age (19%) and gender (25%) appear to be reported with similar frequency (although the former may sometimes be roughly inferred by the education level), while race and socioeconomic status lag behind considerably. The amount of effort (and opportunity) required to collect information about individual participants, combined with the increased risks to being able to personally (re)identify research participants, appears to hinder this kind of data collection. Indeed, the frequency in which different demographic categories were reported was inversely proportional to how sensitive it is considered, with race and SES reported quite infrequently (respectively 9% and 5%).

To be fair, there are altruistic reasons to avoid collecting sensitive data, as some researchers may worry that the increased institutional and regulatory requirements could limit their ability to work with populations who are most in need. For example, sampling bias may increase when consent forms are required (e.g., Noll et al., 1997). However, there are ways to help mitigate these issues (see, for example, discussion in Fletcher & Hunter, 2003), and there is a good reason that this type of personal information is important to consider when investigating issues of equity and fairness.

Other sub-populations of students that we did not include in our survey, such as Second Language Learners and students with learning disabilities, are important to consider as well. We did not specifically code for these categories, as very few EDM papers included information related to them. One exception is Naismith et al. (2018), who showed the importance of using datasets containing data from English Second Learners (ESL) to more accurately measure lexical sophistication. Another exception (Klinger et al., 2017) involved data from students suffering from *developmental dyscalculia*. As was the case for race and SES, information related to special needs status and learning disabilities might be difficult to obtain as it can be very sensitive. However, it is important to consider it as it is likely to impact the student's learning experience.

5.2. Analyzing demographics

Although a majority of publications reported at least one category of demographic information (72%), only 15% *analyzed* demographics. In most publications, demographics were either only *reported* the information (57%) or were *not reported* at all (28%).

The use of demographics does vary by publication type. Journal articles are more likely to use demographic information in their analysis than conference proceedings of either length, and it is possible that the space affordances of journals contribute to these findings. It is important to note, however, that use of demographic information is very uneven from year to year, both for journal and conference publications. (See Table 6, above.)

In terms of how demographic information is *analyzed* (Figure 2), the most common usage for demographic information is as input *features for models* (8%), closely followed by *investigated* (7%). The use of demographic information for *testing and validation* is much less frequent (2%).

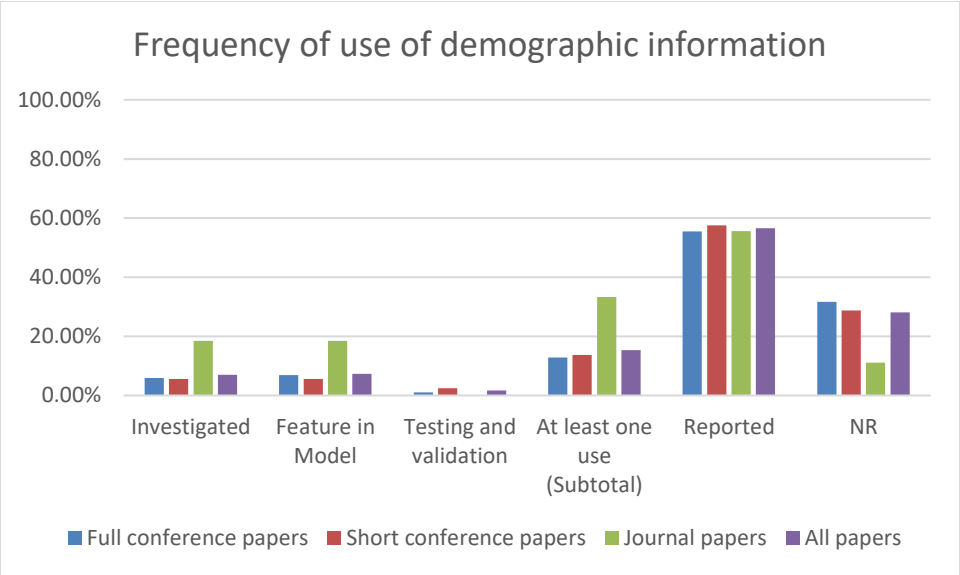

Figure 2: Percentage of time different use for demographic information was included in analyses (among publications for which demographic information was considered applicable), for all papers and for each type of publication.

5.3. Implications for fairness in EDM research

While the survey presented in this article does not directly quantify existing biases in EDM research, it provides a big picture of current efforts that could lead to addressing issues of bias and fairness through the collection, reporting, and analysis of demographics. It appears that there is a general recognition of the importance of demographics in educational research (72% of publications reported at least one demographic category). A majority (59%) of publications provided information related to the education level of the study's participants, and 43% reported the geographic location of the study. This information is important when considering bias in who participates in EDM research and to assess whether results could be expected to generalize across different contexts, but these variables are not sufficient for describing students' sociocultural background.

Moreover, variables that are more likely to be related to historical problems with educational equity are much less likely to be reported. Data related to gender, race, and SES were reported in 25%, 9%, and 5% of publications, respectively. Members of our community (e.g., Holstein et al., 2019) have identified a perceived need for fairness-aware data collection as an important component of addressing fairness, but this work surveyed the machine learning community more broadly. The result of our survey suggests the EDM community has a similar need and should make conscious efforts to consider bias and fairness when collecting, reporting, and testing our data.

Reporting demographic information, by itself, can contribute to fairer research in EDM as it provides a better understanding of the context in which EDM research takes place. However, greater efforts might be required to truly address issues of bias and fairness. The low number of publications (15%) *analyzing* demographic information at all suggests that there is still room for reflection on the utility of demographic variables beyond describing the context in which research takes place. While 7% of publications directly investigated research questions related to demographics, the most frequent use of demographic variables was in the context of model

building. This form of analysis might generally improve the quality of the models when assessed using conventional metrics. However, it is unclear whether it impacts issues of bias and fairness. Only a very small amount of publications (2%) used demographics to directly assess model biases through testing and validation. These numbers are unlikely to convince any researchers who specialize in equity that our field is taking this issue seriously, and in some countries, this oversight might soon become subject to legal consequences.

