JOURNAL OF LEARNING ANALYTICS
SOLAR SOCIETY for LEARNING ANALYTICS RESEARCH

Early Access Articles, 1–20. http://dx.doi.org/10.18608/jla.2026.9001

Beyond "Biased or Not": Reliable Estimation of Algorithmic Bias in Learning Analytics

Jaeyoon Choi[1]*, Shamya Karumbaiah[2], Jeffrey Matayoshi [3]

Abstract
While predictive models are widely used in learning analytics, several studies have shown that the performance of these models can exhibit biases against different demographic groups of students. The first step to audit for and mitigate these group biases is to accurately estimate them. However, the current practices for identifying and measuring group bias often suffer from reliability issues. In this paper, we use simulations and real-world data analysis to explore statistical factors that impact the reliable estimate of group bias and suggest approaches to improve their statistical robustness. 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 researchers to ensure sufficiently large group sizes, construct confidence intervals rather than relying on p-values, 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.

Notes for Practice

- Small group sizes can lead to high variability in group bias estimation. Hence, if possible, we should 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, may come at the cost of privacy. In this case, we recommend using the bootstrapping method.
- When reporting group bias estimates, we should go beyond merely reporting single values and instead compute confidence intervals to quantify the uncertainty in the estimated values. We specifically recommend using confidence intervals instead of p-values, as p-values can be easily misinterpreted and are less reliable with small sample sizes.
- We should move beyond the false dichotomy of the "presence or absence of group bias" because no educational algorithm is entirely free of bias. Instead, we should focus on understanding how much an algorithm is biased for different demographic groups of students.

Keywords
algorithmic bias, fairness, group bias, bias audit, estimation of bias, reliability, predictive models

Submitted: 18/04/2025 — **Accepted:** 13/04/2026 — **Published:** 08/07/2026

[1] **Corresponding author Email: jaeyoon.choi@uci.edu Address: School of Education, University of California, Irvine, Irvine, California, United States*
[2] *Email: shamya.karumbaiah@wisc.edu Address: Department of Educational Psychology, University of Wisconsin-Madison, Madison, Wisconsin, United States*
[3] *Email: jeffrey.matayoshi@mheducation.com Address: McGraw Hills ALEKS, Irvine, California, United States*

1. Introduction

Predictive models have been at the forefront of research and practice in learning analytics since its conception (Ranjeeth et al., 2020). They are used for summative and formative assessment (Tempelaar et al., 2013), predicting students' academic success (Zollanvari et al., 2017); recognizing which students are likely to drop out of a course (Dawson et al., 2017); and detecting disengagement in intelligent tutoring systems (Chen et al., 2021), learning games (Karumbaiah et al., 2018), and collaborative learning (Olsen et al., 2015). 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 (Jeong et al., 2022). 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 and colleagues (2022) 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) than for white students (AUC = 0.80), where AUC stands for area under the receiver operating characteristics (ROC) curve. 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 (Sedgwick, 2012)—for minority groups. For instance, in Zhang and colleagues (2022), there were only 18 Hispanic/Latinx students out of 72 total students. While this may appear reasonably represented as a proportion, such small absolute numbers raise serious questions about how accurate the reported AUC of 0.81 and subsequently the positive bias of 0.01 are for the Hispanic/Latinx student group. Such sampling error occurs more often than not in learning analytics research (Karumbaiah et al., 2022) for reasons such as overrepresentation of Western, educated, industrialized, rich, and Democratic (WEIRD) populations (Scholtz, 2021) or convenience sampling of undergraduate students (Kimble, 1987).

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 (Saleiro et al., 2018). As discussed above, if the estimation of group bias is unreliable, the auditing process could result in spurious results. Hence, current mitigation approaches (e.g., Kearns et al., 2018) cannot "mitigate" bias if the bias estimation is in itself inaccurate. Furthermore, inaccurate estimation of model performance on the minority group could cause the auditing process to wrongly conclude that the model is biased against the majority group. Then, any subsequent attempt to "adjust" for this perceived bias could then inadvertently introduce real 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 that do not themselves introduce bias toward any particular group—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. That is, under the condition that a model treats groups identically, we examine how those factors can inflate or deflate the estimated bias. 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 significantly 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 Algorithmic Bias in Learning Analytics

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 and colleagues (2020) 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., 2021; Mehrabi et al., 2021; Chouldechova, 2017). Mitchell and colleagues (2021) 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 and colleagues (2021) define fairness as the absence of prejudice or favouritism by a model toward certain individuals or groups based on their attributes.

In learning analytics research, given the prevalence of predictive models in the field, many studies focus on how model performance may differ across demographic groups—that is, on group fairness (see the following section for more details). For instance, Zambrano and colleagues (2024) examined group-level AUC differences in Bayesian knowledge-tracing and carelessness detectors. Similarly, Blazkova and colleagues (2025) investigated how AUC disparities in a dropout prediction model evolve over time across demographic groups in the Danish higher education context. H. Lee and colleagues (2025) used root mean squared error to identify bias against students with ADHD in a student performance classifier.

2.2 Group Fairness

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. Generally, a significant portion of this work is based on group fairness (sometimes referred to as statistical fairness, such as in Chouldechova and Roth (2020)), which 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 predicting that a student dropped out is approximately the same across different demographic groups (i.e., statistical parity (Feldman et al., 2015)). Other measures include false positive and false negative rates (FPRs and FNRs; i.e., equalized odds (Hardt et al., 2016)) and positive predictive value (similar to equalized calibration (Chouldechova, 2017)). Therefore, group fairness operates at the group level, guaranteeing fairness to the *average* members of the group instead of individuals.

In contrast, individual fairness emphasizes treating similar individuals similarly, meaning that individuals with similar characteristics should receive similar treatment regardless of their protected attributes (Dwork et al., 2012; Binns, 2020). Put simply, this translates to "less qualified individuals should not be favoured over more qualified individuals." In practice, it is often hard, if not impossible, to define a similarity metric based on how similar two individuals are in terms of relevant characteristics, and attempting to do so can itself introduce societal biases. For this reason, group fairness is more commonly used in empirical studies.

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 (A. Wang et al., 2022; Jo & Gebru, 2020; Li et al., 2022). One contributing factor is the historical distrust of institutions by many marginalized communities, rooted in past exploitation and coercion, such as in the Tuskegee Syphilis Study (Brandt, 1978). Additionally, socioeconomic barriers such as poverty and limited access to transportation, as well as social stigma, further complicate access to individuals from these groups (L. J. Smith, 2008). For example, Codiroli Mcmaster and Cook (2019) highlight that either the information about some intersectional groups is not collected or the sample sizes are insufficient for meaningful analysis. Lastly, convenience sampling, a method to recruit participants who are easily accessible to researchers, often excludes minority groups, as such samples tend to disproportionately consist of WEIRD populations (Bornstein et al., 2013).

Statistically, small sample sizes make it difficult to capture the full variability within a population (Núñez et al., 2023). In other words, small sample sizes increase sampling error—the difference between the sample estimate and the true population parameter (Sedgwick, 2012). Varoquaux (2018) demonstrated that small sample sizes in predictive models lead to large errors, compromising the reliability of the conclusions drawn from these models. Additionally, small sample sizes reduce the statistical power, which is the likelihood of detecting the true effect. In other words, small sample sizes make it harder to detect meaningful effects (Suresh & Chandrashekara, 2012; Button et al., 2013).

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 (Karumbaiah & Brooks, 2021).

2.3.2 Lack of Statistically Reliable Practice to Estimate Group Bias

Another significant challenge is the lack of a 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 and colleagues (2024) compared the 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 (Varoquaux, 2018). 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.

