Bias or Insufficient Sample Size? Improving Reliable Estimation of Algorithmic Bias for Minority Groups

Jaeyoon Choi
University of California, Irvine
Irvine, California, USA
jaeyoon.choi@uci.edu

Shamya Karumbaiah
University of Wisconsin-Madison
Madison, Wisconsin, USA
shamya.karumbaiah@wisc.edu

Jeffrey Matayoshi
McGraw Hill ALEKS
Irvine, California, USA
jeffrey.matayoshi@mheducation.com

Abstract

Despite the prevalent use of predictive models in learning analytics, several studies have demonstrated that these models can show disparate performance across different demographic groups of students. The first step to audit for and mitigate bias is to accurately estimate it. However, the current practice of identifying and measuring group bias faces reliability issues. In this paper, we use simulations and real-world data analysis to explore statistical factors that impact the reliability of bias estimation and suggest approaches to account for it. Our analysis revealed that small group sizes lead to high variability in group bias estimation due to sampling error – an issue that is more likely to impact students from historically marginalized communities. We then suggest statistical approaches, such as bootstrapping, to construct confidence intervals for a more reliable estimation of group bias. Based on our findings, we encourage future learning analytics research to ensure sufficiently large group sizes, construct confidence intervals, use at least two metrics, and move beyond the dichotomy of the presence or absence of bias for a more comprehensive evaluation of group bias.

CCS Concepts

• **Human-centered computing**; • **Social and professional topics** → *Computing / technology policy*; • **Applied computing** → **Education**;

Keywords

Algorithmic bias, Fairness, Group bias, Estimation of bias, Reliability, Predictive models

ACM Reference Format:
Jaeyoon Choi, Shamya Karumbaiah, and Jeffrey Matayoshi. 2025. Bias or Insufficient Sample Size? Improving Reliable Estimation of Algorithmic Bias for Minority Groups. In *LAK25: The 15th International Learning Analytics and Knowledge Conference (LAK 2025), March 03–07, 2025, Dublin, Ireland.* ACM, New York, NY, USA, 11 pages. https://doi.org/10.1145/3706468.3706540

1 Introduction

Predictive models have been at the forefront of learning analytics research and practice since its conception [34]. They are used for summative and formative assessment [43], predicting students' academic success [47], recognizing which students are likely to drop out of a course [13], and detecting disengagement in intelligent tutoring systems [7], learning games [21], and collaborative learning [33]. These high-performing predictive models enable timely interventions for students who need support.

However, the high performance of predictive models is not guaranteed for every student; in fact, predictive models often show disparate performance across different demographic groups, particularly performing worse for students from minoritized backgrounds. For example, studies in learning analytics have demonstrated that some predictive models are more likely to predict that students from historically marginalized groups, such as Black, Hispanic, and Native American students, will struggle or fail, even when they have succeeded [20]. In other words, predictive learning analytics are prone to *group bias* – that is, the models' predictive performance differs across different demographic groups of students.

Current approaches for estimating group bias in the field often involve simply calculating the differences in model performance across groups. For example, Zhang et al. [46] compared the performance of self-regulated learning detectors across racial and ethnic student groups and concluded that one of the detectors performed "somewhat" better for Hispanic/Latinx students (AUC = 0.81) compared to the white students (AUC = 0.80). But how do we determine whether a difference of 0.01 is significant to declare the presence or absence of bias? What is the threshold at which we should take action?

Moreover, given that historically marginalized groups typically have smaller sample sizes, the minority samples may not accurately represent the true population. This increases the likelihood of sampling error – the difference between the sample estimate and the true population parameter [37] – for minority groups. For instance, in Zhang et al. [46], there were only 18 Hispanic/Latinx students in the dataset, raising questions on how accurate the reported AUC of 0.81 and subsequently the positive bias of 0.01 is for the Hispanic/Latinx student group. Such sampling error occurs more often than not in learning analytics research [23] for reasons such as over-representation of Western, Educated, Industrialized, Rich, and Democratic (WEIRD) populations [36] or convenience sampling of undergraduate students [25].

Accurately estimating group bias in predictive models is crucial not only for an accurate bias audit but also to prevent mitigation efforts from further exacerbating bias for minority groups. Bias audits in a predictive model refer to the systematic evaluation of the model to identify potential group biases [35]. As discussed above, if the estimation of group bias is unreliable, the auditing process could result in spurious results. Hence, current mitigation

LAK 2025, Dublin, Ireland
© 2025 Copyright held by the owner/author(s).
ACM ISBN 979-8-4007-0701-8/25/03
https://doi.org/10.1145/3706468.3706540

approaches (e.g., [24]) cannot "mitigate" bias if the bias estimation is in itself inaccurate. Furthermore, if the auditing process wrongly concludes that the model is biased against the majority group due to inaccurate estimation of model performance on the minority group and attempts to "adjust" for bias, it could inadvertently introduce bias against minority students.

We argue that some of the current practices in auditing for or estimating group bias lack reliability. To the best of our knowledge, there is limited research on the factors that contribute to unreliable group bias estimation. To address this gap, our study examines two research questions:

- RQ1: What statistical factors influence the reliable estimation of bias in predictive learning analytics?
- RQ2: How can we make bias estimation more reliable with further statistical evidence?

To answer RQ1, we use simulations to explore how factors unrelated to bias, such as sample size, class distribution, error rate, and performance metric, affect the reliable estimation of group bias. Specifically, we introduce an equal amount of classification error to groups to demonstrate how group bias estimation can become significantly unreliable. Our findings reveal that even when a classifier is designed to perform equally well for two groups, smaller sample sizes can lead to either significantly inflated or deflated bias estimations. Furthermore, we analyze real-world data from a course dropout prediction model to answer RQ2. Specifically, we use Newcombe's Hybrid score method and bootstrapping to construct confidence intervals. This demonstrates that using confidence intervals can provide a more reliable estimate of group bias.

2 Background

2.1 What is Algorithmic Bias?

While the issue of algorithmic bias has gained prominence in recent years, defining it remains a complex task due to the term "bias" being used differently in statistics. Blodgett et al. [1] argue that clarification is needed in several areas, particularly in how bias is defined and what harms it can cause. In general, algorithmic bias in predictive models refers to the disparate predictive performance that algorithms may exhibit across different groups. Mitchell et al. [30] define bias as a model's unjustifiably differing predictive performance across disadvantaged groups. Other studies also have frequently used the term "unfair(ness)" in place of bias. For example, Mehrabi et al. [29] define fairness as the absence of prejudice or favoritism by a model toward certain individuals or groups based on their attributes. Therefore, algorithmic bias in predictive models is used interchangeably with group bias.

2.2 Current Methodological Approaches

Most of the algorithmic bias studies have focused on formalizing fairness, ideally in order to mitigate bias in different stages of the machine learning pipeline to achieve fairness. In general, these studies are based on group fairness: group fairness (sometimes referred to as statistical fairness such as in [10]) asks for parity of some statistical measure across all groups defined by a protected demographic category such as race or gender. For instance, an algorithm that predicts student dropout is deemed fair if the likelihood of being dropped out is approximately the same across different demographic groups (i.e., statistical parity, [16]). Other measures include false positive and false negative rates (i.e., equalized odds, [18]) and positive predictive value (similar to equalized calibration, [9]). Group fairness notion is relatively simple because it does not require making assumptions about the data.

