Hierarchical Dependencies in Classroom Settings Influence Algorithmic Bias Metrics

CLARA BELITZ, School of Information Sciences, University of Illinois Urbana-Champaign, USA
HAEJIN LEE, School of Information Sciences, University of Illinois Urbana-Champaign, USA
NIDHI NASIAR, Graduate School of Education, University of Pennsylvania, USA
STEPHEN E. FANCSALI, Carnegie Learning, Inc., USA
STEVE RITTER, Carnegie Learning, Inc., USA
HUSNI ALMOUBAYYED, Carnegie Learning, Inc., USA
RYAN S. BAKER, Graduate School of Education, University of Pennsylvania, USA
JACLYN OCUMPAUGH, Graduate School of Education, University of Pennsylvania, USA
NIGEL BOSCH, School of Information Sciences and Department of Educational Psychology, University of Illinois Urbana-Champaign, USA

Measuring algorithmic bias in machine learning has historically focused on statistical inequalities pertaining to specific groups. However, the most common metrics (i.e., those focused on individual- or group-conditioned error rates) are not currently well-suited to educational settings because they assume that each individual observation is independent from the others. This is not statistically appropriate when studying certain common educational outcomes, because such metrics cannot account for the relationship between students in classrooms or multiple observations per student across an academic year. In this paper, we present novel adaptations of algorithmic bias measurements for regression for both independent and nested data structures. Using hierarchical linear models, we rigorously measure algorithmic bias in a machine learning model of the relationship between student engagement in an intelligent tutoring system and year-end standardized test scores. We conclude that classroom-level influences had a small but significant effect on models. Examining significance with hierarchical linear models helps determine which inequalities in educational settings might be explained by small sample sizes rather than systematic differences.

CCS Concepts: • **Applied computing** → **Interactive learning environments**; **E-learning**; • **General and reference** → *Metrics*.

Additional Key Words and Phrases: Interactive learning environments, Algorithmic bias, Intelligent tutoring systems, Predictive analytics

ACM Reference Format:
Clara Belitz, HaeJin Lee, Nidhi Nasiar, Stephen E. Fancsali, Steve Ritter, Husni Almoubayyed, Ryan S. Baker, Jaclyn Ocumpaugh, and Nigel Bosch. 2024. Hierarchical Dependencies in Classroom Settings Influence Algorithmic Bias Metrics. In *The 14th Learning Analytics and Knowledge Conference (LAK '24), March 18–22, 2024, Kyoto, Japan.* ACM, New York, NY, USA, 15 pages. https://doi.org/10.1145/3636555.3636869

1 INTRODUCTION

Measuring algorithmic bias in machine learning has historically focused on statistical inequalities pertaining to specific groups [26]. Biased outcomes are generally quantified as some difference

LAK '24, March 18–22, 2024, Kyoto, Japan

ACM ISBN 979-8-4007-1618-8/24/03…$15.00
https://doi.org/10.1145/3636555.3636869

in the predictions between different groups. Depending on the equity concerns, the difference of interest can relate to general accuracy or more specific trends in false (or true) positives or negatives [26]. However, the most common metrics (i.e., those focused on individual- or group-conditioned error rates) [8, 17, 25] are not currently well-suited to educational settings because they assume that each individual observation is independent from the others. This is not statistically appropriate when studying educational outcomes, because such metrics cannot account for the relationship between students in classrooms or multiple observations per student across an academic year.

The nested relationships we see in educational contexts can, however, be modeled statistically by hierarchical linear models (HLMs) [35], which can handle multiple levels of hierarchy as long as there is enough variance and data across each level. In general, mixed-effect linear models allow us to do multilevel analysis when independence assumptions are violated, as is the case for the data in this research [37]. HLMs can also account for the fact that groups influence individuals and individuals influence groups. For example, it is possible to specify a model where level 1 is micro (individuals, such as students), level 2 is secondary (groups, such as classrooms), and level 3 is macro (populations, such as school districts).

In this paper, we use HLMs to adapt algorithmic bias metrics in two novel ways. The first maps four common classification definitions to regression models. The second uses these regression adaptations in hierarchical settings. This work allows us to rigorously measure bias in a machine learning model of the relationship between student engagement in an intelligent tutoring system and year-end standardized test scores.

We examine this problem within the context of MATHia [32], an intelligent tutoring system (ITS) that is widely used in US middle and high schools. Like other intelligent tutoring systems that seek to support learners on an individualized basis [27, 36], MATHia uses machine learning models to serve students questions based on predictions of content mastery [33]. These individualized mastery predictions are built from information extracted from records of students' interactions with the software, and like other ITSs, MATHia is used by large, diverse populations of students that require personalization to address a wide variety of learning needs [20]. The underlying algorithms in these technologies, however, may not be equally effective for all learners—hence the need for analytic approaches to understand the relationships between students' experiences and outcomes, and to measure potential biases in those analyses [7, 18].

In this paper, we explore these issues of algorithmic bias using machine-learned models of gaming the system. "Gaming" occurs in ITSs when students attempt to move through an ITS through exploitation of properties of the system rather than by learning the material [6]. Examples of gaming including trying successive numbers until getting a correct response or repeatedly requesting hints until the answer is provided. This behavior is associated with both lower end-of-year and standardized exam scores, as well as a lower probability of college attendance in the future [3, 15, 29, 34]. Previous work has shown that ITSs improve test scores, though not uniformly across different testing environments [23]. Year-end, state-level standardized tests cannot capture all kinds of success, but they are frequently used for performance prediction in learning analytics settings [16, 29, 30]. Given the relationship between gaming, ITS usage, and exam scores, we use year-end standardized mathematics test scores to measure student performance in relation to the MATHia for this research.

The goal of this work is not to declare certain students as likely to fail but rather to notice where our predictions are misaligned with reality or where they expose a potential systemic failure to support specific groups of students. Our research questions are:

(1) What is the relationship between gaming the system in MATHia (over the year) and standardized mathematics test scores administered at the end of the 2021-2022 academic year?

(2) What kinds of bias, with respect to student race and sex, exist in a machine learning model that uses gaming behaviors to predict standardized test scores?
   (a) What are the best methods for measuring this bias given the hierarchical nature of the dataset, which includes classroom level influences (e.g., teacher characteristics)?