6. CONCLUSION

The result of our survey of the use of demographic information in EDM research reveals a genuine effort in including equity and fairness in EDM research through demographic information. However, this effort is still mostly confined to a restricted number of publications that specifically study these issues, either through direct investigation of demographic variables or through the inclusion of demographic variables as input features for data analysis.

We believe that there is room for these issues to be considered in more publications, especially through the usage of demographic variables to test and validate how general models are across diverse populations of students. Common validation approaches in EDM research have evolved as the field has grown – over the history of EDM, we have seen a shift from no cross-validation or test sets, to the use of simple (flat) cross-validation, to higher-level cross-validation (e.g., student-level cross-validation to evaluate generalization to new students, etc.). We believe that a similar shift could occur to further validate how well models perform when applied to different subpopulations of students, to ensure that results of EDM analyses are fair and equitable.

Given the wide-ranging evidence about the potential for demographic biases to emerge in inferential analysis and the well-documented relationship between students' sociocultural background and their educational outcomes, we would suggest that this might be the only way to ensure that our algorithms are *value-aligned.* That is, it is not enough to treat all students the same (regardless of needs); we must instead ensure that all students are having their needs met by the systems and algorithms we design.

In an ideal scenario, using demographic information to further validate the results of EDM research should not demand a large amount of additional effort. However, in order to conduct such validation, it is necessary to collect the relevant demographic information when conducting studies. This can prove to be challenging, as EDM research sometimes employs historical data that were collected without demographic information or data from platforms without the ability to require students to provide this information. Additionally, the collection of demographic information during new studies can create additional challenges due to the sensitive nature of some demographic variables (e.g., race and SES). Studies might require additional protective measures to ensure the privacy of participants, might require more strict review by institutional review boards, and some participants might be less likely to participate in studies with this type of information is requested from them.

The greater responsibilities of such sensitive data may reduce its availability. For example, it has largely prevented the collection of some forms of data, such as student disability status, which could be very important for guaranteeing equal effectiveness. We simply do not know if learning technology is working well for students with many disabilities. While this protection is understandable, the subsequent avoidance of collecting disability-related information means that any student with a learning disability is essentially a guinea pig, even when they have access to the same technology that has been extensively vetted for other students in his or her class.

This situation is certainly *not* value-aligned, and neither are other uses where we have not shown that a technology or algorithm is socio-culturally appropriate for any student. That is, since we know that demographic categories regularly interact with educational outcomes, we have an obligation to test and validate our models in ways that align with that knowledge.

Though additional effort is needed to collect demographic data, we believe that the additional efforts required to conduct such research will have long term benefits for the EDM community. Since research has shown the existence of biases in other applications of machine learning and EDM research often deals with data representing human behaviors, EDM research is at risk of producing biased research and inequitable systems. While, ongoing research from communities such as the Fairness, Accountability and Transparency in Machine Learning (ACM FAccT)[5] community have been working towards the development of fairness measures (Corbett-Davies, & Goel, 2018; Gaiane & Pechenizkiy, 2018; Yang & Stovanovich, 2017) and algorithmic solutions (Celis et al., 2019; Kamiran, Calders, & Pechenizkiy, 2010; Kamishma, Akaho, & Sakuma, 2011; Zemel et al., 2013) to address biases and fairness in machine learning applications, there is a need to study how those approaches can be applied to educational data, identify the specific characteristics of fairness in educational data and develop methods that are adapted to those characteristics.

We believe that the EDM community, due to the interdisciplinary nature of the field of EDM, combining technical expertise in machine learning and deep knowledge of education issues, is uniquely positioned to investigate and address issues of biases, equity, and fairness in educational data, advancing the fair applications of machine learning to benefit all members of society. Moreover, when there is a group that has historically been denied opportunities, it is important (both legally and morally) that we are not designing learning environments that exasperate these social issues by systematically underserving a particular population. Good intentions are important, but since cultural and behavioral differences often interact in unpredictable ways, it is important that we follow Lynch and Stucker's (2012) advice to always show our data.

ACKNOWLEDGMENTS

The research presented in this paper was partially funded by NSF grant #DRL-1661987.

EDITORIAL STATEMENT

Ryan Baker had no involvement with the journal's handling of this article in order to avoid a conflict with his Associate Editor role. The entire review process was managed by Special Guest Editor Luke G. Eglington.

REFERENCES

AHSEN, M. E., AYVACI, M., AND RAHUNATHAN, S. 2017. When algorithmic predictions use human-generated data: A bias-aware classification algorithm for breast cancer diagnosis. *Information Systems Research* 30, 1, 97-116.

[5] https://facctconference.org/

AGRANOVICH, A.V., PANTER, A.T., PUENTE, A.E., AND TOURADJI, P. 2011. The culture of time in neuropsychological assessment: Exploring the effects of culture-specific time attitudes on timed test performance in Russian and American samples. *Journal of the International Neuropsychological Society* 17, 4, 692-701.

ALVI, M., ZISSERMAN, A., AND NELLAKER, C. 2018. Turning a blind eye: Explicit removal of biases and variation from deep neural network embeddings. In L. Leal-Taixé, & S. Roth (Eds.) *Proceedings of the European Conference on Computer Vision 2018 Workshops*, Munich, Germany, 556-572.

AULCK, L., NAMBI, D., VELAGAPUDI, N., BLUMENSTOCK, J., AND WEST, J. 2019. Mining university registrar records to predict first-year undergraduate attrition. In C.F. Lynch, A. Merceron, M. Desmarais, & R. Nkambou (Eds.) *Proceedings of the 12th International Conference on Educational Data Mining*, Montreal, Canada, 9-18.

BACKENKOHLER, M., SCHERZINGER, F., SINGLA, A., AND WOLF, V. 2018. Data-driven approach towards a personalized curriculum. In K.E. Boyer, & M. Yudelson (Eds.) *Proceedings of the 11th International Conference on Educational Data Mining*, Buffalo, NY, 246-251.

BAKHTIN, M.M. 2010. *The dialogic imagination: Four essays (Vol. 1)*. University of Texas Press.