2.3 Reliability Issues in Algorithmic Bias Studies

In this section, we discuss several challenges related to reliable estimation of algorithmic bias.

2.3.1 Small sample size. While algorithmic bias studies often focus more on model training and optimization, a significant issue in the field may also arise from the data collection process. Specifically, sample sizes for certain groups, particularly historically underrepresented populations, are often relatively small. This is largely due to the historical distrust of institutions by many marginalized communities, rooted in past exploitation and coercion, such as in the the Tuskegee Syphilis Study [2]. Additionally, socioeconomic factors such as poverty and limited access to transportation, and social stigma further complicate access to individuals from these groups [41]. For example, Mcmaster & Cook [11] highlights that the information about some intersectional groups is either not collected or the sample sizes are insufficient for meaningful analysis.

Statistically, small sample sizes make it difficult to capture the full variability within a population [32]. In other words, small sample sizes increase sampling error – the difference between the sample estimate and the true population parameter [37]. Varoquaux [44] demonstrated that small sample sizes in predictive models lead to large errors, compromising reliability of conclusions drawn from these models. Additionally, small sample sizes reduce the statistical power, which is the likelihood of detecting true effect. In other words, small sample sizes make it harder to detect meaningful effects [5, 42].

However, it is important to note that simply increasing data collection from marginalized populations is not a cure-all. Collecting more data could mean heightened surveillance and increased targeted exposure to marginalized populations [22].

2.3.2 Lack of statistically reliable practice to estimate group bias. Another significant challenge is the lack of reliable way to estimate group bias in predictive models. Currently, one of the most common approaches in learning analytics involves comparing the performance of a predictive model across different groups. For example, to assess whether a dropout prediction model is biased against black students, researchers typically compare the AUC for black students against that for white students or the overall population. Zambrano et al. [45] compared AUC of bayesian knowledge tracing models for intersectional student groups and concluded that their models did not show any particular bias against any population. This is because the maximum AUC difference within intersectional groups was 0.033.

However, how can we determine if the AUC difference of 0.033 is statistically significant enough to declare the presence or absence of bias? As discussed above, small sample sizes increase the likelihood of sampling error, ultimately compromising the reliability of the results from predictive models [44]. What if this 0.033 difference resulted from the lack of reliability in the group bias estimation? Without investigating factors that could influence the estimation of group bias, such as sample size, it is challenging to draw definitive conclusions from a single value.

3 Statistical Factors That May Affect The Reliable Estimation of Group Bias

In this section, we synthesize empirical findings from existing literature to explore several factors that may influence the reliability of group bias estimation. That is, we examine factors that in theory do not introduce or mitigate group bias within the machine learning pipelines (particularly during training), but may affect its estimation. Specifically, we examine four factors: group sample size, class distribution, error rate, and performance metric. Ideally, these factors should not influence how biased an algorithm is. For instance, while biased observations by data annotators may lead to algorithmic bias through the training process, we would not expect there to be higher bias just because a dataset has higher positive labels overall. Likewise, while underrepresentation of a group may contribute to bias, the sampling error could also deflate the bias estimation (e.g., when the small group sample is not representative of the variability in the group). Therefore, the factors we analyze here are related to the *reliable measurement* of group bias, not the biases introduced during data collection, training or deployment.

3.1 Group Sample Size

Group sample size refers to the number of instances that belong to the group of interest, that is, it is the *sample size* of the group. Since sample size is the number of instances included in the sample which is drawn from a population, group sample size is the number of the samples drawn from the population of a certain group.

Previous literature has emphasized that sufficient sample size is required to accurately measure or detect true effects [27, 39]. Small sample sizes are often inadequate to fully represent the variability found in a population [32]. As sample size decreases, the likelihood of the sample accurately estimating the true population decreases, leading to increased sampling error. Sampling error refers to the difference between the sample estimate and the population parameter [37]. This is because a small sample has a higher likelihood of being an "unusual" sample of the true population. For instance, if we sample only five students from a classroom, those five students might have significantly different levels of self-regulation, motivation, or prior learning backgrounds. Consequently, it becomes challenging to draw conclusive decisions based on such a small, and potentially un-representative sample.

This implies that small sample sizes can impact the reliable measurement of group performance. Other things being equal, the model performance for a group could be either overestimated or underestimated if the sample size is very small. Furthermore, a small sample size reduces the statistical power, which is the likelihood that a hypothesis test can detect a difference (or relationship) when a true difference (or relationship) exists in the population [42] [1].

This is particularly problematic when exploring algorithmic bias, as historically minoritized groups often have fewer data points (i.e., smaller group size). For instance, in Zambrano et al. [45], there are only four Native American students and 14 Native Hawaiian and Pacific Islander students, compared to 831 white students. Therefore, minoritized groups with smaller group sizes are more likely to have their performance less accurately measured.

3.2 Class Distribution

A class label in classification refers to the category or outcome that a single data point belongs to. In binary classification, there are two class labels: positive (e.g., dropout) and negative (e.g., not dropout). Class distribution in binary classification hence refers to the proportion of positive instances within the given data [17, 19]. If the dataset contains N instances, and N_p instances belong to the positive class, the class distribution is $\frac{N_p}{N}$. A dataset is considered imbalanced when the proportion of positive and negative instances differs significantly.

In practice, much of the data is highly imbalanced, hence class distribution is an important factor to consider in binary classification settings. According to Dablain et al. [12], a lack of balance in class distribution could make the classifiers more biased toward majority class, as the algorithm's parameters are heavily weighted toward more frequently occurring examples during training. However, in this paper, we focus on how the class distribution impacts the reliable estimation of group bias in the evaluation of the model training, validation, and prediction at the deployment process, not how the bias from class distribution is introduced in the training process itself.

Jeni et al. [19] explored how class distribution impacts performance measurement of facial recognition algorithms, particularly with respect to performance metrics. They found that commonly used metrics, such as Accuracy and Area Under the Precision-Recall Curve, are affected by imbalanced class distribution, whereas the Area Under the ROC curve (AUC) is the only exception. Similarly, studies such as Chicco et al. [8] have shown that both the F1 score and Accuracy can be overly inflated with imbalanced data. Therefore, these findings all imply that class distribution could impact the reliable estimation of a classifier's performance, and this could influence the identification and measurement of group bias.

3.3 Error Rate

The amount of classification error determines the classifier's performance on a group. We define this as *error rate* – the fraction of instances that are misclassified. When the error rate differs across groups, we consider it as an evidence for group bias. For instance, an algorithm developed to predict UK students' exam grades in 2020 assigned lower grades to students in public school compared to those in private schools [40], and this becomes the evidence of group bias in the algorithm.

We would expect a classifier's performance to get worse for all groups as the error rate increases (while the extent to which it worsens is dependent on the metrics; see [26]). However, the error

[1] We also acknowledge that the group sample size can impact the model training process and lead to bias. That is, having a relatively small sample size for a certain demographic group (e.g., female students have been underrepresented in STEM courses with only 20% of STEM MOOC learners being female [38]) can cause underrepresentation for that demographic group to be insufficiently trained by the machine learning model. While also relevant to studying bias, this is not the focus of this paper.

rate in and of itself should not affect bias. That is, introducing the same error rate would not lead to different classification performance in different groups. Hence, as long as the same error rate is applied to the groups, varying error rates should not influence bias measurement.