2.4 p-Values and Confidence Intervals for Group Bias Estimation

To improve current practices of reporting group bias in learning analytics research, it is essential to incorporate more rigorous statistical evidence. That is, rather than solely comparing performance metrics—such as AUC—across groups to determine the presence of bias, we should assess whether these bias estimates are reliable and accurate through formal statistical analyses.

Traditionally, p-values and confidence intervals have been used in inferential statistics to evaluate the validity and reliability of statistical estimates (D. K. Lee, 2016). These methods help determine whether an observed result is merely a random occurrence within the studied sample or whether it truly represents the underlying population. A p-value is the probability of obtaining a result at least as extreme as the one observed, under the assumption that the null hypothesis is true. In the context of algorithmic bias studies, p-values can be used to assess whether the measured group bias is statistically significant, with the null hypothesis positing that the difference in performance metrics between certain demographic groups is zero.

Despite their widespread use, p-values may not be the most suitable approach for evaluating algorithmic bias due to the unique challenges inherent in this field. As discussed in Section 2.3.1, algorithmic bias studies often struggle with small sample size issues. For instance, when examining whether a course dropout predictive model is fair across demographic groups, researchers might compare the model's outcomes for a historically minoritized group against those for the rest of the population or a historically privileged group. However, the sample size for the historically minoritized group is often significantly smaller, which increases the likelihood of generating an insignificant p-value (Button et al., 2013; Sullivan & Feinn, 2012), even when meaningful disparities exist. This could lead one to misinterpret the p-value as evidence of *no difference* between two groups, when in reality there could be some performance disparities but the sample size is too small to reach any strong conclusion.

Also, p-values can be easily misinterpreted if not used correctly. For instance, Ranstam (2012) highlights that a statistically insignificant p-value does not imply that an effect is absent in the population from which the sample is drawn. Instead, it often means that the sample size is too small to detect the effect, leading to a failure to reject the null hypothesis. In other words, an insignificant p-value represents an absence of evidence, not evidence of absence. This distinction is crucial, as researchers may mistakenly interpret an insignificant p-value as evidence of no group bias, rather than recognizing it as an absence of evidence due to insufficient statistical power. The fact that p-values are vulnerable to misinterpretation is particularly problematic in algorithmic bias studies, where small sample sizes for minoritized groups can obscure meaningful disparities, leading to false conclusions of fairness.

Therefore, it is a logical decision to use confidence intervals when examining group bias in predictive learning analytics. Confidence intervals provide a range of values within which the true population parameter is likely to lie, offering a more comprehensive understanding of the observed bias. Unlike p-values, which only indicate whether an observed bias is statistically significant, confidence intervals allow researchers to assess the reliability and precision of the bias estimate by presenting a range of plausible values. This added layer of information helps in evaluating the robustness of the findings. For instance, if a confidence interval for the difference in AUC between certain demographic groups is wide and includes zero, we might cautiously interpret the model as fair, but with considerable uncertainty, most likely due to the small sample sizes.

In other words, using confidence intervals supports more transparent reporting by conveying the degree of uncertainty around bias estimates, helping to avoid overconfidence in point estimates. This approach aligns with the broader movement in statistical research that encourages moving beyond p-value-centric analyses to embrace approaches that emphasize estimation accuracy (Gardner & Altman, 1986; D. K. Lee, 2016; Ranstam, 2012). In the context of learning analytics, incorporating confidence intervals can lead to more responsible and nuanced interpretations of fairness.

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 (Singh & Masuku, 2014; Lin, 2018). Small sample sizes are often inadequate to fully represent the variability found in a population (Núñez et al., 2023). 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 (Sedgwick, 2012). 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 make conclusive decisions based on such a small, and potentially unrepresentative, 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 (Suresh & Chandrashekara, 2012)[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 and colleagues (2024), 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 (Gautheron et al., 2019; Jeni et al., 2013). 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, so class distribution is an important factor to consider in binary classification settings. According to Dablain and colleagues (2024), a lack of balance in class distribution could make the classifiers more biased toward the 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 in the deployment process, not how the bias from class distribution is introduced in the training process itself.

Jeni and colleagues (2013) 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 AUC is not affected. Similarly, studies such as Chicco and Jurman (2020) 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 the *error rate*—the fraction of instances that are misclassified. When the error rate differs across groups, we consider it as evidence for group bias. For instance, an algorithm developed to predict UK students' exam grades in 2020 assigned lower grades to students in public schools than those in private schools (H. Smith, 2020), 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 depends on the metrics; see Kwegyir-Aggrey et al. (2023)). However, when the same error is applied uniformly across all subgroups—that is, when a classifier is designed to be fair across all subgroups—we would not expect differences in classification performance between them, assuming all other conditions remain the same. In other words, if the error rate is held constant across groups—thereby eliminating any performance gap by design—its magnitude alone should not introduce bias or lead to disparities in performance.

[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 (Sha et al., 2022)) 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.

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 the algorithmic bias literature, metrics that measure the error—such as the 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 formulas for precision and recall:

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

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 is 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 and colleagues (2013) show 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 and colleagues (2023) argue that AUC is dependent on class distribution and misclassification errors. Therefore, the decision of which metrics to use 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 and colleagues (2018) used equalizing the FPR 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 evidence of differential treatment by the classifier, which constitutes the definition of bias.

Table 1. Factors affecting reliable estimation of group 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)

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, by design, the simulated classifiers 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. (However, note that we also conduct a real data analysis in Section 5.)

Using simulations, we address RQ1: What statistical factors influence the reliable estimation of bias in predictive learning analytics? Specifically, we simulate three cases using different metrics: FPR, FNR, and AUC. Given the distinct characteristics of these metrics, we apply different simulation approaches: FPR and FNR are proportion-based metrics, dependent on binary classification outcomes, whereas AUC requires predicted probabilities to evaluate the model's overall discriminatory ability.

4.1 Simulation for FPR and FNR

4.1.1 Simulation Setup

FPR and FNR are commonly used metrics in fairness research across the learning analytics and machine learning literature (Kearns et al., 2018; Chouldechova, 2017). 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., $FNR = \frac{FN}{FN+TP}$).

To simulate the classification process for FPR and FNR, we assume two datasets exist: (1) A: *Actual* set (true labels) and (2) 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 the *Target*), with all other groups in the dataset collectively referred to as *Others*. In our simulation, the *Target* group is always smaller than the *Others* group, modelling so-called *Minority* (*Target*) and *Majority* (*Others*) groups. We intentionally did not use the terms *Minority* and *Majority* in their conventional sense because they can carry different meanings in statistical and sociohistorical contexts. In this simulation, these terms are defined solely by relative group size (i.e., underrepresentation in numbers).

Lastly, note that the *Target* group and the *Others* group are mutually exclusive in this simulation. While in practice they 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

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

We also introduce different factors discussed in Section 3.

4.1.2 Group Sample Size

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

- When the total dataset size is 1,000 ($|A| = |P| = 1000$), *Target* group size ($|A_{Target}| = |P_{Target}|$) = 10, 20, 50, 100, 200, 300, 400 (*Others* group size $= 1,000 - Target$ group size).
- When the 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 (*Others* group size $= 10,000 - Target$ 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., Zambrano et al., 2024).

4.1.3 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}.

4.1.4 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 its 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 a 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 (Buolamwini & Gebru, 2018)), in this simulation we apply the same error rate to both the *Target* and *Others* groups. 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 imply that factors that are not related to bias could influence the bias estimation.