2 METHOD

In this study, we explore possible differences in how students' demographics (as reported within school district internal data) map to a model of the relationship between gaming the system and end-of-year standardized exam scores. We do so by first constructing a machine-learned detector of gaming the system from 121 students who used MATHia during the 2020–2021 and 2021–2022 school years, applying that detector to measure gaming the system behaviors for 5,856 MATHia students from the 2021–2022 school year, and constructing a machine-learned model that predicts end-of-year outcomes from features derived from gaming the system behaviors.

2.1 Learning Context

Our study examines data from middle and high school students in a small city in the Northeast United States who used MATHia as part of their regular mathematics instruction for grades 6–8 mathematics, algebra, and geometry. MATHia is an ITS which asks students to work in two different types of workspaces: concept builders or mastery workspaces. Concept builders present a fixed sequence of questions associated with content intended to build conceptual understanding and have no fine-grained knowledge components (KCs). Mastery workspaces present multi-step problems on which students demonstrate mastery of KCs, which is measured using Corbett and Anderson's Bayesian Knowledge Tracing (BKT) [13]. Both types of workspaces assist students with context-sensitive hints and just-in-time feedback [2]. All actions are therefore associated with workspace features, which are logged in DataShop format [22], with customized categories for MATHia-specific features.

2.2 Student Demographics Accounted for in this Study

Demographic and education-related categories for each student were provided by the school district, which collects this data as part of their standard practices. These data were shared with the researchers according to an established data-sharing agreement. For this district, demographic data included race, Hispanic ethnicity status, gender, age, grade-level, English Language Learner (ELL) status, Individualized Education Plan (IEP) status (i.e., students who are legally entitled to additional learning supports), and eligibility for free or reduced price lunch (i.e., a proxy for socioeconomic status).

Of the 5,856 students in the test-score data set, 47.0% were labeled female, 52.9% were labeled male, and 0.12% were labeled nonbinary. Race group labels were 60.6% African American; 18.0% Hispanic; 14.2% White; 4.7% Multi-Race, Non-Hispanic; 2.3% Asian; 0.24% Native Hawaiian and Pacific Islander; and 0.07% Native American students. 90.8% of students were in grades 6 through 10. In this paper, we used "African American" to match the school district's terminology, which may not reflect how students would describe themselves.

Additionally, 22.1% of students were English Language Learners (ELLs). Given the racial distribution of the school district in general, the total population of ELL students is 72.8% African American and 20.7% Hispanic, with the remaining students spread across the other race groups. This school district is of particular interest for measuring bias for its majority-Black population, many of whom are first- or second-generation immigrants to the United States, representing distinct Caribbean and West African communities. The ELL population of this school district, even when racially similar, therefore do not always share the same home languages.

Table 1. Descriptive attributes for students in the demographic dataset, including both categorical and binary features.

Attribute	Potential Values
Grade	6–12
Age	11–19
Gender	Male; Female; Nonbinary
Race	African American; Asian; Hispanic; White; Native American; Native Hawaiian, Pacific Islander; Multi-Race, Non-Hispanic
Hispanic Ethnicity	Y, N
Special Education	Y, N
English Language Learner	Y, N
Economically Disadvantaged	Y, N

Some of these demographic attributes serve as the dimensions across which we calculated bias in this study.

2.3 Training Data for Gaming Detectors

We generated reference labels for the gaming detectors using text replays, following the method established in [4]. To label student interactions, log data were first segmented into sequential clips. We defined clips as either 8 actions or 20 seconds in duration, whichever came first. Each sequence included the student ID, timestamps for actions, problem details, student input, relevant knowledge/skill metrics, system evaluations, and the determined outcome (e.g., correct, misconception, wrong answer, hint requests). These sequences were then analyzed for "gaming the system" behavior that included students quickly and repeatedly asking for hints without spending time to read or assimilate the help and attempt the question again [1], using repeated hints to get the correct answer from the system [4], entering a systematic series of answers in quick succession (e.g., 1,2,3,4,...) or selecting every multiple choice answer until the correct answer is identified [4], and trying the same answer for successive questions in a short time period without reading feedback (see [28] for a comprehensive list of the actions classified as gaming behavior). An expert coder, who had previously achieved a Cohen's kappa of 0.62 with another rater to establish inter-rater reliability on the text replay clips of MATHia [24]—which is above the standard 0.6 cut-off for ill-defined/ambiguous constructs like gaming or disengagement—determined whether clips indicated gaming behavior. Using this methodology, the expert coder labeled a random sample of 1,211 coded clips, of which 72 were gaming and 1,095 were not gaming (i.e., 5.94% gaming behavior, which aligns with the usual distribution of gaming [4]). The remaining 44 clips were unclear. As in other research using this method, we did not include unclear clips in the training data [24]. As such, the data used to train the gaming detectors consisted of 1,167 observations across 121 students in two academic years. The training data had an adjusted gaming rate of 6.17%.

Though we had gaming data for two academic years (2020–2021 and 2021–2022) we only received demographic data from the district for the 2021-2022 school year. We did, however, have students' self-reported demographic data for 2020–2021. Some of the 2020–2021 data were therefore unknown, since these self-reports were not standardized by the school district. Overall, the students in the training data were 42.1% female, 38.0% male, and 19.8% unknown. 38.0% of students were labeled African American; 10.7% White; 3.4% Hispanic; 10.7% Multi-Race, Non-Hispanic; 1.7% Asian; and

35.5% unknown. No Native Hawaiian, Pacific Islander or Native American identified students were present in the training data.

2.3.1 Machine Learning for Gaming Detectors. We extracted 162 features for machine learning predictions of gaming the system by applying statistical aggregation functions to the expert-coded clips. Because a clip consists of a sequence of actions, the feature extraction process transformed these sequences into a single row of summary data for each clip. Specifically, these functions included minimum, maximum, quartiles, mean, standard deviation, sum, and count of non-empty values, and they were applied to data from the ITS, including student answer correctness, timing (i.e., time spent per attempt), help requests, and real-time estimates of student knowledge. Further description of the feature engineering process is available in Levin et al. [24]. We rebalanced the training data before training the classifier using Synthetic Minority Oversampling Technique (SMOTE) in order to ensure a uniform balance in class distribution [11].