3.4 Metric

Numerous performance metrics are used in machine learning and learning analytics for binary classification, such as Accuracy, precision, recall, F1 score, and AUC. Furthermore, in algorithmic bias literature, metrics that measure the error – such as False Positive Rate (FPR) – are widely used to examine whether the error is equally distributed across different demographic groups or not.

While the choice of metric does not affect the performance of predictive models in and of itself, metrics do emphasize different aspects of the model performance. For instance, let us compare the formula for Precision and Recall:

$$Precision = \frac{TP}{TP + FP} \quad \text{and} \quad Recall = \frac{TP}{TP + FN}$$

where TP refers to the True Positives, FP refers to the False Positives, and FN refers to the False Negatives.

Precision measures how many of the instances predicted as positive are actually positive. High precision means fewer false positives – thus, precision is used when having high false positives are problematic. On the other hand, Recall measures how a model correctly identifies positives among the actual positive instances. Unlike precision, recall is used when missing positive cases (i.e., false negatives) is costly. Generally, there is a tradeoff between precision and recall – increasing one often leads to a decrease in the other. As in this case, metrics prioritize different aspects of model performance.

In fact, Jeni et al. [19] shows that some performance metrics (such as F1, which is a harmonic mean of Precision and Recall) may reveal more about class distribution than they do about actual performance. Similarly, Kwegyir-Aggrey et al. [26] argues that AUC is dependent on class distribution and misclassification errors. Therefore, the decision to use which metrics could impact the understanding of a classifier's performance.

Furthermore, bias can be defined using these metrics. That is, group bias can be defined by comparing the metrics of two groups. For example, Kearns et al. [24] used equalizing False Positive Rate between the overall population and a target group as a bias metric. That is, the difference between the performance of the group and that of the overall population serves as the evidence of differential treatment by the classifier, which constitutes the definition of bias.

Factor	Description
Group sample size	Sample size of the group
Class distribution	Positive label frequency
Error	Classification error
Metric	Evaluation metric for performance (e.g., FPR)

Table 1: Factors affecting reliable estimation of group bias.

4 Simulation

We investigate whether the statistical factors described in Section 3 influence the reliable estimation of group bias. In this simulation, we introduce an equal amount of error to each group, so that the simulated classifiers by design perform equally across the groups. Therefore, if any performance differences between the groups are observed, this may provide evidence of how these statistical factors could inflate or deflate the estimation of group bias. In other words, any observed performance differences would indicate a lack of reliability in bias estimation.

While examining these factors in real-world datasets would be ideal, it is often impractical to collect datasets that account for all possible variables. For reasons described above, collecting data from historically marginalized groups may be particularly challenging. Therefore, in this section we conduct simulations varying these factors to different degrees and investigate how they influence the measurement of bias (note that we also conduct a real data analysis in Section 5).

4.1 Simulation Setup

Using simulations, we address RQ1: What statistical factors influence the reliable estimation of bias in predictive learning analytics? Specifically, to simulate the classification process, we assume two datasets exist: A : *actual* set (true labels) and P: *prediction* set (predicted labels by classifier), where each set consists of 0 (negative) and 1 (positive) instances. Note that the *actual* set and *prediction* set have the same size. In this simulation, we examine two different dataset sizes: (1)$|A| = |P| = 1000$ (when total dataset size is 1,000), and (2) $|A| = |P| = 10000$ (when total dataset size is 10,000). Furthermore, we assume that there exists a group of interest (hereafter referred to as *target*), with all other groups in the dataset collectively referred to as *others*. Note that *target* and *others* are mutually exclusive in this simulation. While in practice target and others may share attributes (i.e., intersectionality; for instance, 'black' and 'black female' groups), we assume mutual exclusivity in this paper for simplicity. Therefore, an *actual* set A and a *prediction* set P could be defined as such:

$$A = A_{Target} \cup A_{Others} \quad \text{and} \quad P = P_{Target} \cup P_{Others}$$

And we introduce different factors discussed in Section 3.

4.1.1 Group sample size. We set the group sample size for each group as follows.

- When total dataset size is 1,000 ($|A| = |P| = 1000$): *Target* group size ($|A_{Target}| = |P_{Target}|$) = 10, 20, 50, 100, 200, 300, 400 (The group size for *Others* = 1,000 - target group size).
- When total dataset size is 10,000($|A| = |P| = 10000$): *Target* group size ($|A_{Target}| = |P_{Target}|$) = 10, 20, 50, 100, 500, 1,000, 2,000, 3,000, 4,000 (The group size for *Others* = 10,000 - *target* group size).

We chose to start at 10 for *target* group size as it is used as a threshold to filter groups in some algorithmic bias studies (e.g., [45]).

4.1.2 Class distribution. In this simulation, class distribution refers to the proportion of positive instances in the *actual* set A. Hence, class distribution of each group is the proportion of positive instances in A_{Target} or A_{Others}. We investigate five different class distributions for each group: 0.05, 0.1, 0.2, 0.3, and 0.4. For instance, when a *target* group's class distribution is 0.1, that means that

10% of A_{Target} consists of positive instances (i.e., 1). Based on the class distribution, we can simulate the actual labels for A_{Target} and A_{Others}.

4.1.3 Error rate. Error rate refers to the proportion of misclassified instances. Hence, the error rate of a group is the proportion of disagreement between its actual set values and prediction set values (e.g., error rate of *target* group = disagreement between A_{Target} and P_{Target}). Based on this, we can generate each group's prediction set based on its actual set with a certain error rate. Specifically, to operationalize the process of introducing error rate k, we flip each instance (0 to 1, 1 to 0) in the group's actual set with the probability of k.

Although groups in practice may have different error rates in classification tasks (e.g., darker skinned faces are more likely to be misclassified in facial recognition algorithms [4]), in this simulation we apply the same error rate to both the *target* and *others* group. This is because the current study aims to examine the factors that affect the reliable measurement of bias, especially those statistical factors that do not contribute to the origin or mitigation of biases in predictive models. Since the simulated classifiers are designed to perform equally across groups (i.e., equal error rates), any group differences observed later implies that factors that are not related to bias could influence the bias estimation.

4.1.4 Metric. In this simulation, we use False Positive Rate (FPR) and False Negative Rate (FNR), commonly used metrics in fairness research across learning analytics and machine learning literature [9, 24]. FPR is the proportion of negative instances (e.g., students who did not complete the course) that are incorrectly identified as positive instances (e.g., students who completed the course) (i.e., $FPR = \frac{FP}{FP+TN}$). On the other hand, FNR represents the proportion of positive instances that are incorrectly identified as negative instances (i.e., $FNR = \frac{FN}{FN+TP}$).

Furthermore, we define bias as such:

$$FPR_{Diff} = FPR_{Target} - FPR_{Others}$$
$$FNR_{Diff} = FNR_{Target} - FNR_{Others}$$

That is, bias is defined as the difference in a given metric between the *target* group and *others*. A positive FPR_{Diff} indicates greater bias against the *target* group, meaning the classifier is more likely to incorrectly classify actual negatives as positives for the *target* group compared to the *others*. Similarly, a positive FNR_{Diff} implies that the result is more biased against the *target* group. While we examine other metrics such as precision, recall, and F1 in our simulation, due to space limitations, we mostly focus on FPR_{Diff} here; however, the results for other metrics show similar trends.