BANDURA, A. 2001. Social cognitive theory: An agentic perspective. *Annual Review of Psychology* 52, 1, 1-26.

BAROCAS, S., AND SELBST, A.D. 2016. Big data's disparate impact. *California Law Review* 104, 671.

BHARTIYA, D., CONTRACTOR, D., BISWAS, S., SENGUPTA, B., AND MOHANIA, M. 2016. Document segmentation for labeling with academic learning objectives. In T. Barnes, M. Chi, & M. Feng (Eds.) *Proceedings of the 9th International Conference on Educational Data Mining*, Raleigh, NC, 282-287.

BHATNAGAR, S., DESMARAIS, M., WHITTAKER, C., LASRY, N., DUGDALE, M., LENTON, K., AND CHARLES, E. 2015. An analysis of peer-submitted and peer-reviewed answer rationales in a web-based peer instruction based learning environment. In O.C. Santos, J.G. Boticario, C. Romero, M. Pechenizkiy, A. Merceron, J.M. Luna, C. Mihaescu, P. Moreno, A. Hershkovitz, S. Ventura, & M. Desmarais (Eds.) *Proceedings of the 8th International Conference on Educational Data Mining*, Madrid, Spain, 456-459.

BIERDMAN, J., FARAONE, S.V., AND MONUTEAUX, M.C. 2002. Differential effect of environmental adversity by gender: Rutter's index of adversity in a group of boys and girls with and without ADHD. *American Journal of Psychiatry* 159, 9, 1556-1562.

BLODGETT, S.L., AND O'CONNOR, B. 2017. Racial disparity in natural language processing: A case study of social media African-American English. In *Workshop on Fairness, Accountability and Transparency in Machine Learning (FATML)*, Halifax, Nova Scotia, arXiv:1707.00061.

BOLUKBASI, T., CHANG, K.W., ZOU, J., Y., SALIGRAMA, B., AND KALAI, A.T. 2016. Man is to computer as woman is to homemaker? Debiasing word embeddings. In *Advances in Neural Information Processing Systems*, Barcelona, Spain, 4349-4357.

BOSTROM, N. 2016. *Superintelligence: Paths, dangers, strategies*. Oxford University Press: Oxford.

BRAVO, J., ROMERO, S.J., LUNA, M., AND PAMPLONA, S. 2015. Exploring the influence of ICT in online education through data mining tools. In O.C. Santos, J.G. Boticario, C. Romero, M. Pechenizkiy, A. Merceron, J.M. Luna, C. Mihaescu, P. Moreno, A. Hershkovitz, S. Ventura, & M. Desmarais (Eds.) *Proceedings of the 8th International Conference on Educational Data Mining*, Madrid, Spain, 540-543.

BROSCH, T., BAR-DAVID, E., AND PHELPS, E.A. 2013. Implicit race bias decreases the similarity of neural representations of black and white faces. *Psychological Science* 24, 2, 160-166.

BURR, C., CRISTIANINI, N., AND LADYMAN, J. 2018. An analysis of the interaction between intelligent software agents and human users. *Minds and Machines* 28, 4, 735-774.

BYDŽOVSKÁ, H. 2016. Course enrollment recommender system. In T. Barnes, M. Chi, & M. Feng (Eds.) *Proceedings of the 9th International Conference on Educational Data Mining*, Raleigh, NC, 312-317.

CALDERS, T., AND VERWER, S. 2010. Three naïve Bayes approaches for discrimination-free classification. *Data Mining and Knowledge Discovery* 21, 2, 277-292.

CALISKKAN, A., BRYSON, J.J., AND NARAYANAN, A. 2017. Semantics derived automatically from language corpora contain human-line biases. *Science* 356, 6334, 183-186.

CELLIS, L.E., HUANG, L., KESWANI, V., AND VISHNOI, N.K. 2019. Classification with fairness constraints: A meta-algorithm with provable guarantees. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, Atlanta, GA, 319-328.

CHEN, L., MA, R., HANNÁK, A., AND WILSON, C. 2018. Investigating the impact of gender on rank in resume search engines. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, Montréal, Canada, 651.

CHILDS, D.S. 2017. Effects of math identity and learning opportunities on racial differences in math engagement, advanced course-taking, and STEM aspiration. *PhD Dissertation*, Temple University.

CHOPRA, S., GAUTREAU, H., KHAN, A., MIRSAFIAN, M., AND GOLAB, L. 2018. Gender differences in undergraduate engineering applicants: A text mining approach. In K.E. Boyer, & M. Yudelson (Eds.) *Proceedings of the 11th International Conference on Educational Data Mining*, Buffalo, NY, 44-54.

CHOULDECHOVA, A. 2017. Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *Big Data* 5, 2, 153-163.

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.

CROSSLEY, S., ALLEN, L.K., SNOW, E.L., AND MCNAMARA, D.S. 2016. Incorporating learning characteristics into automatic essay scoring models: What individual differences and linguistic features tell us about writing quality. *Journal of Educational Data Mining* 8, 2, 1-19.

CRUES, R.W., BOSCH, N., ANDERSON, C.J., PERRY, M., BHAT, S., AND SHAIK, N. 2018. Who they are and what they want: Understanding the reasons for MOOC enrollment. In K.E. Boyer, & M. Yudelson (Eds.) *Proceedings of the 11th International Conference on Educational Data Mining*, Buffalo, NY, 177-186.

D'ALESSANDRO, B., O'NEIL, C., AND LAGATTA, T. 2017. Conscientious classification: A data scientist's guide to discrimination-aware classification. *Big Data* 5, 2, 120-134.

DEE MILLER, L., SOH, L.-K., SAMAL, A., KUPZYK, K., AND NUGENT, G. 2015. A comparison of educational statistics and data mining approaches to identify characteristics that impact online learning. *Journal of Educational Data Mining* 7, 3, 117-150.

DOROUDI, S., AND BRUNSKILL, E. 2017. The misidentified identifiability problem of bayesian knowledge tracing. In X. Hu, T. Barnes, A. Hershkovitz, & Paquette, L. (Eds.) *Proceedings of the 10th International Conference on Educational Data Mining*, Wuhan, China, 143-149.