4.1.5 Metric: FPR Bias and FNR Bias

We define bias for FPR and FNR using differences:

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

That is, bias is defined as the difference in a given metric between the *Target* group and the *Others* group. 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 than for the *Others* group. Similarly, a positive FNR_{Diff} implies that the result is more biased against the *Target* group.

4.2 Simulation for AUC

4.2.1 Simulation Setup

AUC is a commonly used metric in predictive learning analytics to evaluate the discriminative ability of a model. Specifically, the ROC curve plots the TPR against the FPR at different classification thresholds. Therefore, AUC measures the probability that a randomly chosen positive instance is ranked higher by the model than a randomly chosen negative instance. An AUC of 0.5 indicates that the model performs no better than random guessing.

While simulating FPR and FNR requires generating binary labels, simulating AUC requires a different approach, as it relies on both predicted probabilities and binary labels. To simulate data for AUC, we first specify the overall AUC of the dataset to be generated. This is equivalent to setting the error rate—for instance, a higher AUC overall indicates better performance and, consequently, a lower error rate. We consider AUC values of 0.6, 0.7, 0.8, and 0.9 and then convert each AUC into Cohen's d using the transformation proposed by Salgado (2018). Next, we sample predicted probabilities from two standard normal distributions that are d standard deviations apart, specifically $N(0,1)$ and $N(d,1)$. Samples from $N(0,1)$ correspond to predicted probabilities for actual negative labels (hereinafter denoted as y^-), while samples from $N(d,1)$ correspond to predicted probabilities for actual positive labels (hereinafter denoted as y^+). In the following sections, we describe how this simulation is conducted for different groups. See Algorithm 1 for a detailed algorithm for this simulation.

4.2.2 Group Sample Size

As mentioned in Section 4.1.2, we assume two groups in the total dataset: the *Target* group and the *Others* group. We examine the same group sizes as discussed in Section 4.1.2.

4.2.3 Class Distribution

In this simulation, class distribution refers to the proportion of positive instances in the actual set: $\frac{y^+}{y^+ + y^-}$. Similar to Section 4.1.3, we investigate five different class distributions for each group: 0.05, 0.1, 0.2, 0.3, and 0.4.

4.2.4 Error Rate

As discussed in Section 4.1.4, the error rate refers to the misclassification rate. Unlike the FPR and FNR simulation in Section 4.1, we do not explicitly introduce an error rate. Instead, we simulate different AUC values to generate data, which is equivalent to introducing error rates—that is, a lower AUC corresponds to a higher error rate. This approach also ensures that the same error rate is applied to both the *Target* and the *Others* group.

Algorithm 1: Simulate Data with AUC

Input: *Target* group size n, *Target* class distribution α, *Others* group size m, *Others* class distribution β

for *each AUC value A in* $\{0.6, 0.7, 0.8, 0.9\}$ **do**
- Compute Cohen's d from A;
- **for** *each group G in* $\{$*Target, Others*$\}$ **do**
  - **if** *G is Target* **then**
    - Set group size $N \leftarrow n$ and class distribution $\gamma \leftarrow \alpha$;
  - **else**
    - Set group size $N \leftarrow m$ and class distribution $\gamma \leftarrow \beta$;
  - Compute the number of actual positive labels $|y^+|$ as $N\gamma$ and the number of actual negative labels $|y^-|$ as $N - N\gamma$;
  - **for** *each element in* y^+ **do**
    - Sample predicted probability $p \sim N(d,1)$;
  - **for** *each element in* y^- **do**
    - Sample predicted probability $p \sim N(0,1)$;

return Simulated predicted probabilities and actual labels for both groups

4.2.5 Metric: AUC Bias

Similar to Section 4.1.5, we define AUC bias using difference:

$$AUC_{Diff} = AUC_{Target} - AUC_{Others}.$$

However, the interpretation of AUC_{Diff} differs from that of FPR_{Diff} and FNR_{Diff}. While a positive FPR or FNR bias indicates that the classifier is more likely to be biased against *Target* groups, a positive AUC bias suggests that the classifier is more favourable toward the *Target* group than the *Others* group.

4.3 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

See Table 2 for the full description of factors used in our simulation.

Table 2. Summary of full parameters in simulation.

		Total dataset size = 1000	**Total dataset size = 10000**
Target group size		10, 20, 50, 100, 200, 300, 400	10, 20, 50, 100, 500, 1000, 2000, 3000, 4000
Class distribution		0.05, 0.1, 0.2, 0.3, 0.4	
Error rate	**FPR/FNR simulation**	Error rate = 0.1, 0.2, 0.3, . . . , 0.9	
	AUC simulation	Total AUC = 0.6, 0.7, 0.8, 0.9	

To quantify the variability in our experiment, we repeat the simulation of each unique combination of factors 100 times. In each simulation, we compute the results for the *Target* and *Others* groups. Bias is calculated as the difference between the two groups' results. Finally, to assess the variability of the bias estimates, we compute the 5th and 95th percentiles based on the bias values computed from 100 simulations.

As described in Section 4.1.4, we introduce an equal amount of error to both the *Target* and the *Others* group to ensure that the simulation process itself does not introduce any bias. Therefore, theoretically the results for the *Target* group and the *Others* group should be approximately the same—specifically, the FPR_{Diff}, FNR_{Diff}, or AUC_{Diff} 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.

4.4 Simulation Results

4.4.1 Finding 1: Small Group Sizes Result in High Variability

Figure 1 shows the 5th and 95th percentiles for FPR_{Diff}, with each dot representing the median. That is, each bar covers the middle 90% of the FPR bias estimates. The blue horizontal line represents when the FPR bias estimate is 0 (e.g. $FPR_{Diff} = 0$). 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 bias estimates. 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 (i.e., more biased against *Target*) values, 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 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 argue that the classifier is biased against both the *Target* and *Others* groups.

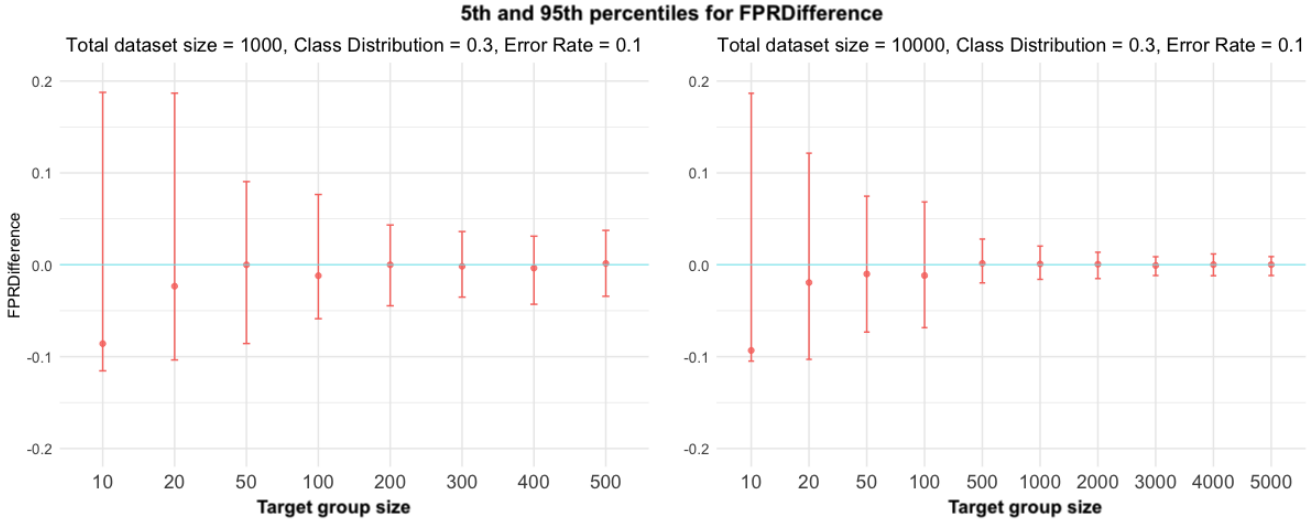

Figure 1. 5th and 95th percentile intervals for FPR_{Diff} ($FPR_{Target} - FPR_{Others}$) with varying *Target* group sizes. The left plot represents a total dataset size of 1,000, while the right plot represents a dataset size of 10,000. Each bar covers from the 5th to 95th percentiles, with a dot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