4.2 Experiment Design

A simulation refers to the unique combination of the following factors:

- Total dataset size (=$|A| = |P|$) e.g., $|A| = |P| = 1000$
- *Target* group size (=$|A_{Target}| = |P_{Target}|$) e.g., $|A_{Target}| = 10$. Then $|A_{Others}| = 990$.
- Class distribution for A_{Target} e.g., 0.1
- Class distribution for A_{Others} e.g., 0.2
- Error rate e.g., 0.1

To quantify the variability in our experiment, we repeat the simulation of each unique combination of factors for 100 times. And in each simulation, we compute the FPR of the *target* group and *others*. Based on FPR_{Target} and FPR_{Others}, we can compute FPR_{Diff}. Lastly, using the FPR_{Diff} values from 100 simulations, we compute the 5th and 95th percentiles to show the variability of the FPR_{Diff}.

As described in 4.1.3, we introduce an equal amount of error to both the *target* group and *others* to ensure that the simulation process itself does not introduce any bias. Therefore, theoretically the results for the *target* group and *others* should be approximately the same – specifically, the FPR_{Diff} between the two should be approximately zero. If this is not observed, it suggests that factors unrelated to bias may be affecting the measurement of group bias. For example, if the FPR_{Diff} deviates significantly from zero in cases where the group size is relatively small, this indicates that group size may impact group bias measurement.

4.3 Simulation Results

4.3.1 Finding 1: Smaller group sizes result in high variability. Fig.1 shows the 5th and 95th percentiles for FPR_{Diff} ($FPR_{Target} - FPR_{Others}$), with each dot representing the median. In other words, each bar covers the middle 90% of the data. The blue horizontal line represents $FPR_{Diff} = 0$, which indicates that there exists no group bias. The left plot corresponds to a scenario with a total data size of 1,000, a class distribution of 0.3, and an error rate of 0.1, while the right plot represents a scenario with a total data size of 10,000, maintaining the same class distribution and error rate. For instance, the leftmost intervals in both plots are the 5th and 95th percentiles of the *target* group of size 10.

In both plots, we observe that the smaller the group size, the wider the interval between the 5th and 95th percentiles. In other words, smaller group sizes lead to greater variability in FPR_{Diff}. For example, in the left plot, when the *target* group size is 10, the interval ranges from -0.115 to 0.188. If a study were to report only a single value, it would be equivalent to selecting one point from this wide range. Given that the interval spans both negative (i.e., more biased against *Others*) and positive values (i.e., more biased against *Target*), this could easily lead to a misinterpretation of bias if only a single value is reported.

Therefore, we argue that this variability could have been mistakenly referred to as an evidence of group bias in previous studies, specifically when any types of intervals to examine variability are not considered and when the group size is smaller. That is, if we were to only compare the single value of FPR_{Diff}, we could wrongly conclude that the classifier exhibits bias against certain groups. As mentioned earlier, when group size is 10, it is possible to either argue that the classifier is biased against *target* group and biased against *others*.

4.3.2 Finding 2: Class distribution and error rate do not impact the variability significantly. Fig. 2 shows the 5th and 95th percentiles for for FPR_{Diff} ($FPR_{Target} - FPR_{Others}$), across different class distributions and error rates, with a total dataset size of 10,000. The top two plots have the same class distribution of 0.3 but differ in error rates – 0.1 (lowest) and 0.9 (highest). The bottom two plots

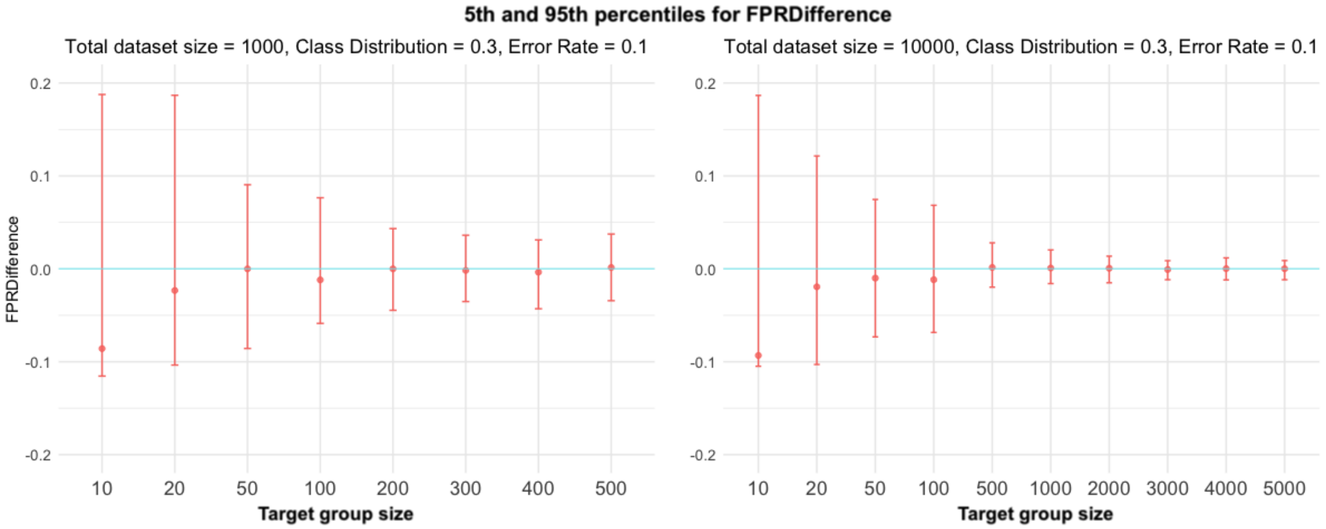

Figure 1: 5th and 95th percentile intervals for FPR_{Diff} ($FPR_{Target} - FPR_{Others}$) with varying *target* group sizes. The left plot represents a total dataset size of 1,000, while the right plot represents a dataset size of 10,000. Each bar covers from the 5th to 95th percentiles , with a dot representing the median value. The horizontal blue line indicates an FPR_{Diff} of 0 ($FPR_{Target} = FPR_{Others}$).