DU, X., DUIVESTEIJN, W., KLABBERS, M., AND PECHENIZKIY, M. 2018. ELBA: Exceptional Learning Behavior Analysis. In K.E. Boyer, & M. Yudelson (Eds.) *Proceedings of the 11th International Conference on Educational Data Mining*, Buffalo, NY, 312-317.

ECKERT, P. 1989. The whole woman: Sex and gender differences in variation. *Language Variation and Change* 1, 3, 245-267.

EZEN-CAN, A., AND BOYER, K.E. 2015. Choosing to interact: Exploring the relationship between learner personality, attitudes, and tutorial dialogue participation. In O.C. Santos, J.G. Boticario, C. Romero, M. Pechenizkiy, A. Merceron, J.M. Luna, C. Mihaescu, P. Moreno, A. Hershkovitz, S. Ventura, & M. Desmarais (Eds.) *Proceedings of the 8th International Conference on Educational Data Mining*, Madrid, Spain, 125-128.

FELDMAN, M., FRIEDLER, S.A., MOELLER, J., SCHEIDEGGER, C., AND VENKATASUBRAMANIAN, S. 2015. Certifying and removing disparate impact. In *Proceedings of the 21st ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, Sydney, Australia, 259-268.

FENG, M., ROSCHELLE, J., MASON, C., AND BHANOT, R. 2016. Investigating gender difference on homework in middle school mathematics. In T. Barnes, M. Chi, & M. Feng (Eds.) *Proceedings of the 9th International Conference on Educational Data Mining*, Raleigh, NC, 364-369.

FERGUSON, R. 2012. Learning Analytics: Drivers, developments and challenges. *International Journal of Technology Enhanced Learning* 4, 5/6, 304-317.

FLETCHER, A.C., HUNTER, A.G. 2003. Strategies for obtaining parental consent to participate in research. *Family Relations* 52, 3, 216-221.

FRIEDMAN, B. 1996. Value-sensitive design. *Interactions* 3, 6, 16-23.

GAIANE, P., AND PECHENIZKIY, M. 2018. On formalizing fairness in prediction with machine learning. *arXiv preprint*, arXiv:1710.03184.

GALHOTRA, S., BRUN, Y., AND MELIOU, A. 2017. Fairness testing: testing software for discrimination. In *Proceedings of the 11th Joint Meeting on Foundations of Software Engineering*, Paderborn, Germany, 498-510.

GARDNER, J., BROOKS, C., AND BAKER, R. 2019. Evaluating the fairness of predictive student models. In *Proceedings of the 9th International Conference on Learning Analytics and Knowledge*, Tempe, AZ.

GAUB, M., CARLSON, C.L. 1997. Gender differences in ADHD: A meta-analysis and critical review. *Journal of the American Academy of Child & Adolescent Psychiatry* 36, 8, 1036-1045.

GELLERT R., DE VRIES, K., DE HERT, P., AND GUTWIRTH, S. 2013. A comparative analysis of anti-discrimination and data protection legislations. In *Discrimination and Privacy in the*

Information Society, B. Custers, T. Calders, B. Schermer, and T. Zarsky, Eds. Springer, Berlin, 61-89.

GIANFRANCESCO, M.A., TAMANG, S., YAZDANY, J., AND SCHMAJUK, G. 2018. Potential biases in machine learning algorithms using electronic health record data. *JAMA Internal Medicine* 178, 11, 1544-1547.

GOODMAN, B., AND FLAXMAN, S. 2017. European Union regulations on algorithmic decision-making and a "right to explanation". *AI Magazine* 38, 3, 50-57.

GOODMAN, S.N., GOEL, S., AND CULLEN, M.R. 2018. Machine learning, health disparities, and causal reasoning. *Annals of Internal Medicine* 169, 12, 883-884.

GOSSE, D., AND ARNOCKY, S. 2012. The state of Canadian boyhood–beyond literacy to a holistic approach. *in Education* 18, 2, 67-97.

HACKER, P., AND WIEDMANN, E. 2017. A continuous framework for fairness. *arXiv preprint*, arXiv:1712.07924.

GUTIÉRREZ, K.D., AND ROGOFF, B. 2003. Cultural ways of learning: Individual traits or repertoires of practice. *Educational* Researcher 32, 5, 19-25.

HADFIELD-MENELL, D., RUSSELL, S.J., ABBEEL, P., AND DRAGAN, A. 2016. Cooperative inverse reinforcement learning. In *Advances in Neural Information Processing Systems*, Barcelona, Spain, 3909-3917.

HAJIAN, S., AND DOMINGO-FERRER. 2013. A methodology for direct and indirect discrimination prevention in data mining. *IEEE Transactions on Knowledge and Data Engineering* 25, 7, 1445-1459.

HENDRIX, L.A., BURNS, K., SAENKO, K., DARRELL, T., AND ROHRBACH, A. 2018. Women also snowboard: Overcoming bias in captioning models. In *European Conference on Computer Vision*, Munich, Germany, 793-811.

HOLSTEIN, K., WORTMAN VAUGHAN, J., DAUMÉ III, H., DUDIK, M., WALLACH, H. 2019. Improving fairness in machine learning systems: What do industry practitioners need? In *Proceedings of the ACM CHI Conference on Human Factors in Computer Systems*, Glasgow, UK, 1-16.

HOVY, D. 2015. Demographic factors improve classification performance. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics at the 7th International Joint Conference on Natural Language Processing volume 1*, Beijing, China, 752-762.

HOWARD, A., ZHANG, C., AND HORVITZ, E. 2017. Addressing bias in machine learning algorithms: A pilot study on emotion recognition for intelligent systems. In *IEEE Workshop on Advanced Robotics and Its Social Impacts*, Genoa, Italy, 1-7.

HUTT, S., GARDNER, M., DUCKWORTH, A.L., AND D'MELLO, S. 2019. Evaluating fairness and generalizability in models of predicting on-time graduation from college applications. In C.F. Lynch, A. Merceron, M. Desmarais, & R. Nkambou (Eds.) *Proceedings of the 12th International Conference on Educational Data Mining*, Montreal, Canada, 79-88.