We trained four classification models on these data, extreme gradient boosting (i.e., XGBoost), a decision tree, a random forest, and a multilayer perceptron neural net [12, 31]. Three of these models are available in *scikit-learn*, a commonly used machine learning package for Python [31]. The fourth, XGBoost, is another popular adaptation of decision trees [12]. These models were chosen to allow comparison with previously published work [24]. We set a maximum depth of 5 for the tree-based models and used log loss for the evaluation metric for XGBoost. The random forest model had the highest performance in terms of area under the receiver operating characteristic curve (AUC ROC = .80) and recall (.67). Though XGBoost had slightly higher precision (0.5, compared to 0.43 for the random forest model), we were more concerned with false negatives, since the overall gaming rate was quite low. As such, the random forest model was selected for subsequent use as our measure of gaming the system.

2.3.2 Distribution of Gaming Detector Predictions Across Demographic Categories. We applied the random forest detector of gaming to the entire dataset of all 5,856 students (6,300,569 clips) to measure gaming the system for each of these students across the academic year (M = 1,075 clips per student, SD = 1,054). The random forest model labeled each clip with a probability.

The overall predicted gaming rate was higher than the labeled data at 21.9%. Asian students had a predicted gaming rate of 15.9%; African American 22.4%; Hispanic 24.1%; Multi-race, non-Hispanic 22.0%; Native American 17.5%; Native Hawaiian and Pacific Islander 17.7%; and White 20.5%. Female students had a predicted gaming rate of 21.0%, male students 23.4%, and nonbinary students 0.54%. Predicted gaming rates are likely too high (relative to the training data) due to the change in class balance from SMOTE, but this does not affect our downstream analysis given that we extract features from the probabilities in a threshold-free manner rather than binarizing them [5].

2.4 Standardized Test Score Prediction

For the state-level standardized test used in our study, the possible scores are scaled to a standardized range of 440 to 560. These scores are grouped into four 30-point ranges by the state, representing broad categories of student performance in relation to grade-level standards for mathematics. Student scores are reported back to not only the school district, but also the student and their family. Students classified as "not meeting expectations" are explicitly slated for additional assistance and/or instruction. Students who partially meet expectations see language suggesting that the school, in consultation with their family, "consider whether" the student needs such additional support, while students who meet expectations are described as "academically on track to succeed." Students classified as "exceeding expectations" are described as having "demonstrated mastery of the subject matter." Note that this definition of mastery, provided by the state which administers

the exam, is different than how mastery is used in MATHia workspaces, where it instead is used to describe students who have met expectations for a particular workspace.

To predict these year-end scores (solely from gaming), we extracted 16 features from the gaming predictions to create a student-level dataset (i.e., one row per student) which described each student's general gaming behaviors across the year. Features consisted of summary statistics of the gaming predictions, such as the number of observations per student; the predicted gaming probability mean, standard deviation, and quartiles. We then used 10-fold cross-validation to fit a machine learning model to these data and predict the standardized math test scores. We considered five models: decision tree, random forest, Extra-Trees (a random forest variant [19]), XGBoost, and linear regression. We tuned hyperparameters via nested cross-validation for the tree-based models, including minimum samples per leaf (1, 2, 3, 8, 16, or 32) and maximum number of features considered for each decision in a tree (proportions of .1, .25, .5, .75, 0.9, or 1.0, the square root of the number of features, or the base-2 log of the number of features).

In all, we generated standardized test score predictions for 3,964 students, because some (n = 1,892) did not have a recorded standardized math test score. The mean score for all students in our dataset was 481.3. Male and female students scored similarly on year-end standardized tests: female students had a mean score of 481.4 (*SD* = 18.9) and male students had a mean score of 481.2 (*SD* = 19.8). No students labeled as nonbinary were present in the dataset containing test scores.

There was more variance in mean test scores by race. African American students had an average score of 479.1 (*SD* = 18.5); Asian students 501.9 (*SD* = 23.3); Hispanic students 479.7 (*SD* = 18.9); White students 488.2 (*SD* = 19.3); Native American 480.0 (*SD* = 26.1); Native Hawaiian and Pacific Islander 483.6 (*SD* = 10.7); and Multi-Race, Non-Hispanic students 481.2 (*SD* = 18.7).

2.5 Measuring Bias

In our study, we adapt the following four commonly used classification metrics [8–10, 14], where TP = true positives, FP = false positives, FN = false negatives, and TN = true negatives.

- Overall Accuracy Equality (OAE): OAE is satisfied when the proportion of correct predictions (both true positives and negatives) made by model is equal across each interest group. Therefore, we say overall accuracy equality is ensured when $\frac{TP+TN}{TP+FP+TN+FN}$ is the same across all groups of interest. OAE looks at differences in predictive accuracy across different groups to try to establish whether these differences are driven by systematic bias in the model or by differences in the underlying data. OAE assumes that true positives and true negatives are equally important, though this assumption is not always true in real-world scenarios [10].
- Statistical Parity (SP): SP measures the bias of a model by ensuring the predicted positive and negative class proportions are equal across every group. Specifically, the ratio of predicted positives $\frac{TP+FP}{TP+FP+TN+FN}$ and predicted negatives $\frac{FN+TN}{TP+FP+TN+FN}$ should be consistent across all groups. SP tries to establish that there are similar predicted rates of success (or failure) across groups. This particular definition has been critiqued because it can lead to forcing dissimilar groups to have similar outcomes, even when this may not be desirable or make sense. [14].
- Conditional Procedure Accuracy Equality (CPA): CPA assesses a model's predictions by ensuring actual positives and negatives are consistent across all groups. Therefore, to meet the conditional procedure accuracy equality, the model's rate of correctly predicting positives $\frac{TP}{TP+FN}$ and negatives $\frac{TN}{TN+FP}$ should be consistent for all groups to meet the conditional procedure accuracy equality. CPA is equivalent to comparing the recall of the model for each group.
- Conditional Use Accuracy Equality (CUA): CUA measures the model's predictions rather than the actual outcomes. Thus, it focuses on ensuring the proportion of all predicted positives

$\frac{TP}{TP+FP}$ and negatives $\frac{TN}{TN+FN}$ are equivalent for all groups. CUA is equivalent to comparing the precision of the model for each group.