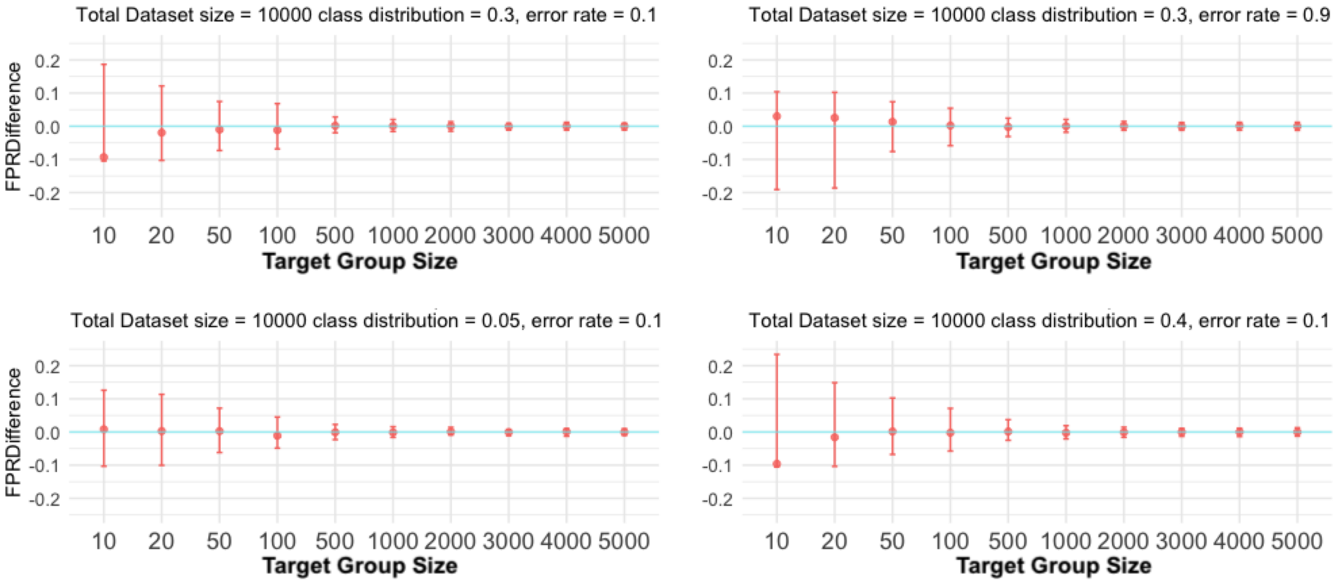

Figure 2: 5th and 95th percentile intervals for FPR_{Diff} ($FPR_{Target} - FPR_{Others}$) with varying *target* group sizes across different class distributions and error rates, with a total dataset size of 10,000. The top two plots have same class distribution of 0.3, but each has different error rate (0.1 and 0.9). The bottom two plots have same error rate of 0.1, but each has different class distribution (0.05 and 0.4). The horizontal blue line indicates an FPR_{Diff} of 0 (i.e., $FPR_{Target} = FPR_{Others}$).

show intervals for the same error rate of 0.1 but differ in class distributions – 0.05 (lowest) and 0.4 (highest).

The overall pattern observed in Fig. 2 is consistent with finding 1, showing that the smaller group sizes result in high variability, even with varying class distribution and error rate. That is, different class distribution and error rate did not significantly impact the variability of bias measurement, and there was no interaction effect observed . For instance, when group size is 10, the percentile intervals from Fig. 2 are as following: (-0.105, 0.187) for class distribution = 0.3, error rate = 0.1; (-0.191, 0.104) for class distribution = 0.3, error rate = 0.9; (-0.103, 0.126) for class distribution = 0.05, error rate = 0.1; and (-0.105, 0.234) for class distribution = 0.4, error rate = 0.1. The lengths of those intervals are approximately 0.3 for all cases. However, when the group size is 500, the intervals become shorter: (-0.019, 0.028) for class distribution = 0.3, error rate = 0.1; (-0.031, 0.024) for class distribution = 0.3, error rate = 0.9; (-0.023, 0.028) for class distribution = 0.05, error rate = 0.1; and (-0.024, 0.037) for class distribution = 0.4, error rate = 0.1. Now the lengths of the intervals are around 0.05. And this pattern is also observed when the total dataset size is 1,000.

4.3.3 Finding 3: Reliability issue in bias estimation persists across different metrics. Fig. 3 shows the 5th and 95th percentile intervals for FNR_{Diff} with a total dataset size of 10,000 with varying class distributions and error rates. Consistent with Finding 1, FNR_{Diff} exhibits a similar trend: smaller group sizes lead to higher variability in bias estimates. Additionally, we found that the intervals for FNR_{Diff} become exceptionally wide when the class distribution is highly imbalanced (0.05) and the *target* group size is extremely small (10 or 20), as illustrated in the bottom-leftmost plot of Fig 3. For instance, when the target group size is 10 and the class distribution is 0.05, the interval spans from -0.12 to 0.902. This suggests that when the group size is small and positives are scarce, FNR_{Diff} can be substantially inflated.

We also observed a similar pattern of unreliable bias estimation with smaller group sizes across other metrics, including precision, recall, and F1 score. The simulation in this paper focused solely on proportion-based metrics; therefore, we did not examine AUC, a threshold-independent metric, and plan to investigate it in the future research. However, we did evaluate AUC in the real data analysis in Section 5, and based on these results, we anticipate that AUC would exhibit similar patterns.

5 Real-World Data Analysis with Bootstrapping

In this section, we address RQ2: How can we make bias estimation more reliable with further statistical evidence? To answer this question, we use a real-world dataset and build a machine learning model to predict student success/dropout. Note that this section is for illustrative purposes – that is, optimizing the machine learning models for best performance is beyond the scope of this paper.

5.1 Data

We chose a dataset named *Predict Students' Dropout and Academic Success* from the UC Irvine Machine Learning Repository [2]. This dataset was collected to develop a machine learning model that predicts academic success and failure in higher education [28]. The data were originally collected from 4,424 students but we used 3,630 students' information to binarize the output variable (graduate vs. dropout). Among 36 features, we use a binary variable gender to construct groups (female and male)[3], and selected 7 variables (previous qualification, grade from previous qualification, mother job, father job, admission grade, educational special needs, scholarship holder) as predictors based on [28] to predict students' success (i.e., graduation).

Among 3,630 students, around 60% of the students graduated while 40% dropped out. In other words, the class distribution is 0.6. Specifically, the data consists of 2,381 female students (~65%) and 1,249 male students (~35%). Among female students, around 70% of students graduated while 30% of them dropped out (i.e., the class distribution for female students was 0.7). Among male students, around 44% of students graduated while 56% of them dropped out (i.e., the class distribution for male students was 0.44).

5.2 Modeling

We first built a logistic regression model that predicts students' graduation using the 7 variables mentioned above. We split the data into a training set (70%) and a test set (30%), and used stratified sampling to make sure each set maintains a similar proportion of gender. Because optimization is not the purpose of this paper, we skipped the hyperparameter tuning or the feature selection process. For evaluation, we selected FPR and AUC as metrics, representing proportion-based and threshold-independent measures respectively. The baseline model performance was FPR = 0.64, and AUC = 0.72. For given groups divided by gender variable, the female group had AUC of 0.722 and FPR of 0.679. For the male group, AUC was 0.69, and FPR was 0.616.

5.3 Bias Analysis

To construct confidence intervals for bias analysis, we propose two approaches based on whether the performance metric is based on proportions (e.g., FPR) or not (e.g., AUC):

(1) **Confidence Interval for FPR_{Diff}: Newcombe's Hybrid Score Method**
We use Newcombe's hybrid score method [31] to construct a confidence interval for FPR_{Diff}. Newcombe's hybrid method is designed to estimate the difference between two binomial proportions, and is known for its robust performance even with small sample sizes (See [3, 15, 31] for a detailed explanation of Newcombe's method). Since FPR is a binomial proportion ($FPR = \frac{FP}{FP+TN}$) and our bias metric is the difference between two groups, this method is well-suited for our analysis. Note that Newcombe's hybrid score method makes confidence intervals without bootstrapping. That is, we compute the confidence interval of the computed FPR_{Diff} between female and male groups. If the confidence interval for FPR_{Diff} includes 0, we cannot reject the null hypothesis that the FPRs for the female and male groups are the same. Conversely, if the confidence interval does not include 0, it suggests a statistically significant difference in FPR between the two groups.
However, Newcombe's hybrid score method is only used for constructing confidence intervals with two binomial proportions. Therefore, non-proportion based metrics like AUC could not use this method to construct confidence intervals. To construct confidence intervals for AUC, we use bootstrapping.