JATON, F. 2017. We get the algorithms of our ground truths: Designing referential databases in digital image processing. *Social Studies of Science* 47, 6, 811-840.

JENSEN, E., HUTT, S., AND D'MELLO, S.K. 2019. Generalizability of sensor-free affect detection models in a longitudinal dataset of tens of thousands of students. In C.F. Lynch,

A. Merceron, M. Desmarais, & R. Nkambou (Eds.) *Proceedings of the 12th International Conference on Educational Data Mining*, Montreal, Canada, 324-329.

JURGENS, D., TSVETKOV, Y., AND JURAFSKY, D. 2017. Incorporating dialectal variability for socially equitable language identification. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, Vancouver, Canada, 51-57.

KAI, S., ANDRES, J.M., PAQUETTE, L., BAKER, R., MOLNAR, K., WATKINS, H., MOORE, M. 2017. Course enrollment recommender system. In *Proceedings of the 10th International Conference on Educational Data Mining*, Wuhan, China, 250-255.

KAMIRAN, F., AND CALDERS, T. 2012. Data preprocessing techniques for classification without discrimination. *Knowledge and Information Systems* 33, 1, 1-33.

KAMIRAN, F., CALDERS, T., AND PECHENIZKIY, M. 2012. Discrimination aware decision tree learning. In *Proceedings of the 10th IEEE International Conference on Data Mining*, Sydney, Australia, 869-874.

KAMISHIMA, T., AKAHO, S., AND SAKUMA, J. 2011. Fairness-aware learning through regularization approach. In *Proceedings of the 11th IEEE International Conference on Data Mining Workshops*, Vancouver, Canada, 643-650.

KAMISHIMA, T., AKAHO, S., ASOH, H., AND SAKUMA, J. 2012. Fairness-aware classifier with prejudice remover regularizer. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, Bristol, United Kingdom, 35-50.

KARUMBAIAH, S., OCUMPAUGH, J., AND BAKER, R.S. 2019. The influence of school demographics on the relationship between students' help-seeking behavior and performance and motivational measures. In C.F. Lynch, A. Merceron, M. Desmarais, & R. Nkambou (Eds.) *Proceedings of the 12th International Conference on Educational Data Mining*, Montreal, Canada, 99-108.

KHAJAH, M., LINDSEY, R.B., & MOZER, M.C. 2016. How deep is knowledge tracing? In T. Barnes, M. Chi, & M. Feng (Eds.) *Proceedings of the 9th International Conference on Educational Data Mining*, Raleigh, NC, 94-102.

KINZIE, AND M.B., JOSEPH, D.R. 2008. Gender differences in game activity preferences of middle school children: Implications for educational game design. *Educational Technology Research and Development* 56, 5-6, 643-663.

KIRITCHENKO, S., AND MOHAMMAD, S.M. 2018. Examining gender and race bias in two hundred sentiment analysis systems. In *Proceedings of the 7th Joint Conference on Lexical and Computational Semantics*, New Orleans, LA, 43-53.

KLARE, B.F., BURGE, M.J., KLONTZ, J.C., VORDER BRUEGGE, R.W., AND JAIN, A.K. 2012. Face recognition performance: Role of demographic information. *IEEE Transactions on Information Forensics and Security* 7, 6, 1789-1801.

KNOWLES, J.E. 2015. Of needles and haystacks: Building an accurate statewide dropout early warning system in Wisconsin. *Journal of Educational Data Mining* 7, 3, 1-17.

LABARTHE, H., BOUCHET, F., BACHELET, R., AND YACEF, K. 2016. Does a peer recommender foster students' engagement in MOOCs? In T. Barnes, M. Chi, & M. Feng (Eds.) *Proceedings of the 9th International Conference on Educational Data Mining*, Raleigh, NC, 418-423.

LABARTHE, H., LUENGO, V., AND BOUCHET, F. 2018. Analyzing the relationships between learning analytics, educational data mining and A.I. for education. In *Workshop on Learning Analytics: Building Bridges Between the Education and Computing Communities at the 14th International Conference on Intelligent Tutoring Systems*, Montreal, Canada, 10-19.

LADSON-BILLINGS, G. 1995. Towards a theory of culturally relevant pedagogy. *American Educational Research Journal* 32, 3, 465-491.

LADSON-BILLINGS, G. 1998. Just what is critical race theory and what's it doing in a nice field like education? *International Journal of Qualitative Studies in Education* 11, 1, 7-24.

LABUTOV, V., AND LIPSON, H. 2016. Web as a textbook: Curating targeted learning paths through the heterogeneous learning resources on the web. In T. Barnes, M. Chi, & M. Feng (Eds.) *Proceedings of the 9th International Conference on Educational Data Mining*, Raleigh, NC, 110-118.

LAUDERDALE, B.E., AND CLARK, T.S. 2014. Scaling politically meaningful dimensions using texts and votes. *American Journal of Political Science* 58, 3, 754-771.

LAVE, J. 1991. Situating learning in communities of practice. *Perspectives on Socially Shared Cognition* 2, 63-82.

LIU, Z., BROWN, R., LYNCH, C., BARNES, T., BAKER, R.S., BERGNER, Y., AND MCNAMARA, D. 2016. MOOC Learners behaviors by country and culture; an exploratory analysis. In T. Barnes, M. Chi, & M. Feng (Eds.) *Proceedings of the 9th International Conference on Educational Data Mining*, Raleigh, NC, 127-134.

LIU, Z., CODY, C., BARNES, T., LYNCH, C., AND RUTHERFORD, T. 2017. The antecedent of and associations with elective replay in an educational game: Is replay worth it? In X. Hu, T. Barnes, A. Hershkovitz, & Paquette, L. (Eds.) *Proceedings of the 10th International Conference on Educational Data Mining*, Wuhan, China, 40-47.