Because we are working with a regression problem, we adapted the above definitions into the following analogous formulas:

- Overall Accuracy Equality: OAE is measured with the root mean squared error (RMSE) between the actual and predicted values for the interest group. Variance in accuracy is explained by the category of interest (e.g., gender or race). This metric aims to ensure the predictions are equally accurate across protected groups of interest, but does not differentiate between positive and negative error. A value of 0 indicates equal error across predictions per group.
- Statistical Parity: This metric examines whether the groups of interest receive similar model-predicted outcomes (e.g., test score) on average. We measured SP by taking the mean of predicted values for each group of interest. Statistical parity ensures the means of predicted values are consistent across protected interest groups. This measure lets us see if the model systematically predicts higher or lower for groups. A value of 0 indicates equal predictions of scores per group.
- Conditional Procedure Accuracy Equality: We assessed CPA analogously to the measurement of OAE. However, this metric conditions on the actual values, focusing on instances across groups where the actual outcome (e.g., test score) is above or below a specified threshold. Specifically, this metric allows us to investigate the model's prediction accuracy for specific cases meeting the test score threshold. Thresholding at a specific value is required here, given that we cannot condition easily on a continuous outcome. A value of 0 indicates equal error across predictions per group.
- Conditional Use Accuracy Equality: CUA is also measured similarly to how CPA is measured. However, this metric conditions on the predicted value of the test scores, rather than the true values. We use the same test score threshold used for CPA. A value of 0 again indicates equal error across predictions per group.

For both CPA and CUA, we use thresholds of both 470 (the cutoff for partially meeting expectations) and 500 (the cutoff for meeting expectations). These thresholds allow us to compare how our model performs for students seen as needing extra support versus students seen as succeeding. High school students in this state must score 486 or higher on their mathematics exam to graduate.

Our study deals with datasets which violate independence assumptions; we therefore again adapt the regression bias metrics to account for these violations. HLMs can be more flexible than traditional linear models and can support the kinds of complicated data structures which occur in the real world [35]. In this study, we use a random intercept to model standardized test score per classroom, allowing us to capture the teacher effect on classrooms [37]. By allowing each classroom to have a unique intercept, we can account for baseline differences between classrooms to isolate the effect that categories like race or sex have on test scores. The 3,964 students included in our dataset represented 267 classrooms (M = 14.9 students per classroom, SD = 8.5). We fit both a traditional and a hierarchical model to compare measures of bias that do and do not account for variance in this secondary relationship.

For categories with more than two groups, we calculate an overall score for each metric by taking the difference between the maximum and minimum predicted values for the groups. Doing this calculation allows us to have one number per metric as well as quantify the extent of potential bias in the model's predictions across these groups. This is true for both the traditional and hierarchical metrics.

Table 2. Regression-based bias metrics, both traditional and hierarchical, for student sex. All results are significant ($\alpha = 0.05$) in the HLMs other than those marked with $^{\#}$.

	OAE	SP	CPA ≥ 470	CPA $< 470^{b}$	CPA ≥ 500	CPA < 500	CUA ≥ 470	CUA $< 470^{a,b}$	CUA $\geq 500^{a,b}$	CUA < 500
Hierarchical	1.214	2.176	2.273	$0.788^{\#}$	4.476	$0.289^{\#}$	1.212	1.442	$0.717^{\#}$	1.242
Traditional	1.148	2.951	1.928	0.634	4.048	0.330	1.142	1.444	0.671	1.197

3 RESULTS

Random forest had the highest R^2 (0.494) score of the models by a small margin; linear regression had the lowest R^2 of 0.464. The random forest model also had the smallest RMSE of the trained models: 13.8. We therefore used the predictions from the random forest model for our bias analysis.

We initially evaluated the results with both traditional statistical analysis and algorithmic bias metrics. We visualized predicted test scores compared to actual year-end scores, categorizing students by race and sex (Figure 1). Overall, the model predicted that male students had lower test scores than female students: female students had a mean predicted score of 482.8; male students had a mean predicted score of 479.9. The mean predicted score for male students was 1.3 points lower than the true mean Asian students had the highest predicted mean score (493.4) while Hispanic students had the lowest predicted mean score (479.2). The remaining predicted mean scores were African American: 480.8; White: 483.8; Native American: 484.7; Native Hawaiian and Pacific Islander: 485.1; and Multi-Race, Non-Hispanic: 480.7.

Female students had better predictive accuracy than male students for OAE, CPA, and CUA across both the hierarchical and traditional metrics (Table 2). When calculating these values across sex, the traditional and hierarchical metrics were largely similar, though in some cases the traditional method overestimated bias (up to 35.6% higher for SP) or underestimated bias (15.2% lower for CPA ≥ 470 and 19.5% lower for CPA ≥ 500).

In the hierarchical models, African American was the baseline (largest) group for race, meaning differences in the model were in relation to African American students rather than an overall mean. Statistical parity was generally quite small across both traditional and hierarchical models, while mean square error rates varied more widely (Table 3).

For the hierarchical models, which race group had the smallest or largest error varied per metric. For example, Asian students had the largest predicted RMSE for CPA and CUA for students scoring, or predicted to score, above 500, while Native American students had the largest RMSE for the other error-based metrics. Native Hawaiian and Pacific Islander students had the smallest predicted error for three of the error-based metrics, while Hispanic students had the smallest predicted error for the other two. White, African American, and multi-race, non-Hispanic students were consistently none of the extremes across metrics. Of note is that zero Native American or Native Hawaiian and Pacific Islander students were predicted to score at or above 500 points, even though they are present in the true scores. As such, we also looked at the threshold for students who partially meet expectations: 470 points. Using this cutoff, Native American students had the largest discrepancy in error rate, leading to a CPA of 28.370, the largest spread in predictive error for any of the error-based metrics.