(2) **Confidence Interval for AUC_{Diff}: Bootstrapping (Bias-Corrected and Accelerated Bootstrap interval method)**

[2] https://archive.ics.uci.edu/dataset/697/predict+students+dropout+and+academic+success

[3] While gender is understood as a spectrum, the dataset used in this study classifies gender as a binary variable. This limitation reflects the structure of the data and not our perspective on gender.

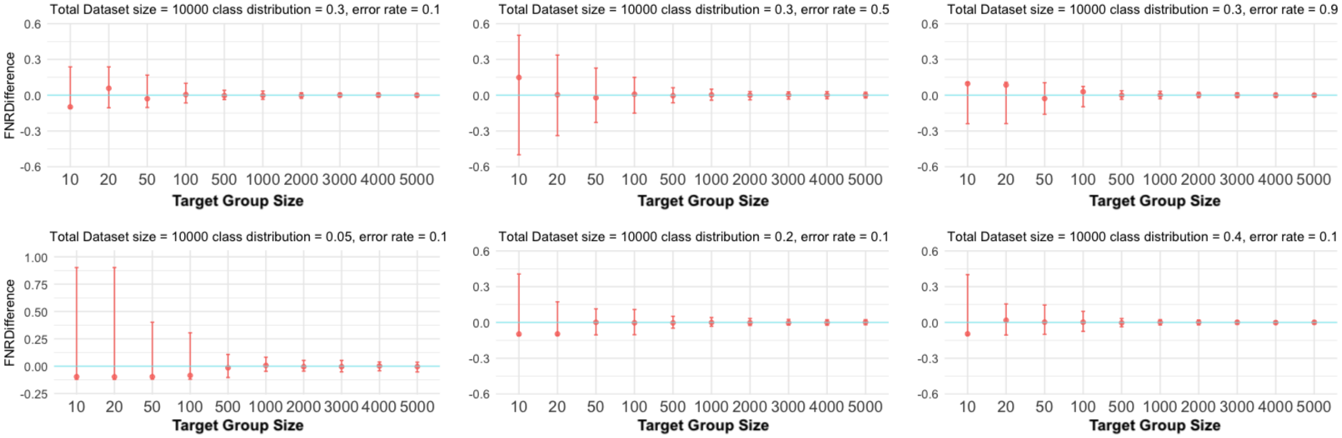

Figure 3: 5th and 95th percentile intervals for FNR_{Diff}, with total dataset size 10000 and varying class distributions and error rates. Note that the bottom-leftmost plot uses a different y-axis scale for clarity.

After completing the modeling process, we conducted a bootstrapping analysis to construct a confidence interval for AUC_{Diff}. Bootstrapping is a statistical technique to estimate the distribution of sample statistics by repeatedly resampling with replacement from the original data [6]. By using bootstrapping, we can repeatedly resample small amounts of data points multiple times and construct confidence intervals empirically for metrics like AUC, which is not a binomial proportion.

We conduct bootstrapping by resampling the data with replacement for 10,000 iterations. The size of each resample is equal to the size of the original dataset, which is 3,630. For each of the resamples, we compute the AUC for both female and male groups, and then calculate the AUC_{Diff}. Based on the bootstrapped results, we use the Bias-Corrected and Accelerated (BCa) bootstrap interval method to construct a confidence interval for AUC_{Diff}, which adjusts for both bias and skewness in the bootstrap distribution [14].

5.4 Results

The top interval in Fig. 4 is the 95% confidence interval for the FPR_{Diff} and is computed using the Newcombe Hybrid score method, with the red dot indicating the computed FPR_{Diff} of 0.063. The confidence interval is (-0.028,0.153), including 0. Therefore, we cannot reject the null hypothesis that the two groups have the same FPR.

The bottom interval in Fig. 4 is the 95% BCa confidence interval for the AUC_{Diff} and is computed using bootstrapping, with the red dot representing the computed AUC_{Diff} of 0.032. The confidence interval is (-0.004, 0.069), also including 0. Therefore, similar to the FPR_{Diff}, we cannot reject the null hypothesis that the two groups have the same AUC.

Whether using the Newcombe hybrid score method or bootstrapping, constructing confidence intervals in this way can help interpret observed group differences at hand. For instance, suppose we only examine the computed difference values ($FPRDiff$ = 0.063, $AUCDiff$ = 0.032). We would lack statistical evidence to determine whether these computed differences are meaningful. Worse, these values could be misinterpreted as evidence of group bias. Confidence interval shows a range of values for bias estimation, and helps us understand if the computed group differences at hand could indicate group bias or not. Hence, estimating bias becomes more reliable with confidence intervals.

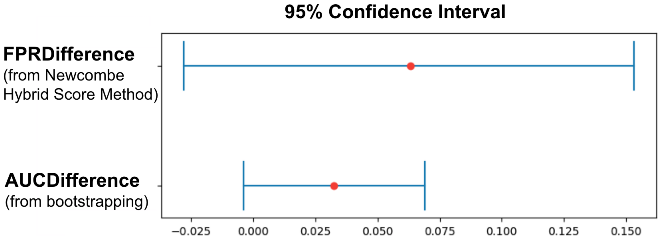

Figure 4: 95% Confidence Intervals for FPR_{Diff} and AUC_{Diff}. FPR_{Diff} is computed using the Newcombe Hybrid score method, while AUC_{Diff} is computed using bootstrapping and BCa. Each red dot represents the computed model performance difference between female and male.

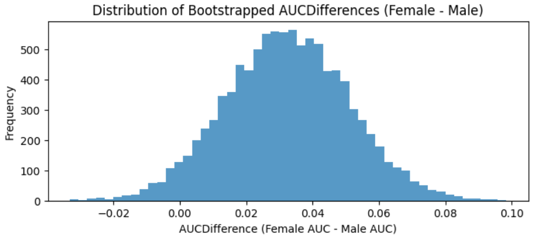

Figure 5: Distribution of bootstrapped AUC_{Diff}s between female and male groups.

6 Discussion

6.1 Summary of Findings

In this paper, we examine two research questions: 1) What statistical factors influence the reliable estimation of bias in predictive learning analytics, and 2) How can we make bias estimation more reliable with further statistical evidence? We first conducted a simulation study to explore how group sample size, class distribution, error rate, and metrics – those that in theory do not introduce or mitigate bias within the machine learning pipelines – could affect the reliable estimation of group bias, while ensuring that an equal amount of classification error was applied to each group. We found that group sample size affects the estimation of group bias. Specifically, smaller group sizes lead to high variability in bias estimation, making the bias estimation process unreliable.

Furthermore, we used real-world data and applied Newcombe's Hybrid score method and bootstrapping to construct confidence intervals for FPR Difference and AUC Difference. Our findings show that while the computed group differences might indicate group bias, the confidence intervals implied that we cannot reject the null hypothesis that the model performances for the two groups are the same.

Therefore, this paper demonstrates that the current approaches of simply computing differences in model performance across groups lack reliability. In the worst-case scenario, the reported "bias" is not actually group bias, but a result of insufficient samples – that is, the computed difference arises from sampling error. Therefore, we need to make the bias estimation process more reliable with further statistical evidence, such as using confidence intervals.

6.2 Implications to Research and Practice

Based on the findings from sections 4 and 5, we make some recommendations for learning analytics researchers and practitioners who study algorithmic bias.