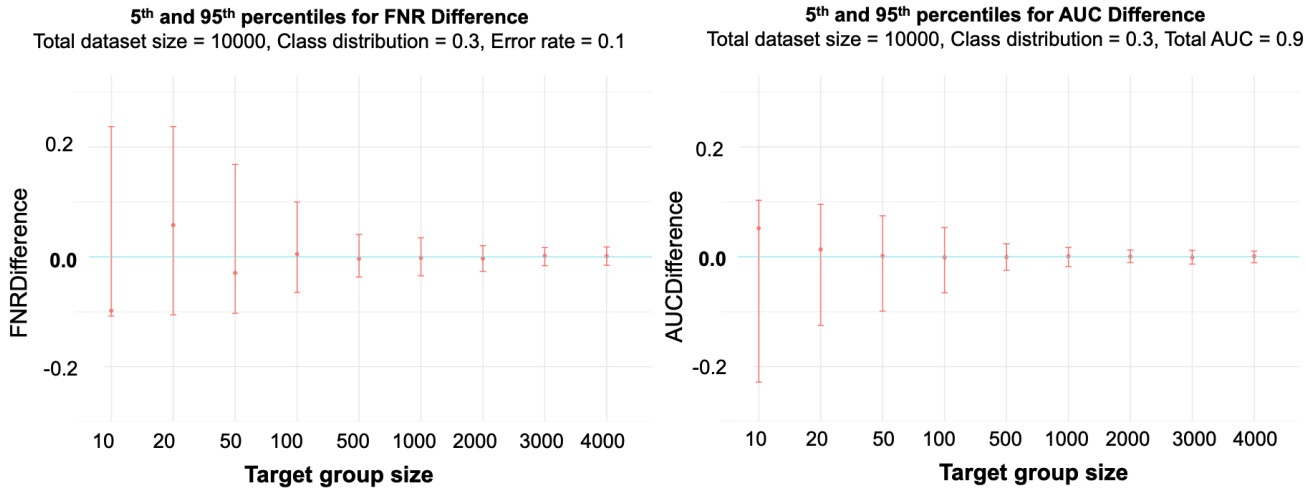

Figure 2. 5th and 95th percentile intervals for FNR_{Diff} and AUC_{Diff}. All cases with total dataset size = 10000, class distribution = 0.3, and error rate = 0.3 (for AUC simulation, total AUC = 0.9).

4.4.2 Finding 2: The Reliability Issue in Bias Estimation Persists Regardless of the Chosen Metric

Figure 2 shows the 5th and 95th percentile intervals for FNR_{Diff} and AUC_{Diff}. Consistent with Finding 1, all of the metrics exhibit a similar pattern: smaller group sizes lead to greater variability in bias estimates. Moreover, the intervals extend into both positive and negative values. This means that although our simulations are designed to perform equally for both groups, the point estimate for group bias may suggest bias in both directions—the classifier might be biased either against or in favour of the *Target* group when its size is particularly small. Lastly, although not shown due to space limitations, we found that the same pattern was observed for precision and recall as well.

4.4.3 Finding 3: Class Distribution and Error Rate Do Influence Bias Estimate Variability

Figures 3 and 4 show the 5th and 95th percentile intervals for AUC_{Diff} with varying class distributions and error rates (i.e., total AUC simulated). In general, the intervals in Figure 3 are wider than those in Figure 4. For example, when both the *Target* and *Others* groups have a class distribution of 0.1 and the *Target* group size is 10 (as shown in the top left plot of Figure 3), the bias estimate for the *Target* group ranges from -0.561 to 0.404 at a total AUC of 0.6. Under the same conditions but with a higher total AUC of 0.9, the interval narrows to -0.339 to 0.121 (as shown in the top left plot of Figure 4). In other words, introducing higher error rates overall (i.e., a lower total AUC) leads to more variable bias estimates. However, it should be noted that this pattern was observed only for AUC—that is, increasing error rates did not consistently lead to higher variability for FPR or FNR.

Also, from the left plots in Figure 4, we observe that when the *Target* group size is small (10) and the class distribution is low (0.1), the distribution of AUC differences is skewed negatively. Specifically, the percentile intervals for the *Target* group size of 10 are $(-0.339, 0.121)$ in the top left plot and $(-0.554, 0.111)$ in the bottom left plot. In other words, under these conditions, it is likely, just by chance, to observe a negative AUC_{Diff}, which could lead to mistaken interpretation that there exists bias against the *Target* group.

Furthermore, by comparing the top two plots in Figure 3 and 4, we observe that a higher class distribution decreases the bias estimate interval. For instance, with a total AUC of 0.6, when both the *Target* and *Others* groups have a class distribution of 0.4 and the *Target* group size is 10, the bias estimate becomes -0.232 to 0.286, compared to -0.561 to 0.404 when both groups have a class distribution of 0.1.

Moreover, when the *Target* and *Others* groups have different class distributions, we find that whenever the *Target* group's class distribution is relatively low, the bias estimate interval becomes wider. For example, with a total AUC of 0.6, when the *Target* class distribution is 0.1 and the *Others* class distribution is 0.4, the bias estimate for a *Target* group size of 10 ranges from -0.601 to 0.416. However, when the class distributions are reversed, the range changes to -0.388 to 0.275. This suggests that a lower class distribution for a smaller group can increase the bias estimate range.

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 (Martins et al., 2021). 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 the binary variable *gender* to construct groups (female and male[3]) and selected seven variables (previous qualification, grade from previous qualification, mother job, father job[4], admission grade, educational special needs, scholarship holder) as predictors based on Martins and colleagues (2021) to predict students' success (i.e., graduation).

Among the 3,630 students, around 60% graduated, while 40% dropped out. In other words, the class distribution is 0.6. Specifically, the data consists of 2,381 female students ($\sim$65%) and 1,249 male students ($\sim$35%). Among female students,

[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 original study classified gender as a binary variable. This limitation reflects the structure of the data and not our perspective on gender.

[4]Given that prior research has demonstrated a strong association between parents' education and occupation and their children's academic performance (Eccles, 2005; W. Wang et al., 2020), including parental occupation as a predictor variable may risk introducing bias into the model. However, the purpose of using this data and the prediction model is not to conduct a bias audit, but rather to illustrate how bias estimation can be made more reliable through additional statistical evidence (e.g., constructing confidence intervals).