A benefit of HLMs is their ability to measure statistical significance of differences in the predictive model. For sex, the relatively small differences in predictive error rates for CPA ($score < 470$; $p = .189$), CPA ($score < 500$; $p = .433$), and CUA ($score \geq 500$; $p = .374$) are not significant (Table 2). Measuring statistical significance is additionally useful for examining biases related to race, since there were several groups in these data. Seven of the metrics are changed when we calculate hierarchical bias using only race groups with results that were significantly different than the

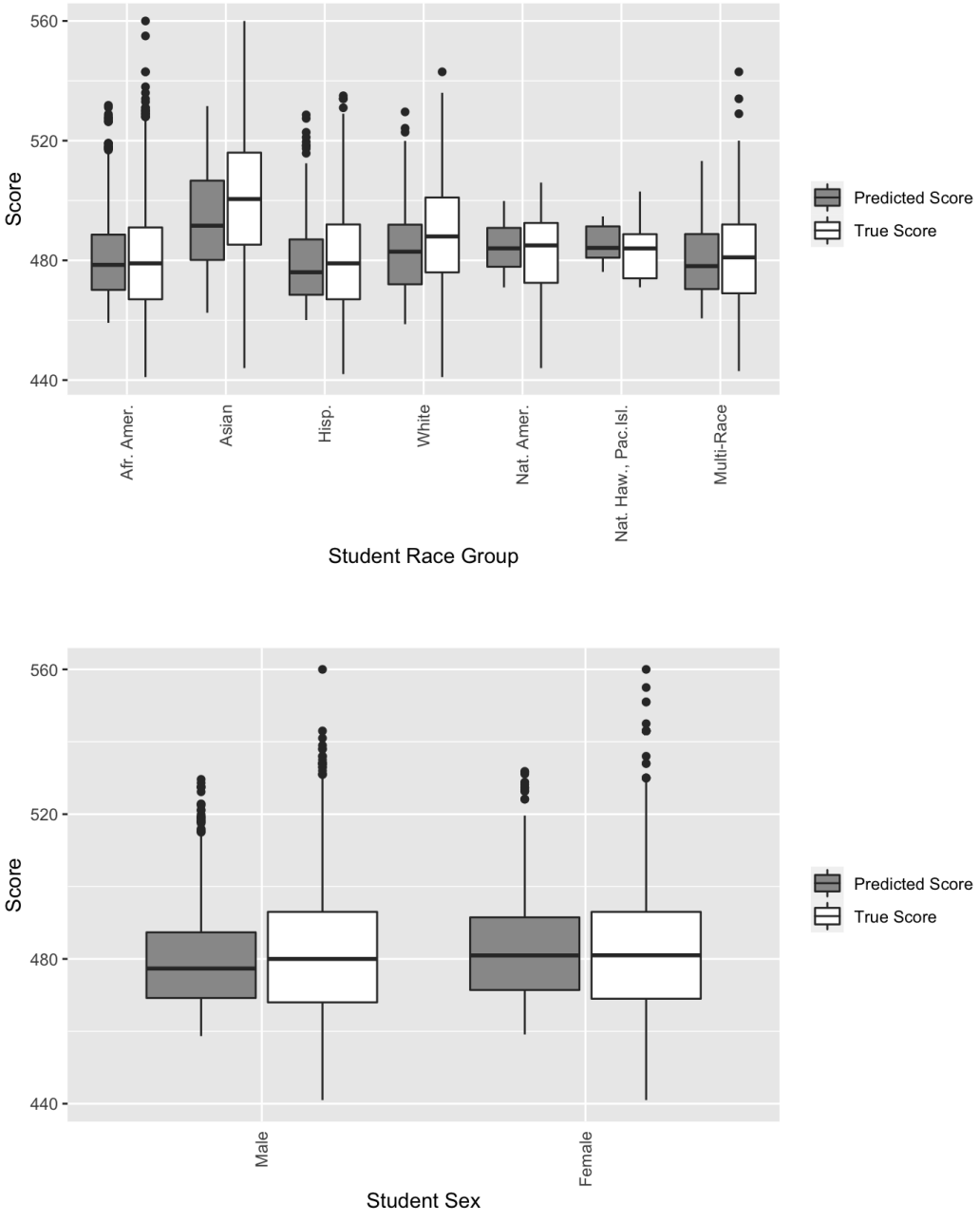

Fig. 1. Boxplots of true and predicted (RF model) mean test scores for students by race and sex.

Table 3. Regression-based bias metrics, both traditional and hierarchical, for student race. *a* indicates that no Native American students were present in this group. *b* indicates that no or Native Hawaiian and Pacific Islander students were present in this group.

	OAE	**SP**	**CPA** ≥ 470	**CPA** $< 470^{b}$	**CPA** ≥ 500	**CPA** < 500	**CUA** ≥ 470	**CUA** $< 470^{a,b}$	**CUA** $\geq 500^{a,b}$	**CUA** < 500
Hierarchical	16.224	11.969	5.952	28.370	12.271	19.918	16.031	1.839	6.612	10.658
Traditional	15.103	14.252	6.919	28.924	12.578	19.092	15.107	1.603	6.840	15.107

Table 4. Hierarchical bias metrics for student race using only statistically significant group differences ($\alpha = 0.05$). *a* indicates that no Native American students were present in this group. *b* indicates that no or Native Hawaiian and Pacific Islander students were present in this group.

OAE	**SP**	**CPA** ≥ 470	**CPA** $< 470^{b}$	**CPA** ≥ 500	**CPA** < 500	**CUA** ≥ 470	**CUA** $< 470^{a,b}$	**CUA** $\geq 500^{a,b}$	**CUA** < 500
10.084	11.969	2.688	28.370	1.900	13.844	9.936	0	4.288	10.058

baseline (Table 4). OAE, the second largest difference of the error-based metrics for race, only had a statistically significant difference in the predictive error rate (compared with the baseline error) for Native American students (RMSE = 24.0, $p = .004$). The original smallest predictive error, for Native Hawaiian and Pacific Islander students (RMSE = 7.7), was not significant ($p = .162$). If we recalculate OAE using only significant differences—that is, using the statistically significant baseline RMSE of 13.9 rather than the smallest observed RMSE—it was reduced from 16.2 to 10.1. SP for race was, however, unchanged by accounting for statistical significance, since the largest and smallest differences in predicted scores were both significant ($p = .043$ and $p = .021$, respectively), with a nearly 12-point spread between the means of the highest and lowest predicted scores. CUA (< 470) had no statistically different error rates across different race groups (the smallest p-value was .071, while the largest was .852).