6.2.1 Use a large enough group size. In section 4, we showed that small group sizes lead to high variability in group bias estimation. This indicates that the bias estimation, and any interpretations based on it, would be unreliable. In the worst case, the observed bias could simply result from insufficient sample sizes. That is, the variability caused by sampling error may have been mistakenly interpreted as bias in previous literature.

Therefore, it is crucial to ensure that group sizes are large enough when estimating group bias. This is not a new idea – in statistics, having a large enough sample size is essential for making accurate estimates about the population [27, 32, 39]. Furthermore, small group sizes reduce statistical power, making it harder to detect true effects. Therefore, it is important to use a large enough group size to ensure accurate identification and measurement of group bias.

However, we also acknowledge that collecting more data points, particularly for historically marginalized groups, is not an easy task. As Karumbaiah et al. [22] discussed, collecting more data for marginalized people could come at the cost of increased surveillance and compromised privacy and individual agency. Hence, if it is impossible to collect more samples, we suggest a bootstrapping method (see section 5). Even when you have a small number of samples collected, you can resample with replacement and construct a confidence interval to help identify if the observed bias with the samples at hand is reliable.

6.2.2 Construct confidence intervals for reliability. To the best of our knowledge, algorithmic bias audits do not report confidence intervals. Instead, a common practice is reporting a single value (e.g., the FPR difference between a *target* group and *others* is 0.01). However, as discussed in section 4, there exists sampling error in bias measurement, particularly when the group size is small. Hence, the current practice lacks reliability: without reliability, it is impossible to declare the presence or absence of group bias, or make any conclusive statements about bias. This inaccurate estimation of group bias not only undermines the accuracy of bias auditing but could also complicate the bias mitigation process, potentially harming minority groups.

Hence, reporting group bias should go beyond merely reporting single difference values and instead compute confidence intervals to quantify the uncertainty in the estimated values, and if desired, also test for statistical significance.

Specifically, we suggest constructing confidence intervals as follows.

- Metrics based on proportion (e.g., FPR, FNR): You could use the Newcombe Hybrid Score method to construct confidence intervals when you compute two independent groups' differences.
- Metrics that are not based on proportion (e.g., AUC): Conduct bootstrapping (resample with replacement your collected samples for multiple iterations). Then construct BCa confidence intervals.

6.2.3 Examining at least two metrics for comprehensive bias analysis. Furthermore, we recommend estimating group bias using more than one performance metric. This recommendation arises from our observation from the simulation that FNR Difference can be inflated when the sample size is small and actual positives are scarce. Therefore, we advise using at least two different performance metrics to ensure a comprehensive evaluation and to verify whether they result in consistent conclusions.

6.2.4 Beyond the false dichotomy of the "presence or absence of bias.". Our study also emphasizes the need to move beyond the goal of declaring "presence or absence of group bias" based on computed group differences. We argue that no non-trivial algorithm in education that is useful is entirely free of bias – societal biases can be introduced at every step of the design and deployment of predictive models. Prematurely declaring absence of bias when there are insufficient samples in minority groups raises additional concerns on validity. Hence, we recommend that methodological research on group bias should focus instead on understanding "how much" an algorithm is biased against certain demographic groups.

6.3 Limitations

While this study investigates factors affecting the reliability of group bias estimation and explores how to make bias estimation more reliable with further statistical evidence, a few limitations should be considered for future research. First, we assume the presence of only two mutually exclusive groups in both the simulation and real-world analysis. However, in practice, groups may share attributes, leading to intersectional groups within the data (e.g., black and black female). Therefore, it would be valuable to explore how intersectional bias measurement might be influenced by group size and other relevant factors. Additionally, our simulation only used proportion based metrics such as FPR and FNR for simplicity. Using other non-proportion based metrics such as AUC could provide a more comprehensive evaluation of group bias.

In conclusion, our study demonstrated that the current practices for identifying and estimating group bias in predictive learning analytics face significant reliability issues, likely due to sampling error in minority groups. To address this, we recommend approaches such as constructing confidence interval with Newcombe Hybrid Score or bootstrapping. Improving methods used for bias research is crucial to prevent our efforts from further exacerbating bias for minority groups.

Acknowledgments

Support for this research was provided by the Office of the Vice Chancellor for Research and Graduate Education at the University of Wisconsin–Madison with funding from the Wisconsin Alumni Research Foundation. We also appreciate Daniel Bolt for his valuable feedback on the methods used in this study.

References

[1] Su Lin Blodgett, Solon Barocas, Hal Daumé III, and Hanna Wallach. 2020. Language (technology) is power: A critical survey of" bias" in nlp. *arXiv preprint arXiv:2005.14050* (2020).

[2] Allan M Brandt. 1978. Racism and research: the case of the Tuskegee Syphilis Study. *Hastings center report* (1978), 21–29.

[3] Lawrence Brown and Xuefeng Li. 2005. Confidence intervals for two sample binomial distribution. *Journal of Statistical Planning and Inference* 130, 1-2 (2005), 359–375.

[4] Joy Buolamwini and Timnit Gebru. 2018. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Conference on fairness, accountability and transparency*. PMLR, 77–91.

[5] Katherine S Button, John PA Ioannidis, Claire Mokrysz, Brian A Nosek, Jonathan Flint, Emma SJ Robinson, and Marcus R Munafò. 2013. Power failure: why small sample size undermines the reliability of neuroscience. *Nature reviews neuroscience* 14, 5 (2013), 365–376.

[6] Susan Carey. 2004. Bootstrapping & the origin of concepts. *Daedalus* 133, 1 (2004), 59–68.

[7] Su Chen, Ying Fang, Genghu Shi, John Sabatini, Daphne Greenberg, Jan Frijters, and Arthur C Graesser. 2021. Automated disengagement tracking within an intelligent tutoring system. *Frontiers in artificial intelligence* 3 (2021), 595627.

[8] Davide Chicco and Giuseppe Jurman. 2020. The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. *BMC genomics* 21 (2020), 1–13.

[9] Alexandra Chouldechova. 2017. Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *Big data* 5, 2 (2017), 153–163.

[10] Alexandra Chouldechova and Aaron Roth. 2020. A snapshot of the frontiers of fairness in machine learning. *Commun. ACM* 63, 5 (2020), 82–89.

[11] Natasha Codiroli Mcmaster and Rose Cook. 2019. The contribution of intersectionality to quantitative research into educational inequalities. *Review of Education* 7, 2 (2019), 271–292.

[12] Damien Dablain, Bartosz Krawczyk, and Nitesh Chawla. 2022. Towards a holistic view of bias in machine learning: Bridging algorithmic fairness and imbalanced learning. *arXiv preprint arXiv:2207.06084* (2022).

[13] Shane Dawson, Jelena Jovanovic, Dragan Gašević, and Abelardo Pardo. 2017. From prediction to impact: Evaluation of a learning analytics retention program. In *Proceedings of the seventh international learning analytics & knowledge conference*. 474–478.

[14] Thomas J Diciccio and Joseph P Romano. 1988. A review of bootstrap confidence intervals. *Journal of the Royal Statistical Society Series B: Statistical Methodology* 50, 3 (1988), 338–354.

[15] Morten W Fagerland, Stian Lydersen, and Petter Laake. 2015. Recommended confidence intervals for two independent binomial proportions. *Statistical methods in medical research* 24, 2 (2015), 224–254.