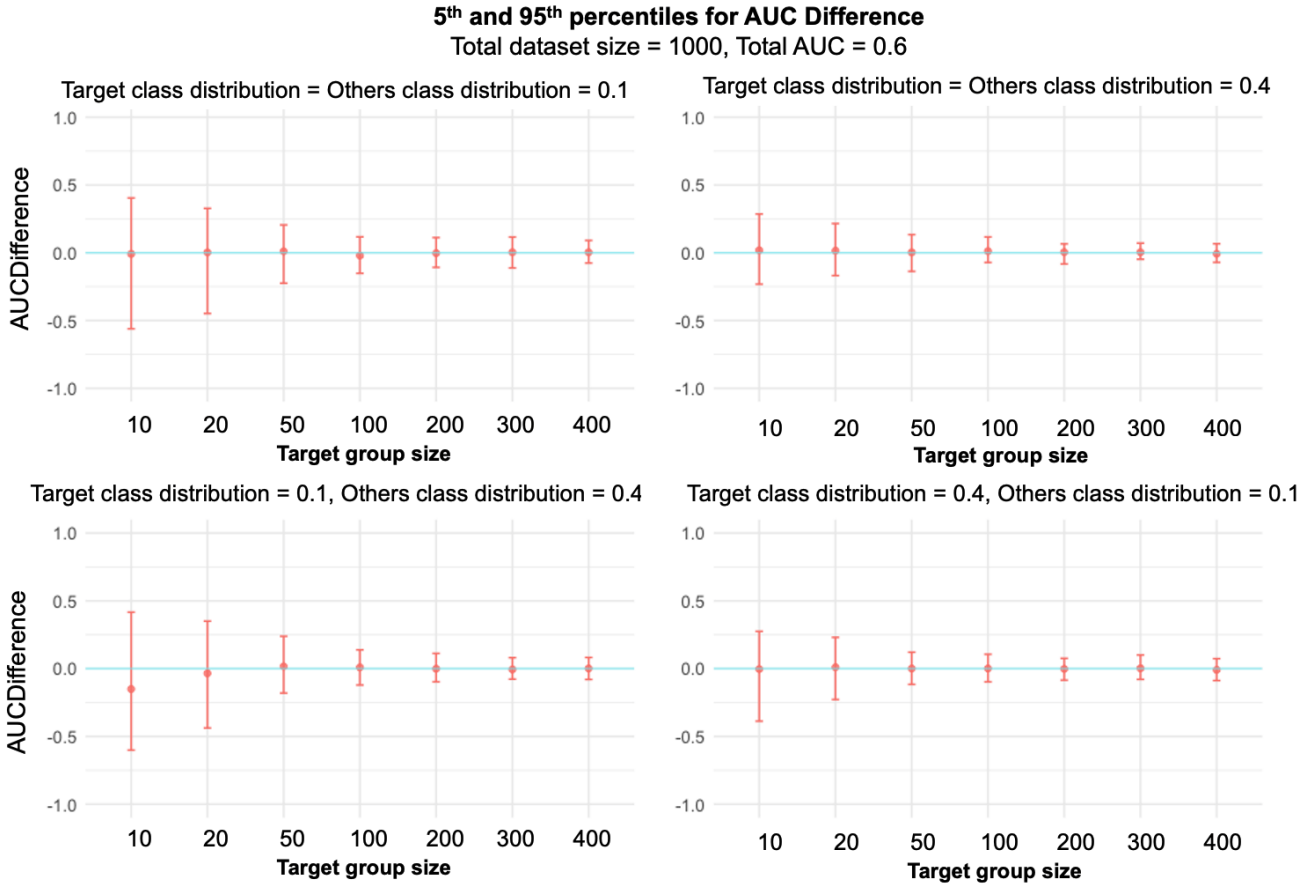

Figure 3. 5th and 95th percentile intervals for AUC_{Diff} with total dataset size = 1000 and total AUC = 0.6. Specifically, the top two plots correspond to cases where the *Target* and *Others* groups have the same class distribution (left: 0.1, right: 0.4), while the bottom two plots correspond to cases where the class distributions differ between the two groups (0.1 and 0.4).

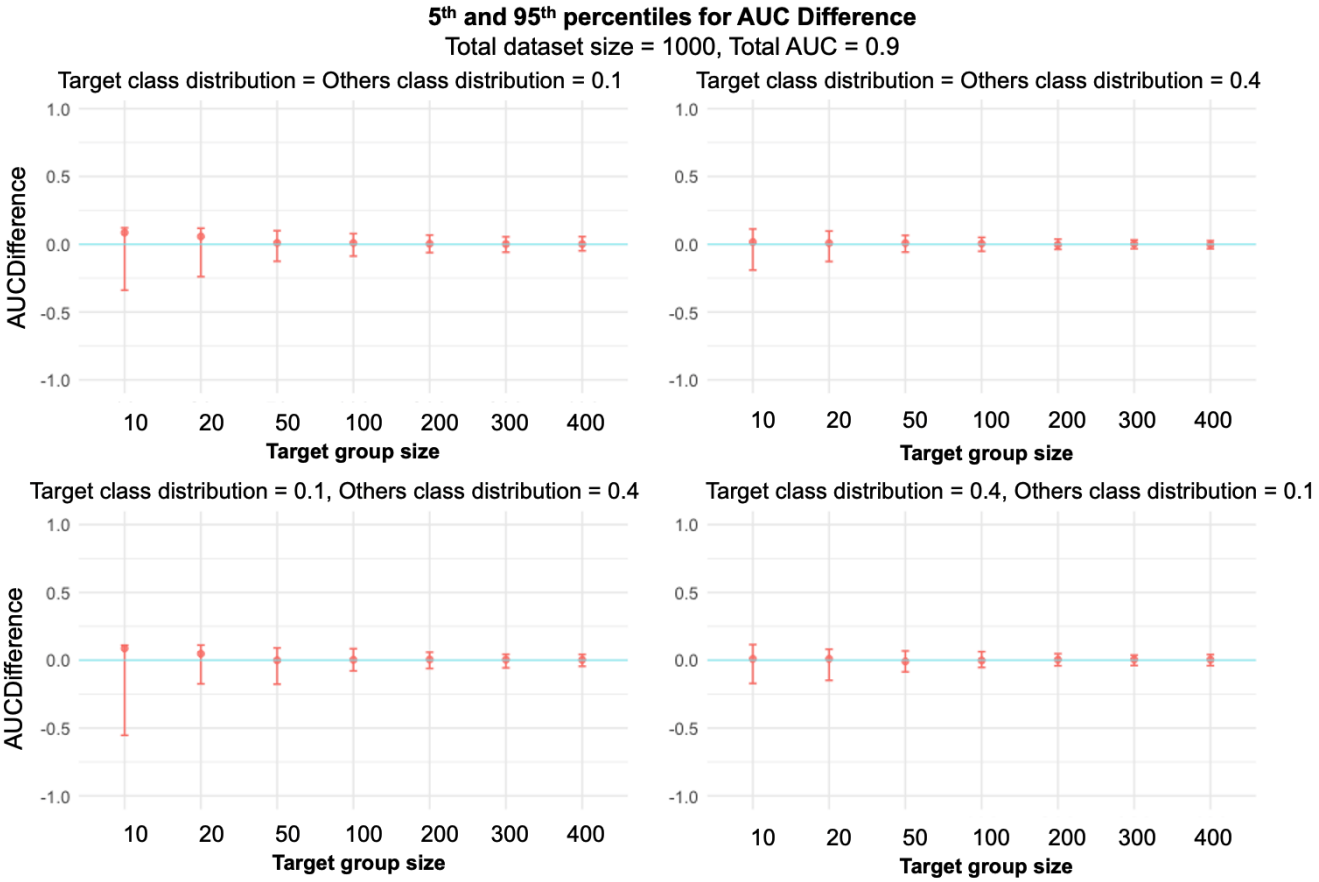

Figure 4. 5th and 95th percentile intervals for AUC_{Diff} with total dataset size = 1000 and total AUC = 0.9. Specifically, the top two plots correspond to cases where the *Target* and *Others* groups have the same class distribution (left: 0.1, right: 0.4), while the bottom two plots correspond to cases where the class distributions differ between the two groups (0.1 and 0.4).

around 70% graduated, while 30% dropped out (i.e., the class distribution for female students was 0.7). Among male students, around 44% graduated, while 56% dropped out (i.e., the class distribution for male students was 0.44).