LUO, L., KOPRINSKA, I., AND LIU, W. 2015. Discrimination-aware classifiers for student performance prediction. In O.C. Santos, J.G. Boticario, C. Romero, M. Pechenizkiy, A. Merceron, J.M. Luna, C. Mihaescu, P. Moreno, A. Hershkovitz, S. Ventura, & M. Desmarais (Eds.) *Proceedings of the 8th International Conference on Educational Data Mining*, Madrid, Spain, 384-387.

LYNCH, J., AND STUCKLER, D. 2012. In God we trust, all others (must) bring data. *International Journal of Epidemiology* 41, 6, 1503-1506.

MADAIO, M., LASKO, R., CASSELL, J., AND OGAN, A. 2017. Using temporal association rule mining to predict dyadic rapport in peer tutoring. In X. Hu, T. Barnes, A. Hershkovitz, & Paquette, L. (Eds.) *Proceedings of the 10th International Conference on Educational Data Mining*, Wuhan, China, 318-323.

MADDOX, T.M., RUMSFELD, J.S., AND PAYNE, P.R.O. 2018. Questions for artificial intelligence in health care. *JAMA* 321, 1, 31-32.

MAKKONEN, T., 2007. Measuring discrimination: Data collection and the E.U. Equality Law. European Network of Legal Experts in Anti-Discrimination. (http://www.migpolgroup.com)

MCBRIDE, L., AND NICHOLS, A. 2016. Retooling poverty targeting using out-of-sample validation and machine learning. *World Bank Economic Review* 32, 3, 531-550.

MILNER IV, H.R. 2012. Beyond a test score: Explaining opportunity gaps in educational practice. *Journal of Black Studies* 43, 6, 693-718.

MISRA, I., ZITNICK, C.L., MITCHELL, M., AND GIRSHICK, R. 2016. Seeing through the human reporting bias: Visual classifiers from noisy humancentric labels. In *Conference on Computer Vision and Pattern Recognition*, Las Vegas, NV, 2930-2939.

MITCHELL, S., POTASH, E., AND BAROCAS, S. 2018. Prediction-based decisions and fairness: A catalogue of choices, assumptions and definitions. *arXiv preprint*, arXiv:1811.07867.

NAISMITH, B., HAN, N.-R., JUFFS, A., HILL, B., AND ZHENG, D. 2018. Accurate measurement of lexical sophistication with reference to ESL learning data. In K.E. Boyer, & M. Yudelson (Eds.) *Proceedings of the 11th International Conference on Educational Data Mining*, Buffalo, NY, 259-265.

NASIR, N.I.S., AND HAND, V.M. 2006. Exploring sociocultural perspective on race, culture, and learning. *Review of Educational Research* 76, 4, 449-475.

NGUYEN, H., AND LIEW, C.W. 2018. Using student logs to build Bayesian models of student knowledge and skills. In K.E. Boyer, & M. Yudelson (Eds.) *Proceedings of the 11th International Conference on Educational Data Mining*, Buffalo, NY, 312-317.

NIEVA, R. 2015. Google apologizes for algorithm mistakenly calling black people 'gorillas'. *CNet*, July 4, 2015.

NIŽNAN, J., PELÁNEK, R., AND RIHÁK, J. 2015. Student models for prior knowledge estimation. In O.C. Santos, J.G. Boticario, C. Romero, M. Pechenizkiy, A. Merceron, J.M. Luna, C. Mihaescu, P. Moreno, A. Hershkovitz, S. Ventura, & M. Desmarais (Eds.) *Proceedings of the 8th International Conference on Educational Data Mining*, Madrid, Spain, 109-116.

NOLL, R.B., ZELLER, M.H., VANNATTA, K., BUKOWSKI, W.M., AND DAVIES, W.H. 1997. Potential bias in classroom research: Comparison of children with permission and those who do not receive permission to participate. *Journal of Clinical Child Psychology* 26, 36-42.

OCUMPAUGH, J., BAKER, R., GOWDA, S., HEFFERNAN, N., AND HEFFERNAN, C. 2014. Population validity for Educational Data Mining models: A case study in affect detection. *British Journal of Educational Technology* 45, 3, 487-501.

PALAEZ, K., LEVINE, R., FAN, J., GUARCELLO, M., AND LAUMAKIS, M. 2019. Using a latent class forest to identify at-risk students in higher education. *Journal of Educational Data Mining* 11, 1, 18-46.

PARIS, S.G., AND BYRNES, J.P. 1989. The constructivist approach to self-regulation and learning in the classroom. In B.J. Zimmerman, D.H. Schunk (Eds.) *Self-Regulated Learning and Academic Achievement*, Springer, NY, 169-200.

PARK, J., YU, R., RODRIGUEZ, F., BAKER, R., SMYTH, P., AND WARSCHAUER, M. 2018. Understanding student procrastination via mixture models. In K.E. Boyer, & M. Yudelson (Eds.) *Proceedings of the 11th International Conference on Educational Data Mining*, Buffalo, NY, 187-197.

PITLARZ, I., PU, S., PATEL, M., AND PRABHU, R. 2018. What can we learn from college students' network transactions? Constructing useful features for student success prediction. In K.E. Boyer, & M. Yudelson (Eds.) *Proceedings of the 11th International Conference on Educational Data Mining*, Buffalo, NY, 444-448.

PINKARD, N. 2005. How the perceived masculinity and/or femininity of software applications influences students' software preferences. *Journal of Educational Computing Research* 32, 1, 57-78.

PEDRESCHI, D., RUGGIERI, AND S., TURINI, F. 2008. Discrimination-aware data mining. In *Proceedings of the 14th International Conference on Knowledge Discovery and Data Mining*, Las Vegas, Nevada, 560-568.

RAJKOMAR, A., HARDT, M., HOWELL, M.D., CORRADO, G., AND CHIN, M.H. 2018. Ensuring fairness in machine learning to advance health equity. *Annals of Internal Medicine* 169, 12, 866-872.

REN, Z., NING, X., LAN, A., AND RANGWALA, H. 2019. Grade prediction based on cumulative knowledge and co-taken courses. In C.F. Lynch, A. Merceron, M. Desmarais, & R. Nkambou (Eds.) *Proceedings of the 12th International Conference on Educational Data Mining*, Montreal, Canada, 158-167.