[16] Michael Feldman, Sorelle A Friedler, John Moeller, Carlos Scheidegger, and Suresh Venkatasubramanian. 2015. Certifying and removing disparate impact. In *proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining*. 259–268.

[17] Léo Gautheron, Amaury Habrard, Emilie Morvant, and Marc Sebban. 2019. Metric learning from imbalanced data. In *2019 IEEE 31st International Conference on Tools with Artificial Intelligence (ICTAI)*. IEEE, 923–930.

[18] Moritz Hardt, Eric Price, and Nati Srebro. 2016. Equality of opportunity in supervised learning. *Advances in neural information processing systems* 29 (2016).

[19] László A Jeni, Jeffrey F Cohn, and Fernando De La Torre. 2013. Facing imbalanced data–recommendations for the use of performance metrics. In *2013 Humaine association conference on affective computing and intelligent interaction*. IEEE, 245–251.

[20] Haewon Jeong, Michael D Wu, Nilanjana Dasgupta, Muriel Médard, and Flavio Calmon. 2022. Who gets the benefit of the doubt? Racial bias in machine learning algorithms applied to secondary school math education. *Math AI for Education: Bridging the Gap Between Research and Smart Education* (2022).

[21] Shamya Karumbaiah, Ryan S Baker, and Valerie Shute. 2018. Predicting Quitting in Students Playing a Learning Game. *International Educational Data Mining Society* (2018).

[22] Shamya Karumbaiah and Jamiella Brooks. 2021. How colonial continuities underlie algorithmic injustices in education. In *2021 Conference on Research in Equitable and Sustained Participation in Engineering, Computing, and Technology (RESPECT)*. IEEE, 1–6.

[23] Shamya Karumbaiah, Jaclyn Ocumpaugh, and Ryan S Baker. 2022. Context matters: Differing implications of motivation and help-seeking in educational technology. *International Journal of Artificial Intelligence in Education* 32, 3 (2022), 685–724.

[24] Michael Kearns, Seth Neel, Aaron Roth, and Zhiwei Steven Wu. 2018. Preventing fairness gerrymandering: Auditing and learning for subgroup fairness. In *International conference on machine learning*. PMLR, 2564–2572.

[25] Gregory A Kimble. 1987. The scientific value of undergraduate research participation. *American Psychological Association* (1987).

[26] Kweku Kwegyir-Aggrey, Marissa Gerchick, Malika Mohan, Aaron Horowitz, and Suresh Venkatasubramanian. 2023. The misuse of AUC: What high impact risk assessment gets wrong. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*. 1570–1583.

[27] Lifeng Lin. 2018. Bias caused by sampling error in meta-analysis with small sample sizes. *PloS one* 13, 9 (2018), e0204056.

[28] Mónica V Martins, Daniel Tolledo, Jorge Machado, Luís MT Baptista, and Valentim Realinho. 2021. Early prediction of student's performance in higher education: a case study. In *Trends and Applications in Information Systems and Technologies: Volume 1 9*. Springer, 166–175.

[29] Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. 2021. A survey on bias and fairness in machine learning. *ACM computing surveys (CSUR)* 54, 6 (2021), 1–35.

[30] Shira Mitchell, Eric Potash, Solon Barocas, Alexander D'Amour, and Kristian Lum. 2021. Algorithmic fairness: Choices, assumptions, and definitions. *Annual review of statistics and its application* 8, 1 (2021), 141–163.

[31] Robert G Newcombe. 1998. Two-sided confidence intervals for the single proportion: comparison of seven methods. *Statistics in medicine* 17, 8 (1998), 857–872.

[32] Anne-Marie Núñez, Matthew J Mayhew, Musbah Shaheen, and Eric McChesney. 2023. Critical quantitative intersectionality: Maximizing integrity in expanding tools and applications. In *Handbook of Critical Education Research*. Routledge, 430–451.

[33] Jennifer K Olsen, Vincent Aleven, and Nikol Rummel. 2015. Predicting Student Performance in a Collaborative Learning Environment. *International Educational Data Mining Society* (2015).

[34] Sama Ranjeeth, Thamarai Pugazhendhi Latchoumi, and P Victer Paul. 2020. A survey on predictive models of learning analytics. *Procedia Computer Science* 167 (2020), 37–46.

[35] Pedro Saleiro, Benedict Kuester, Loren Hinkson, Jesse London, Abby Stevens, Ari Anisfeld, Kit T Rodolfa, and Rayid Ghani. 2018. Aequitas: A bias and fairness audit toolkit. *arXiv preprint arXiv:1811.05577* (2018).

[36] Salome E Scholtz. 2021. Sacrifice is a step beyond convenience: A review of convenience sampling in psychological research in Africa. *SA Journal of Industrial Psychology* 47, 1 (2021), 1–12.

[37] Philip Sedgwick. 2012. What is sampling error? *Bmj* 344 (2012).

[38] Lele Sha, Mladen Raković, Angel Das, Dragan Gašević, and Guanliang Chen. 2022. Leveraging class balancing techniques to alleviate algorithmic bias for predictive tasks in education. *IEEE Transactions on Learning Technologies* 15, 4 (2022), 481–492.

[39] Ajay S Singh and Micah B Masuku. 2014. Sampling techniques & determination of sample size in applied statistics research: An overview. *International Journal of economics, commerce and management* 2, 11 (2014), 1–22.

[40] Helen Smith. 2020. Algorithmic bias: should students pay the price? *AI & society* 35, 4 (2020), 1077–1078.

[41] Linda J Smith. 2008. How ethical is ethical research? Recruiting marginalized, vulnerable groups into health services research. *Journal of Advanced nursing* 62, 2 (2008), 248–257.

[42] KP Suresh and Sachin Chandrashekara. 2012. Sample size estimation and power analysis for clinical research studies. *Journal of human reproductive sciences* 5, 1 (2012), 7–13.

[43] Dirk T Tempelaar, André Heck, Hans Cuypers, Henk van der Kooij, and Evert van de Vrie. 2013. Formative assessment and learning analytics. In *Proceedings of the third international conference on learning analytics and knowledge*. 205–209.

[44] Gaël Varoquaux. 2018. Cross-validation failure: Small sample sizes lead to large error bars. *Neuroimage* 180 (2018), 68–77.

[45] Andres Felipe Zambrano, Jiayi Zhang, and Ryan S Baker. 2024. Investigating Algorithmic Bias on Bayesian Knowledge Tracing and Carelessness Detectors. In *Proceedings of the 14th Learning Analytics and Knowledge Conference*. 349–359.

[46] Jiayi Zhang, Juliana Ma Alexandra L Andres, Stephen Hutt, Ryan S Baker, Jaclyn Ocumpaugh, Nidhi Nasiar, Caitlin Mills, Jamiella Brooks, Sheela Sethuaman, Tyron Young, et al. 2022. Using machine learning to detect SMART model cognitive operations in mathematical problem-solving process. *Journal of Educational Data Mining* 14, 3 (2022), 76–108.

[47] Amin Zollanvari, Refik Caglar Kizilirmak, Yau Hee Kho, and Daniel Hernández-Torrano. 2017. Predicting students' GPA and developing intervention strategies based on self-regulatory learning behaviors. *IEEE Access* 5 (2017), 23792–23802.