5.2 Modelling

We first built a logistic regression model that predicts students' graduation using the seven 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 maintained a similar proportion of gender. Because optimization is not the purpose of this paper, we skipped hyperparameter tuning and 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 an AUC of 0.722 and an 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 (1998) 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 Newcombe, 1998; Brown & Li, 2005; Fagerland et al., 2015, 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)**

   After completing the modelling 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 (Carey, 2004). By using bootstrapping, we can repeatedly resample small numbers 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 (Diciccio & Romano, 1988).

5.4 Results

The top interval in Figure 5 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 Figure 5 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 the observed group differences at hand. For instance, suppose we only examine the computed difference values (FPR_{Diff} = 0.063, AUC_{Diff} = 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 intervals show a range of values for bias estimation and help us understand if the computed group differences at hand could indicate group bias or not. Hence, estimating bias becomes more reliable with confidence intervals.

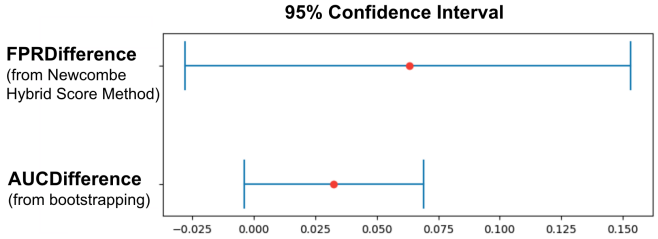

Figure 5. 95% confidence intervals for FPR_{Diff} and AUC_{Diff}. FPR_{Diff} is computed using the Newcombe hybrid score method, while AUC_{Diff} is computed using bootstrapping and BCa. Each red dot represents the computed model performance difference between female and male.

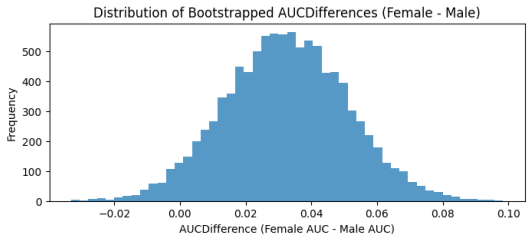

Figure 6. Distribution of bootstrapped AUC_{Diff} between female and male groups.

6. Discussion

6.1 Summary of Findings

In this paper, we examine 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?

We first conducted a simulation study to explore how group sample size, class distributions, 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. Most importantly, we found that group sample sizes significantly impact bias estimation: smaller group sizes lead to high variability in bias estimation, making the bias estimation process unreliable. Furthermore, we also observed that both error rates and class distributions influence bias estimations; specifically, higher error rates and lower class distributions could result in less reliable bias estimates, particularly when small groups show low class distributions. However, overall, group size remains the most influential factor.

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 related to the population (Singh & Masuku, 2014; Lin, 2018; Núñez et al., 2023). 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 and Brooks (2021) 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 collected a small number of samples, you can resample with replacement and construct a confidence interval to help identify whether 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 to report a single value (e.g., the FPR difference between the *Target* group and the *Others* group 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. Furthermore, we recommend using confidence intervals instead of p-values for algorithmic bias analysis, as p-values can be easily misinterpreted, especially when dealing with small sample sizes (see Section 2.4).

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 your collected samples with replacement for multiple iterations). Then construct BCa confidence intervals.

6.2.3 Examine at Least Two Metrics for Comprehensive Bias Analysis

Furthermore, we recommend estimating group bias using more than one performance metric. While this is a common practice in machine learning pipelines, doing so is particularly important in algorithmic bias analysis because different metrics can interact differently with various factors. For example, as noted in Section 4.4.3, increasing error rates tends to widen the bias intervals for AUC, but this effect is not necessarily observed for FPR or FNR bias. Additionally, while not included in the results section, we also observed that FNR bias can be inflated in cases where the sample size is very small and the number of actual positives is scarce. Therefore, we advise using at least two different performance metrics to ensure a comprehensive evaluation of bias and to verify whether they result in consistent conclusions.

6.2.4 Go 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.

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 intervals with the Newcombe hybrid score or bootstrapping. Improving methods used for bias research is crucial to prevent our efforts from further exacerbating bias for minority groups.

Acknowledgements

We appreciate Daniel Bolt for his valuable feedback on the methods used in this study.

Declaration of Conflicting Interest

The authors declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The publication of this article received financial support from the Office of the Vice Chancellor for Research and Graduate Education at the University of Wisconsin–Madison and the Wisconsin Alumni Research Foundation.

References

Binns, R. (2020). On the apparent conflict between individual and group fairness. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency* (FAT* 2020), 27–30 January 2020, Barcelona, Spain (pp. 514–524). ACM. https://doi.org/10.1145/3351095.3372864

Blazkova, T., Kizilcec, R. F., Nielsen, M. L., Lassen, D. D., & Bjerre-Nielsen, A. (2025). Fairness over time: A nationwide study of evolving bias in dropout prediction. In *Proceedings of the 12th ACM Conference on Learning at Scale* (L@S 2025), 21–23 July 2025, Palermo, Italy (pp. 246–250). ACM. https://doi.org/10.1145/3698205.3733933

Blodgett, S. L., Barocas, S., Daumé III, H., & Wallach, H. (2020). Language (technology) is power: A critical survey of "bias" in NLP. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* (ACL 2020), 5–10 July 2020, online (pp. 5454–5476). Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.acl-main.485

Bornstein, M. H., Jager, J., & Putnick, D. L. (2013). Sampling in developmental science: Situations, shortcomings, solutions, and standards. *Developmental Review*, *33*(4), 357–370. https://doi.org/10.1016/j.dr.2013.08.003

Brandt, A. M. (1978). Racism and research: The case of the Tuskegee Syphilis Study. *Hastings Center Report*, *8*(6), 21–29. https://doi.org/10.2307/3561468

Brown, L., & Li, X. (2005). Confidence intervals for two sample binomial distribution. *Journal of Statistical Planning and Inference*, *130*(1-2), 359–375. https://doi.org/10.1016/j.jspi.2003.09.039

Buolamwini, J., & Gebru, T. (2018). Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Proceedings of the First International Conference on Fairness, Accountability and Transparency* (FAT* 2018), 23–24 February 2018, New York, New York, USA (pp. 77–91). PMLR. https://proceedings.mlr.press/v81/buolamwini18a.html

Button, K. S., Ioannidis, J. P., Mokrysz, C., Nosek, B. A., Flint, J., Robinson, E. S., & Munafò, M. R. (2013). Power failure: Why small sample size undermines the reliability of neuroscience. *Nature Reviews Neuroscience*, *14*(5), 365–376. https://doi.org/10.1038/nrn3475

Carey, S. (2004). Bootstrapping & the origin of concepts. *Daedalus*, *133*(1), 59–68. https://doi.org/10.1162/001152604772746701

Chen, S., Fang, Y., Shi, G., Sabatini, J., Greenberg, D., Frijters, J., & Graesser, A. C. (2021). Automated disengagement tracking within an intelligent tutoring system. *Frontiers in Artificial Intelligence*, *3*, 595627. https://doi.org/10.3389/frai.2020.595627