RIDDLE, T., BHAGAVATULA, S., GUO, W., MURESAN, S., COHEN, G., COOK, J., AND PURDIE-VAUGHNS, V. 2015. Mining a written values affirmation intervention to identify the unique linguistic features of stigmatized groups. In O.C. Santos, J.G. Boticario, C. Romero, M. Pechenizkiy, A. Merceron, J.M. Luna, C. Mihaescu, P. Moreno, A. Hershkovitz, S. Ventura, & M. Desmarais (Eds.) *Proceedings of the 8th International Conference on Educational Data Mining*, Madrid, Spain, 274-281.

ROMEI, A., AND RUGGIERI, S. 2014. A multidisciplinary survey on discrimination analysis. *Engineering Review* 29, 5, 582-638.

ROSENBAUM, P.R. 2001. Replicating effect and biases. *The American Statistician* 55, 3, 223-227.

ROTH, W.M., AND LEE, Y.J. 2007. "Vygotsky's neglected legacy": Cultural-historical activity theory. *Review of Educational Research* 77, 2, 186-232.

ROWE, E., ASBELL-CLARKE, J., EAGLE, M., HICKS, A., BARNES, T., BROWN, R. AND EDWARDS, T. 2016. Validating game-based measures of implicit science learning. In T. Barnes, M. Chi, & M. Feng (Eds.) *Proceedings of the 9th International Conference on Educational Data Mining*, Raleigh, NC, 490-495.

ROWE, E., BAKER, R., AND ASBELL-CLARKE, J. 2015. Strategic game moves mediate implicit science learning. In O.C. Santos, J.G. Boticario, C. Romero, M. Pechenizkiy, A. Merceron, J.M. Luna, C. Mihaescu, P. Moreno, A. Hershkovitz, S. Ventura, & M. Desmarais (Eds.) *Proceedings of the 8th International Conference on Educational Data Mining*, Madrid, Spain, 432-435.

RUGGIERI, S., PEDRESCHI, D., AND TURINI, F. 2010a. Data mining for discrimination discovery. *ACM Transactions on Knowledge Discovery from Data (TKDD)* 4, 2, 1-40.

RUGGIERI, S., PEDRESCHI, D., AND TURINI, F. 2010b. DCUBE: Discrimination discovery in databases. In *Proceedings of the 2010 SIGMOD International Conference on Management of Data*, Indianapolis, Indiana, 1127-1130.

SAARELA, M., AND KÄRKKÄINEN, T. 2015. Do country stereotypes exist in educational data? A clustering approach for large, sparse, and weighted data. In O.C. Santos, J.G. Boticario, C. Romero, M. Pechenizkiy, A. Merceron, J.M. Luna, C. Mihaescu, P. Moreno, A. Hershkovitz, S. Ventura, & M. Desmarais (Eds.) *Proceedings of the 8th International Conference on Educational Data Mining*, Madrid, Spain, 156-163.

SAMEI, B., OLNEY, A., KELLY, S., NYSTRAND, M., D'MELLO, S., BLANCHARD, N., AND GRAESSER, A.C. 2015. Modeling classroom discourse: Do models of predicting dialogic instruction properties generalize across populations? In O.C. Santos, J.G. Boticario, C. Romero, M. Pechenizkiy, A. Merceron, J.M. Luna, C. Mihaescu, P. Moreno, A. Hershkovitz, S. Ventura, & M. Desmarais (Eds.) *Proceedings of the 8th International Conference on Educational Data Mining*, Madrid, Spain, 444-447.

SANDVIG, C. 2015. Seeing the sort: The aesthetic and industrial defence of 'the algorithm'. *Media-N* 11, 1, 35-51.

SCHNEIDER, B., AND BLIKSTEIN, P. 2015. Unraveling students' interaction around a tangible interface using multimodal learning analytics. *Journal of Educational Data Mining* 7, 3, 89-116.

SHAFFER, I.R. 2018. Exploring the performance of facial expression recognition technologies on deaf adults and their children. In *Proceedings of the 20th International ACM SIGACCESS Conference on Computers and Accessibility*, Galway, Ireland, 474-476.

SIEMENS, G. 2005. Connectivism: Learning as network-creation. *ASTD Learning News* 10, 1, 1-28.

SIEMENS, G., AND BAKER, R.S. 2012. Learning analytics and educational data mining: Towards communication and collaboration. In *Proceedings of the 2nd International Conference on Learning Analytics and Knowledge*, Vancouver, Canada, 252-254.

SIMOIU, C., CORBETT-DAVIES, AND GOEL, S. 2017. The problem of infra-marginality in outcome tests for discrimination. *The Annals of Applied Statistics* 11, 3, 1193-1216.

SLIM, A., HUSH, D., OJAH, T., AND BABBITT, T. 2018. Predicting student enrollment based on student and college characteristics. In K.E. Boyer, & M. Yudelson (Eds.) *Proceedings of the 11th International Conference on Educational Data Mining*, Buffalo, NY, 383-389.

SPOON, K., BEEMER, J., WHITMER, J.C., FAN, J., FRAZEE, J.P., STRONACH, J., BOHONAK, A.J., AND LEVINE, R.A. 2016. Random forests for evaluating pedagogy and informing personalized learning. *Journal of Educational Data Mining* 8, 2, 20-50.

STOYANOVICH, J., HOWE, B., JAGADISH, H.V., AND MIKLAY, G. 2018. Panel: A debate on data and algorithmic ethics. *Proceedings of the VLDB Endowment* 11, 12, 2165-2167.

STRECHT, P., CRUZ, L., SOARES, C., MENDES-MOREIRA, J., AND ABREU, R. 2015. A comparative study of regression and classification algorithms for modelling students' academic performance. In *Proceedings of the 8th International Conference on Educational Data Mining*, Madrid, Spain, 392-395.

SWEENEY, L. 2013. Discrimination in online ad delivery. *Queue* 11, 3, 10-29.

SWEENEY, M., LESTER, J., RANGWALA, H., AND JOHRI, A. 2016. Next-term student performance prediction: A recommender systems approach. *Journal of Educational Data Mining* 8, 1, 22-51.