For English language learner status, using HLMs, error rates were not statistically significantly different across conditional definitions nor for overall accuracy equality. SP was, however, five points lower for ELL students. We can likely conclude the predictive model reflects increased difficulties in test-taking for students who are still learning English, rather than bias in the predictive model. This is not a surprising result, given that students are being instructed in a new language, and so allows us to conclude that ELL students likely need additional academic support, but are not gaming the system at very different rates than native English speakers nor was our model systematically mis-predicting their expected test scores.

4 DISCUSSION

4.1 RQ1: What is the relationship between gaming the system in MATHia (over the year) and standardized mathematics test scores administered at the end of the 2021-2022 academic year?

Male and female students scored nearly identically on year-end standardized tests, but when gaming the system was used to predict these scores, male students were predicted to do nearly 3 points worse on average (female: 482.8; male: 479.9). A spread of 3 points can be enough to place students into different categories of achievement. For example, 319 students (over 8% of students) scored between 497 and 503 points (inclusive).

More male students were present in the data than female students, so this is not likely accounted for by basic differences in representation in the training data. Features were extracted from these data without any demographic information, nor anything that would arguably be a proxy for demographic information, and yet they are not well calibrated for all different groups. Male students were predicted to game the system more frequently than female students, perhaps accounting for their lower predicted scores. The three race groups with the lowest predicted gaming rates (Asian, Native American, Native Hawaiian and Pacific Islander) had the highest predicted mean test scores. In the actual test scores, however, White students were among the three highest-scoring groups, rather than Native American students. These findings may be artifacts of focusing solely on the relationship between gaming and test score outcomes, though many other factors also contribute to outcomes.

4.2 RQ2: What kinds of bias, with respect to student race and sex, exist in a machine learning model that uses gaming behaviors to predict standardized test scores? What are the best methods for measuring this bias given the hierarchical nature of the dataset?

The differences between the traditional metrics and the hierarchical difference were not often large, though there are noticeable exceptions for a small number of metrics for race (SP, CUA<500). We can likely conclude that classroom-level influences had a smaller effect on models than we hypothesized in this school district. However, the HLM approach we propose is still more comprehensive in its combination of statistical rigor with established algorithmic bias metrics. It allows us to observe both the nested nature of the data as well as the differences in predictive accuracy across groups, which are both relevant to understanding the relationship of predictive bias to other potential sources of bias and difference in educational outcomes.

Perhaps most important is that HLMs include valid measures of statistical significance that are important for making inferential claims about bias in machine learning models. While traditional measures of algorithmic bias could likely be adapted to offer significance as well, they have not done so previously. Moreover, adding a measure of statistical significance to non-hierarchical metrics would violate assumptions of non-independence in cases where there is reason to believe biases might be correlated between students (e.g., within a classroom). Examining significance helps distinguish conclusions, such as which inequalities might be explained by small sample sizes rather than true difference. Statistical validity can then inform our future work and conclusions. For example, conditional procedure accuracy by sex, conditioned on students who met or exceeded expectations, was the largest difference of any of the error-based metrics, and was significant in the HLMs. The smaller inequality in accuracy for students who did not meet expectations was not significant. This is useful for developing next steps; we can validate the discrepancy in predicted scores for students who are generally seen as "succeeding" and investigate why this occurs. We used test scores as one measure of student outcome; false negatives mean the predictive model failed to notice students who may not meet performance thresholds. This particular model, however, has no significant difference in accuracy with respect to false negatives (CPA), when conditioned on sex, for both thresholds used.

For race, the recalibrated OAE score was similar to the predicted error rates for students who scored or were predicted to score below 500. Because the predictive error rates for other race groups were not statistically significant, we can interpret this as having achieved OAE for most groups, but not for Native American students. The number of Native American students in our dataset was quite small, however, so this finding should be considered tentative until replicated. The recalibrated CPA and CUA scores for students scoring at or above 500 were also much smaller. The recalibrated CUA score for students scoring below 470 was 0, as there were no significant

differences in error rates, indicating equal predicted error rates across all race groups. These results help us to focus on a pattern of increased predictive error for students scoring between 470 and 500 (i.e., those classified as partially meeting expectations) on these exams.

This predictive discrepancy (and perhaps others seen here) is likely partially attributable to (a lack of) representation in the training data. African American students were the most represented in the dataset and in the school district and had a mean predictive value closest to the true mean. Native American students only represented 0.07% of the students, and our predictive accuracy for this group was poor, including that none of these students were predicted to meet or exceed expectations despite their having done so in the actual exam. Similarly, Asian students represented 2.3% of students, and their scores were underpredicted compared to the baseline, likely representing an issue with predictive regression to the mean of the student population. This could be remedied with additional, fairness-focused, preprocessing techniques which allow recalibration of the training data to eliminate algorithmic bias, even beyond SMOTE [21]. Future work might also seek to include these under-represented groups in the training data using an explicit demographic sampling strategy.

4.3 Limitations

Our values for certain demographic categories for this work are incomplete, representing the school district's current attempts to classify students, rather than the true possible range. For example, gender was reported as female, male, and nonbinary. While a third gender category is useful, these three categories do not include whether a student is cisgender or transgender. Moreover, nonbinary may not fully capture the range of gender expressions in this data, and it is possible that even students who might identify with the term may not feel comfortable or safe reporting it. In the end, the small number of nonbinary students—none of whom had reported test scores—meant we were not able to include them in our analysis.

Similarly, the racial groups provided do not capture the full breadth of possible racial identities and ethnicities. For example, "Multi-Race, Non-Hispanic" does not allow us to know which races are represented in this group. Similarly, this particular school district's Black population has a variety of ethnic backgrounds, including distinct immigrant communities with distinct language practices, which is not reflected in the label of "African American.". Also of note is that this dataset includes Hispanic as both a race group and ethnicity, with the Hispanic Ethnicity flag perfectly correlating to whether the student's listed race group was Hispanic. Relatedly, students from Brazil (a sizeable group in this district) would not be identified by the school as Hispanic, even though they are Latin American.