Chicco, D., & Jurman, G. (2020). The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. *BMC Genomics*, *21*(6). https://doi.org/10.1186/s12864-019-6413-7

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

Chouldechova, A., & Roth, A. (2020). A snapshot of the frontiers of fairness in machine learning. *Communications of the ACM*, *63*(5), 82–89. https://doi.org/10.1145/3376898

Codiroli Mcmaster, N., & Cook, R. (2019). The contribution of intersectionality to quantitative research into educational inequalities. *Review of Education*, *7*(2), 271–292. https://doi.org/10.1002/rev3.3116

Dablain, D., Krawczyk, B., & Chawla, N. (2024). Towards a holistic view of bias in machine learning: Bridging algorithmic fairness and imbalanced learning. *Discover Data*, *2*(1), 4. https://doi.org/10.1007/s44248-024-00007-1

Dawson, S., Jovanovic, J., Gašević, D., & Pardo, A. (2017). From prediction to impact: Evaluation of a learning analytics retention program. In *Proceedings of the Seventh International Conference on Learning Analytics and Knowledge* (LAK 2017), 13–17 March 2017, Vancouver, British Columbia, Canada (pp. 474–478). ACM. https://doi.org/10.1145/3027385.3027405

Diciccio, T. J., & Romano, J. P. (1988). A review of bootstrap confidence intervals. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, *50*(3), 338–354. https://doi.org/10.1111/j.2517-6161.1988.tb01732.x

Dwork, C., Hardt, M., Pitassi, T., Reingold, O., & Zemel, R. (2012). Fairness through awareness. In *Proceedings of the Third Innovations in Theoretical Computer Science Conference* (ITCS 2012), 8–10 January 2012, Cambridge, Massachusetts, USA (pp. 214–226). ACM. https://doi.org/10.1145/2090236.2090255

Eccles, J. S. (2005). Influences of parents' education on their children's educational attainments: The role of parent and child perceptions. *London Review of Education*, *3*(3), 191–204. https://doi.org/10.1080/14748460500372309

Fagerland, M. W., Lydersen, S., & Laake, P. (2015). Recommended confidence intervals for two independent binomial proportions. *Statistical Methods in Medical Research*, *24*(2), 224–254. https://doi.org/10.1177/0962280211415469

Feldman, M., Friedler, S. A., Moeller, J., Scheidegger, C., & Venkatasubramanian, S. (2015). Certifying and removing disparate impact. In *Proceedings of the 21st ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (KDD 2015), 10–13 August 2015, Sydney, Australia (pp. 259–268). ACM. https://doi.org/10.1145/2783258.2783311

Gardner, M. J., & Altman, D. G. (1986). Confidence intervals rather than p values: Estimation rather than hypothesis testing. *British Medical Journal (Clinical Research Edition)*, *292*(6522), 746–750. https://doi.org/10.1136/bmj.292.6522.746

Gautheron, L., Habrard, A., Morvant, E., & Sebban, M. (2019). Metric learning from imbalanced data. In *Proceedings of the 2019 IEEE 31st International Conference on Tools with Artificial Intelligence* (ICTAI 2019), 4–6 November 2019, Portland, Oregon, USA (pp. 923–930). IEEE. https://doi.org/10.1109/ICTAI.2019.00131

Hardt, M., Price, E., & Srebro, N. (2016). Equality of opportunity in supervised learning. In D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, & R. Garnett (Eds.), *Proceedings of the 2016 Conference on Advances in Neural Information Processing Systems* (NIPS 2016), 5–10 December 2016, Barcelona, Spain. NeurIPS Proceedings. https://papers.neurips.cc/paper/6374-equality-of-opportunity-in-supervised-learning

Jeni, L. A., Cohn, J. F., & De La Torre, F. (2013). Facing imbalanced data—recommendations for the use of performance metrics. In *2013 Humaine Association Conference on Affective Computing and Intelligent Interaction* (ACII 2013), 2–5 September 2013, Geneva, Switzerland (pp. 245–251). IEEE. https://doi.org/10.1109/ACII.2013.47

Jeong, H., Wu, M. D., Dasgupta, N., Médard, M., & Calmon, F. (2022). Who gets the benefit of the doubt? Racial bias in machine learning algorithms applied to secondary school math education [Poster Presented at the 2021 Workshop Math AI for Education: Bridging the Gap between Research and Smart Education, 14 December 2021, online]. https://neurips.cc/virtual/2021/33829

Jo, E. S., & Gebru, T. (2020). Lessons from archives: Strategies for collecting sociocultural data in machine learning. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency* (FAT* 2020), 27–30 January 2020, Barcelona, Spain (pp. 306–316). ACM. https://doi.org/10.1145/3351095.3372829

Karumbaiah, S., Baker, R. S., & Shute, V. (2018). Predicting quitting in students playing a learning game. *Proceedings of the 11th International Conference on Educational Data Mining* (EDM 2018), 15–18 July 2018, Buffalo, New York, USA. https://educationaldatamining.org/files/conferences/EDM2018/papers/EDM2018_paper_39.pdf

Karumbaiah, S., & Brooks, J. (2021). How colonial continuities underlie algorithmic injustices in education. In *Proceedings of the 2021 Conference on Research in Equitable and Sustained Participation in Engineering, Computing, and Technology* (RESPECT 2021), 23–27 May 2021, online (pp. 1–6). IEEE. https://doi.org/10.1109/RESPECT51740.2021.9620605

Karumbaiah, S., Ocumpaugh, J., & Baker, R. S. (2022). Context matters: Differing implications of motivation and help-seeking in educational technology. *International Journal of Artificial Intelligence in Education*, *32*(3), 685–724. https://doi.org/10.1007/s40593-021-00272-0

Kearns, M., Neel, S., Roth, A., & Wu, Z. S. (2018). Preventing fairness gerrymandering: Auditing and learning for subgroup fairness. *Proceedings of Machine Learning Research*, *80*, 2564–2572. https://proceedings.mlr.press/v80/kearns18a.html

Kimble, G. A. (1987). The scientific value of undergraduate research participation. *American Psychologist*, *42*(3), 267–268. https://doi.org/10.1037/0003-066X.42.3.267.b

Kwegyir-Aggrey, K., Gerchick, M., Mohan, M., Horowitz, A., & Venkatasubramanian, S. (2023). The misuse of AUC: What high impact risk assessment gets wrong. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency* (FAccT 2023), 12–15 June 2023, Chicago, Illinois, USA (pp. 1570–1583). ACM. https://doi.org/10.1145/3593013.3594100

Lee, D. K. (2016). Alternatives to P value: Confidence interval and effect size. *Korean Journal of Anesthesiology*, *69*(6), 555. https://doi.org/10.4097/kjae.2016.69.6.555

Lee, H., Belitz, C., Nasiar, N., & Bosch, N. (2025). XAI reveals the causes of attention deficit hyperactivity disorder (ADHD) bias in student performance prediction. In *Proceedings of the 15th International Conference on Learning Analytics and Knowledge* (LAK 2025), 3–7 March 2025, Dublin, Ireland (pp. 418–428). ACM. https://doi.org/10.1145/3706468.3706521

Li, W., Sun, K., Schaub, F., & Brooks, C. (2022). Disparities in students' propensity to consent to learning analytics. *International Journal of Artificial Intelligence in Education*, *32*(3), 564–608. https://doi.org/10.1007/s40593-021-00254-2

Lin, L. (2018). Bias caused by sampling error in meta-analysis with small sample sizes. *PLoS One*, *13*(9), e0204056. https://doi.org/10.1371/journal.pone.0204056