TATMAN, R. 2017. Gender and dialect bias in YouTube's automatic captions. In *Workshop on Ethics in Natural Language Processing volume 1*, Valencia, Spain, 53-59.

TODA, A.M., OLIVEIRA, W., SHI, L., BITTENCOURT, I.I., ISOTANI, S., AND CRISTEA, A. 2019. Planning gamification strategies based on user characteristics and D.M.: A gender-based case study. In C.F. Lynch, A. Merceron, M. Desmarais, & R. Nkambou (Eds.) *Proceedings of the 12th International Conference on Educational Data Mining*, Montreal, Canada, 438-443.

ÜLTANIR, E. 2012. An epistemological glance at the constructivist approach: Constructivist learning in Dewey, Piaget, and Montessori. *International Journal of Instruction* 5, 2, 195-212.

VAN MILTERNBURG, E. 2016. Stereotyping and bias in the flickr30k dataset. In J. Edlund, D. Heylen, & P. Paggio (Eds.) *Multimodal Corpora: Computer Vision and Language Processing (MMC 2016) Workshop*, 1-4.

VEALE, M., AND BINNS, R. 2017. Fairer machine learning in the real world: Mitigating discrimination without collecting sensitive data. *Big Data & Society* 4, 2, 1-17.

VERBERT, K., MANOUSELIS, N., DRACHSLER, H., AND DUVAL, E. 2012. Dataset-driven research to support learning and knowledge analytics. *Journal of Educational Technology & Society* 15, 3, 133-148.

VERGHESE, A., SHAH, N.H., AND HARRINGTON, R.A. 2018. What this computer needs is a physician: Humanism and artificial intelligence. *JAMA* 319, 1, 19-20.

WADSWORTH, B.J. 1996. *Piaget's theory of cognitive and affective development: Foundations of constructivism*. Longman Publishing.

WARNER, J., DOORENBOS, J., MILLER, B., AND GUO, P. 2015. How high school, college, and online students differentially engage with an interactive digital textbook. In O.C. Santos, J.G. Boticario, C. Romero, M. Pechenizkiy, A. Merceron, J.M. Luna, C. Mihaescu, P. Moreno, A. Hershkovitz, S. Ventura, & M. Desmarais (Eds.) *Proceedings of the 8th International Conference on Educational Data Mining*, Madrid, Spain, 528-531.

WELNER, K.G., AND CARTER, P.L. 2013. Achievement gaps arise from opportunity gaps. In P.L. Carter, & K.G. Welner (Eds.) *Closing the Opportunity Gap: What America Must Do to Give Every Child an Even Chance*, Oxford University Press, U.K., 1-10.

WILTZ, C. 2017. Bias in, bias out: How A.I. can become racist. https://www.designnews.com/bias-bias-out-how-ai-can-become-racist

WOLFF, A., ZDRAHAL, Z., NIKOLOV, A., AND PANTUCEK, M. 2013. Improving retention: Predicting at-risk students by analyzing clicking behaviour. In *Proceedings of the 3rd International Conference on Learning Analytics and Knowledge*, Leuven, Belgium, 145-149.

YANG, D., KRAUT, R., AND ROSE, C. 2016. Exploring the effect of student confusion in massive open online courses. *Journal of Educational Data Mining* 8, 1, 52-83.

YANG, K., AND STOVANOVICH, J. 2017. Measuring fairness in ranked outputs. In *Proceedings of the 29th International Conference on Scientific and Statistical Database Management*, Chicago, IL, 1-6.

YAO, S., AND HUANG, B. 2017. New fairness metrics for recommendation that embrace differences. In *Workshop on Fairness, Accountability and Transparency in Machine Learning (FATML)*, Halifax, Nova Scotia. arXiv:1706.09838.

ZADROZNY, W., BUDZIKOWSKA, M., CHAI, J., KAMBHATLA, N., LEVESQUE, S., AND NICOLOV, N. 2000. Natural language dialogue for personalized interaction. *Communications of the ACM* 43, 8, 116-120.

Zafar, M.B., Valera, I., Rodriguez, M.G., and Gummadi, K.P. 2015. Fairness constraints: Mechanisms for fair classification. In A. Singh, & J. Zhu (Eds.) *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*, 962-970.

Zafar, M.B., Valera, I., Rodriguez, M.G., and Gummadi, K.P. 2017. Fairness beyond disparate treatment & disparate impact: Learning classification without disparate mistreatment. In *Proceedings of the 26th International Conference on World Wide Web*, Perth, Australia, 1171-1180.

ZEMEL, R., WU, Y., SWERSKY, K., PITASSI, T., AND DWORK, C. 2013. Learning fair representations. In *Proceedings of the International Conference on Machine Learning*, Atlanta, GA, 325-333.

ZHENG, Z., VOGELSANG, T., AND PINKWART, N. 2015. The impact of small learning group composition on drop-out rate and learning performance in a MOOC. In O.C. Santos, J.G. Boticario, C. Romero, M. Pechenizkiy, A. Merceron, J.M. Luna, C. Mihaescu, P. Moreno, A. Hershkovitz, S. Ventura, & M. Desmarais (Eds.) *Proceedings of the 8th International Conference on Educational Data Mining*, Madrid, Spain, 500-503.

ZIEWITZ, M. 2016. Governing algorithms: Myth, mess, and methods. *Science, Technology, & Human Values* 41, 1, 3-16.

ZIMMERMANN, J., BRODERSEN, K.H., HEINIMANN, H.R., AND BUHMANN, J.M. 2015. A model-based approach to predicting graduate-level performance using indicators of undergraduate-level performance. *Journal of Educational Data Mining* 7, 3, 151-176.

ŽLIOBAITĖ, I. 2017. Measuring discrimination in algorithmic decision making. *Data Mining and Knowledge Discovery* 31, 4, 1060-1089.

ŽLIOBAITĖ, I, AND CUSTERS, B. 2016. Using sensitive personal data may be necessary for avoiding discrimination in datadriven decision models. *Artificial Intelligence and Law* 24, 2, 183-201.