There are related limitations in the training data itself. The first is that the data used to train the gaming detector was not directly representative of the total student population. Though a demonstration of generality, this could be a source of bias and inaccuracy in our test score predictions. The training data was also collected during both the 2020–2021 and 2021–2022 school years, though the testing data was from only the 2021–2022 school year. Students experienced varying amounts of in person and remote instruction across these two years due to COVID-19, but we saw similar rates of gaming in our training data across both years. Machine learning models inherently draw predictions on incomplete models of the world, which underlines the importance of considering whether these models create biased predictive outcomes.

5 CONCLUSION

The development of hierarchical bias metrics allows the learning analytics community to better understand where our predictive models—and the analyses and adaptive learning platforms they enable—fall short of equity goals. When a predictive model is not well calibrated for distinct groups,

we risk neglecting those already at the margins. Our novel contribution of algorithmic bias metrics for regression and HLMs allows more accurate exploration of biases in situations where not all predictions are independent of each other, which is also of interest to a broad machine learning community. Hierarchical structure, modeled here in terms of students within classrooms, can also account for common scenarios such as multiple predictions per student (which are statistically dependent) while measuring bias and providing statistical significance testing. Future work can, for example, model the introduction of biases at each level of a process (e.g., during gaming the system predictions made many times per student in this study). This research used only gaming indicators to predict year-end test scores. Future work could also improve these models by incorporating additional predictive factors and relevant dimensions in which biases could occur, like finer-grained ethnicity data, to better identify which students may not be seeing the full benefits of using an ITS.

ACKNOWLEDGMENTS

This research was supported by NSF grant no. 2000638. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

REFERENCES

[1] Vincent Aleven and Kenneth R Koedinger. 2001. Investigations into help seeking and learning with a cognitive tutor. In *Papers of the AIED-2001 Workshop on Help Provision and Help Seeking in Interactive Learning Environments*. Springer Cham, San Antonio, TX, 47–58.

[2] Husni Almoubayyed, Stephen E. Fancsali, and Steve Ritter. 2023. Instruction-embedded assessment for reading ability in adaptive mathematics software. In *LAK23: 13th International Learning Analytics and Knowledge Conference*. ACM, Arlington TX USA, 366–377. https://doi.org/10.1145/3576050.3576105

[3] Joshi Ambarish, Stephen E. Fancsali, Steven Ritter, Tristan Nixon, and Susan R. Berman. 2014. Generalizing and extending a predictive model for standardized test scores based on Cognitive Tutor interactions. In *Proceedings of the 7th International Conference on Educational Data Mining*. Educational Data Mining Society (IEDMS), Online, 369–370.

[4] Ryan Baker and Adriana de Carvalho. 2008. Labeling student behavior faster and more precisely with text replays. In *Proceedings of the 1st International Conference on Educational Data Mining*. Educational Data Mining Society (IEDMS), Montréal, Canada, 38–47.

[5] Ryan S. Baker. 2023. Big Data and Education. 7th Edition.

[6] Ryan Shaun Baker, Albert T. Corbett, Kenneth R. Koedinger, and Angela Z. Wagner. 2004. Off-Task behavior in the Cognitive Tutor classroom: When students "Game the System". In *Proceedings of the SIGCHI conference on Human factors in computing systems*, Vol. 6. Association for Computing Machinery, Vienna, Austria, 383–390.

[7] Ryan S. Baker and Aaron Hawn. 2021. Algorithmic bias in education. *International Journal of Artificial Intelligence in Education* 32 (Nov. 2021), 1052–1092. https://doi.org/10.1007/s40593-021-00285-9

[8] Solon Barocas, Moritz Hardt, and Arvind Narayanan. 2019. *Fairness and Machine Learning*. fairmlbook.org, Online. http://www.fairmlbook.org

[9] Clara Belitz, Lan Jiang, and Nigel Bosch. 2021. Automating procedurally fair feature selection in machine learning. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*. Association for Computing Machinery, New York, NY, 379–389. https://doi.org/10.1145/3461702.3462585

[10] Richard Berk, Hoda Heidari, Shahin Jabbari, Michael Kearns, and Aaron Roth. 2021. Fairness in criminal justice risk assessments: The state of the art. *Sociological Methods & Research* 50, 1 (2021), 3–44. https://doi.org/10.1177/0049124118782533

[11] Nitesh V. Chawla, Kevin W. Bowyer, Lawrence O. Hall, and W. Philip Kegelmeyer. 2002. SMOTE: Synthetic Minority Over-sampling Technique. *Journal of Artificial Intelligence Research* 16 (June 2002), 321–357. https://doi.org/10.1613/jair.953

[12] Tianqi Chen and Carlos Guestrin. 2016. XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. Association for Computing Machinery, San Francisco, CA, USA, 785–794. https://doi.org/10.1145/2939672.2939785

[13] Albert T. Corbett and John R. Anderson. 1994. Knowledge tracing: Modeling the acquisition of procedural knowledge. *User Modeling and User-Adapted Interaction* 4, 4 (Dec. 1994), 253–278. https://doi.org/10.1007/BF01099821 Company: Springer Distributor: Springer Institution: Springer Label: Springer Number: 4 Publisher: Kluwer Academic Publishers.

[14] Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. 2012. Fairness through awareness. In *Proceedings of the 3rd Innovations in Theoretical Computer Science Conference (ITCS '12)*. Association for Computing Machinery, New York, NY, USA, 214–226. https://doi.org/10.1145/2090236.2090255

[15] Stephen E. Fancsali. 2014. Causal discovery with models: Behavior, affect, and learning in Cognitive Tutor Algebra. In *Proceedings of the 7th International Conference on Educational Data Mining*. Educational Data Mining Society (IEDMS), Online, 28–35.