Martins, M. V., Tolledo, D., Machado, J., Baptista, L. M., & Realinho, V. (2021). Early prediction of student's performance in higher education: A case study. In Á. Rocha, H. Adeli, G. Dzemyda, F. Moreira, & A. Ramalho Correia (Eds.), *Trends and applications in information systems and technologies. WorldCIST 2021. Advances in intelligent systems and computing* (pp. 166–175, Vol. 1365). Springer. https://doi.org/10.1007/978-3-030-72657-7_16

Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., & Galstyan, A. (2021). A survey on bias and fairness in machine learning. *ACM Computing Surveys (CSUR)*, *54*(6), 1–35. https://doi.org/10.1145/3457607

Mitchell, S., Potash, E., Barocas, S., D'Amour, A., & Lum, K. (2021). Algorithmic fairness: Choices, assumptions, and definitions. *Annual Review of Statistics and Its Application*, *8*(1), 141–163. https://doi.org/10.1146/annurev-statistics-042720-125902

Newcombe, R. G. (1998). Two-sided confidence intervals for the single proportion: Comparison of seven methods. *Statistics in Medicine*, *17*(8), 857–872. https://doi.org/10.1002/(SICI)1097-0258(19980430)17:8⟨857::AID-SIM777⟩3.0.CO;2-E

Núñez, A.-M., Mayhew, M. J., Shaheen, M., & McChesney, E. (2023). Critical quantitative intersectionality: Maximizing integrity in expanding tools and applications. In M. D. Young & S. Diem (Eds.), *Handbook of critical education research* (pp. 430–451). Routledge. https://doi.org/10.4324/9781003141464-25

Olsen, J. K., Aleven, V., & Rummel, N. (2015). Predicting student performance in a collaborative learning environment. In O. Santos, J. G. Boticario, C. Romero, M. Pechenizkiy, A. Merceron, P. Mitros, J. Luna, C. Mihaescu, P. Moreno, A. Hershkovitz, S. Ventura, & M. Desmarais (Eds.), *Proceedings of the Eighth International Conference on Educational Data Mining* (EDM 2015), 26–29 June 2015, Madrid, Spain (pp. 211–217). International Educational Data Mining Society. https://www.educationaldatamining.org/EDM2015/proceedings/edm2015_proceedings.pdf

Ranjeeth, S., Latchoumi, T. P., & Paul, P. V. (2020). A survey on predictive models of learning analytics. *Procedia Computer Science*, *167*, 37–46. https://doi.org/10.1016/j.procs.2020.03.180

Ranstam, J. (2012). Why the *P*-value culture is bad and confidence intervals a better alternative. *Osteoarthritis and Cartilage*, *20*(8), 805–808. https://doi.org/10.1016/j.joca.2012.04.001

Saleiro, P., Kuester, B., Hinkson, L., London, J., Stevens, A., Anisfeld, A., Rodolfa, K. T., & Ghani, R. (2018). Aequitas: A bias and fairness audit toolkit. *arXiv preprint arXiv:1811.05577*. https://doi.org//10.48550/arXiv.1811.05577

Salgado, J. F. (2018). Transforming the area under the normal curve (AUC) into Cohen's *d*, Pearson's r_{pb}, odds-ratio, and natural log odds-ratio: Two conversion tables. *European Journal of Psychology Applied to Legal Context*, *10*(1), 35–47. https://doi.org/10.5093/ejpalc2018a5

Scholtz, S. E. (2021). Sacrifice is a step beyond convenience: A review of convenience sampling in psychological research in Africa. *SA Journal of Industrial Psychology*, *47*(1), 1–12. https://doi.org/10.4102/sajip.v47i0.1837

Sedgwick, P. (2012). What is sampling error? *BMJ*, *344*. https://doi.org/10.1136/bmj.e4285

Sha, L., Raković, M., Das, A., Gašević, D., & Chen, G. (2022). Leveraging class balancing techniques to alleviate algorithmic bias for predictive tasks in education. *IEEE Transactions on Learning Technologies*, *15*(4), 481–492. https://doi.org/10.1109/TLT.2022.3196278

Singh, A. S., & Masuku, M. B. (2014). Sampling techniques & determination of sample size in applied statistics research: An overview. *International Journal of Economics, Commerce and Management*, *2*(11), 1–22. https://ijecm.co.uk/wp-content/uploads/2014/11/21131.pdf

Smith, H. (2020). Algorithmic bias: Should students pay the price? *AI & Society*, *35*(4), 1077–1078. https://doi.org/10.1007/s00146-020-01054-3

Smith, L. J. (2008). How ethical is ethical research? Recruiting marginalized, vulnerable groups into health services research. *Journal of Advanced Nursing*, *62*(2), 248–257. https://doi.org/10.1111/j.1365-2648.2007.04567.x

Sullivan, G. M., & Feinn, R. (2012). Using effect size—or why the P value is not enough. *Journal of Graduate Medical Education*, *4*(3), 279–282. https://doi.org/10.4300/JGME-D-12-00156.1

Suresh, K., & Chandrashekara, S. (2012). Sample size estimation and power analysis for clinical research studies. *Journal of Human Reproductive Sciences*, *5*(1), 7–13. https://doi.org/10.4103/0974-1208.97779

Tempelaar, D. T., Heck, A., Cuypers, H., van der Kooij, H., & van de Vrie, E. (2013). Formative assessment and learning analytics. In *Proceedings of the Third International Conference on Learning Analytics and Knowledge* (LAK 2013), 8–13 April 2013, Leuven, Belgium (pp. 205–209). ACM. https://doi.org/10.1145/2460296.2460337

Varoquaux, G. (2018). Cross-validation failure: Small sample sizes lead to large error bars. *Neuroimage*, *180*, 68–77. https://doi.org/10.1016/j.neuroimage.2017.06.061

Wang, A., Ramaswamy, V. V., & Russakovsky, O. (2022). Towards intersectionality in machine learning: Including more identities, handling underrepresentation, and performing evaluation. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency* (FAccT 2022), 21–24 June 2022, Seoul, Korea (pp. 336–349). ACM. https://doi.org/10.1145/3531146.3533101

Wang, W., Dong, Y., Liu, X., Bai, Y., & Zhang, L. (2020). The effect of parents' education on the academic and non-cognitive outcomes of their children: Evidence from China. *Children and Youth Services Review*, *117*, 105307. https://doi.org/10.1016/j.childyouth.2020.105307

Zambrano, A. F., Zhang, J., & Baker, R. S. (2024). Investigating algorithmic bias on Bayesian knowledge tracing and carelessness detectors. In *Proceedings of the 14th International Conference on Learning Analytics and Knowledge* (LAK 2024), 18–22 March 2024, Tokyo, Japan (pp. 349–359). ACM. https://doi.org/10.1145/3636555.3636890

Zhang, J., Andres, J. M. A. L., Hutt, S., Baker, R. S., Ocumpaugh, J., Nasiar, N., Mills, C., Brooks, J., Sethuaman, S., & Young, T. (2022). Using machine learning to detect smart model cognitive operations in mathematical problem-solving process. *Journal of Educational Data Mining*, *14*(3), 76–108. https://doi.org/10.5281/zenodo.7304763

Zollanvari, A., Kizilirmak, R. C., Kho, Y. H., & Hernández-Torrano, D. (2017). Predicting students' GPA and developing intervention strategies based on self-regulatory learning behaviors. *IEEE Access*, *5*, 23792–23802. https://doi.org/10.1109/ACCESS.2017.2740980