[16] Mingyu Feng, Neil Heffernan, and Kenneth Koedinger. 2009. Addressing the assessment challenge with an online system that tutors as it assesses. *User Modeling and User-Adapted Interaction* 19, 3 (Aug. 2009), 243–266. https://doi.org/10.1007/s11257-009-9063-7

[17] Sorelle A. Friedler, Carlos Scheidegger, and Suresh Venkatasubramanian. 2021. The (Im)possibility of fairness: Different value systems require different mechanisms for fair decision making. *Commun. ACM* 64, 4 (April 2021), 136–143. https://doi.org/10.1145/3433949

[18] Josh Gardner, Christopher Brooks, and Ryan Baker. 2019. Evaluating the fairness of predictive student models through slicing analysis. In *Proceedings of the 9th International Conference on Learning Analytics & Knowledge (LAK19)*. Association for Computing Machinery, New York, NY, USA, 225–234. https://doi.org/10.1145/3303772.3303791

[19] Pierre Geurts, Damien Ernst, and Louis Wehenkel. 2006. Extremely randomized trees. *Machine Learning* 63, 1 (April 2006), 3–42. https://doi.org/10.1007/s10994-006-6226-1

[20] Neil T. Heffernan and Cristina Lindquist Heffernan. 2014. The ASSISTments ecosystem: Building a platform that brings scientists and teachers together for minimally invasive research on human learning and teaching. *International Journal of Artificial Intelligence in Education* 24, 4 (Dec. 2014), 470–497. https://doi.org/10.1007/s40593-014-0024-x

[21] Faisal Kamiran and Toon Calders. 2012. Data preprocessing techniques for classification without discrimination. *Knowledge and Information Systems* 33, 1 (Oct. 2012), 1–33. https://doi.org/10.1007/s10115-011-0463-8

[22] Kenneth R. Koedinger, Elizabeth A. McLaughlin, and Neil T. Heffernan. 2010. A quasi-experimental evaluation of an on-line formative assessment and tutoring system. *Journal of Educational Computing Research* 43, 4 (Dec. 2010), 489–510. https://doi.org/10.2190/EC.43.4.d Publisher: SAGE Publications Inc.

[23] James A. Kulik and J.D. Fletcher. 2016. Effectiveness of intelligent tutoring systems: A meta-analytic review. *Review of Educational Research* 86, 1 (March 2016), 42–78. http://journals.sagepub.com/doi/full/10.3102/0034654315581420

[24] Nathan Levin, Ryan S. Baker, Nidhi Nasiar, Stephen Fancsali, and Stephen Hutt. 2022. Evaluating gaming detector model robustness over time. In *Proceedings of the 15th International Conference on Educational Data Mining*. International Educational Data Mining Society, Durham, UK, 398–405. https://doi.org/10.5281/ZENODO.6852962 Publisher: Zenodo.

[25] Kristian Lum, Yunfeng Zhang, and Amanda Bower. 2022. De-biasing "bias" measurement. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency (FAccT '22)*. Association for Computing Machinery, New York, NY, USA, 379–389. https://doi.org/10.1145/3531146.3533105

[26] Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. 2021. A survey on bias and fairness in machine learning. *Comput. Surveys* 54, 6 (July 2021), 115:1–115:35. https://doi.org/10.1145/3457607

[27] Roger Nkambou, Jacqueline Bourdeau, Riichiro Mizoguchi, and Janusz Kacprzyk (Eds.). 2010. *Advances in Intelligent Tutoring Systems*. Studies in Computational Intelligence, Vol. 308. Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-642-14363-2

[28] Luc Paquette and Ryan S. Baker. 2019. Comparing machine learning to knowledge engineering for student behavior modeling: A case study in gaming the system. *Interactive Learning Environments* 27, 5-6 (Aug. 2019), 585–597. https://doi.org/10.1080/10494820.2019.1610450

[29] Zachary A. Pardos, Ryan S. J. D. Baker, Maria O. C. Z. San Pedro, Sujith M. Gowda, and Supreeth M. Gowda. 2014. Affective states and state tests: Investigating how affect and engagement during the school year predict end-of-year learning outcomes. *Journal of Learning Analytics* 1, 1 (2014), 107–128. https://eric.ed.gov/?id=EJ1127034 Publisher: Society for Learning Analytics Research ERIC Number: EJ1127034.

[30] Zachary A. Pardos, Qing Yang Wang, and Shubhendu Trivedi. 2012. The real world significance of performance prediction. In *Proceedings of the 5th International Conference on Educational Data Mining*. International Educational Data Mining Society, Chania, Greece, 192–195. https://eric.ed.gov/?id=ED537229 Publication Title: International Educational Data Mining Society ERIC Number: ED537229.

[31] Fabian Pedregosa, Gael Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, and David Cournapeau. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research* 12, 85 (2011), 2825–2830.

[32] Steven Ritter, John R. Anderson, Kenneth R. Koedinger, and Albert Corbett. 2007. Cognitive Tutor: Applied research in mathematics education. *Psychonomic Bulletin & Review* 14, 2 (April 2007), 249–255. https://doi.org/10.3758/BF03194060

[33] Steven Ritter and Stephen E. Fancsali. 2016. MATHia X: The next generation Cognitive Tutor. In *Proceedings of the EDM 2016 Workshops and Tutorials*. Raleigh, North Carolina, 624–625.

[34] Maria O. C. Z. San Pedro, Jaclyn L. Ocumpaugh, Ryan S. Baker, and Neil T. Heffernan. 2014. Predicting STEM and non-STEM college major enrollment from middle school interaction with mathematics educational software. In *Proceedings of the 7th International Conference on Educational Data Mining*. International Educational Data Mining Society, Online, 276–279.

[35] Tom A.B. Snijders and Roel Bosker. 2012. *Multilevel Analysis: An Introduction to Basic and Advanced Multilevel Modeling* (2nd edition ed.). Sage, Thousand Oaks, CA.

[36] Kurt VanLehn. 2011. The relative effectiveness of human tutoring, intelligent tutoring systems, and other tutoring systems. *Educational Psychologist* 46, 4 (2011), 197–221. http://www.tandfonline.com/doi/abs/10.1080/00461520.2011.611369

[37] William J. Webster, Robert L. Mendro, Timothy H. Orsak, and Dash Weerasinghe. 1998. An application of hierarchical linear modeling to the estimation of school and teacher effect. In *1998 Annual Meeting Program*. American Educational Research Association, San Diego, CA, 33. https://eric.ed.gov/?id=ED424300 ERIC Number: ED424300.