Jiang, B. et al. (Eds.) (2025). Proceedings of the 33rd International Conference on Computers in Education. Asia-Pacific Society for Computers in Education

Investigating Algorithmic Bias in Affect Detectors with Constructed Categories of Student Identity

Nidhi NASIAR[a*], Clara BELITZ[b], HaeJin LEE[b], Frank STINAR[b], Ryan S. BAKER[c], Jaclyn OCUMPAUGH[d], Stephen E. FANCSALI[e], Steve RITTER[e], Nigel BOSCH[e]
[a]*University of Pennsylvania, USA*
[b]*University of Illinois Urbana–Champaign, USA*
[c]*Adelaide University, Australia*
[d]*University of Houston, USA*
[e]*Carnegie Learning, Inc., USA*
*nasiar@upenn.edu

Abstract: Algorithmic bias research often evaluates models in terms of traditional demographic categories (e.g., U.S. Census), but these categories may not capture nuanced, context-dependent identities relevant to learning. This study evaluates four affect detectors (boredom, confusion, engaged concentration, and frustration) developed for an adaptive math learning system. Metrics for algorithmic fairness (AUC, weighted F1, MADD) show subgroup differences across several categories that emerged from a free-response social identity survey (Twenty Statements Test; TST), including both those that mirror demographic categories (i.e., race and gender) as well as novel categories (i.e., Learner Identity, Interpersonal Style, and Sense of Competence). For demographic categories, the confusion detector performs better for boys than for girls and underperforms for West African students. Among novel categories, biases are found related to learner identity (boredom, engaged concentration, and confusion) and interpersonal style (confusion), but not for sense of competence. Results highlight the importance of using contextually grounded social identities to evaluate bias.

Keywords: algorithmic bias, affect, fairness, social identity, twenty statements test

1. Introduction

The scaled adoption of adaptive learning technologies, which in many cases have led to improved student outcomes (Rai & Murthy, 2022), has raised concerns about algorithmic fairness (Kizilcec & Lee, 2022). A critical step in addressing these concerns is detecting these biases, which have been investigated in terms of race (Hu & Rangwala, 2020), gender (Christie et al., 2019), neurodivergence (e.g., ADHD; Lee et al., 2025), socioeconomic status (Yu et al., 2021), urbanicity (Ocumpaugh et al., 2015), and regional groups (Svabensky et al., 2024). The range of subgroups potentially affected by these biases accentuates the complexity of this issue. To date, many studies rely solely on institutionally defined demographic labeling systems, such as race or gender, based on legal protections in some countries. These labels often deliberately merge smaller categories in order to facilitate large-scale comparisons, but may fail to capture the nuanced, context-dependent dimensions of student identity in the classroom. In practice, heterogeneous groups (e.g., English language learners) are treated as monoliths, and significant within-group diversity is left unanalyzed or even misrepresented (Wang et al., 2022).

In addition to collapsing heterogeneity, institutionally defined labels may also miss the unique, context-specific dimensions of students' self-perceived identities (e.g., learner identity), which may be especially relevant in classroom settings (Cribbs et al., 2015; Crossley et al., 2018). Research shows that a student's sense of identity—like seeing oneself as "good at math"—can drive learning outcomes and academic success (Cribbs et al., 2015). Likewise,

while there are well-documented links between large-scale demographic categories (e.g., race, gender) and learners' performance, as well as help-seeking and motivational behaviors(e.g., Karumbaiah et al., 2022), incorporating non-traditional, locally meaningful categories can provide a richer, more nuanced picture of students and their learning experiences. These context-specific identities particularly matter in the case of modeling emotions, which are inherently social and situated (Barrett et al., 2019).

To investigate this, we examine whether affect detectors built for MATHia (an adaptive learning system) show biases against middle school students in terms of their self-reported identities, as elicited from a free-response survey. Doing so helps us understand whether any of these less-studied categories of identities are important for determining if automated detectors are less effective for some groups of students.

2. Literature Review

2.1 Algorithmic Bias in Education

Research on algorithmic bias in education has largely focused on broad demographic categories like race, ethnicity, nationality, and gender (Baker & Hawn, 2022). Such studies have revealed racial disparities across models predicting college retention (Kai et al., 2017), high school dropout (Christie et al., 2019), and course failure risk (Hu & Rangwala, 2020), with inconsistent results highlighting the complexity of intra-group heterogeneity. Geographic and linguistic differences have also emerged in automated essay scoring (Bridgeman et al., 2009) and in help-seeking models favoring domestic learners (Ogan et al., 2015). Gender disparities also appear but are often intertwined with other sociocultural factors (Yu et al., 2021).

However, reliance on these traditional categories risks over-simplification, as aggregating heterogeneous subgroups can obscure contextually salient identities (Belitz et al., 2023). Emerging work has explored less-studied categories, such as urbanicity (Ocumpaugh et al., 2014), neurodivergent groups (Lee et al., 2025), and regional groups in the Philippines (Svabensky et al., 2024), revealing biases tied to dimensions of identity that are finer-grained and more contextually situated than what is often studied. Yet these efforts remain fragmented, and few frameworks address how self-reported identities affect algorithmic biases. This gap highlights the need for more nuanced approaches to subgroup definition, ensuring that fairness evaluations reflect the lived experiences of learners rather than overly broad, externally chosen demographic variables.

2.2 Twenty Statements Test (TST)

Determining which factors are relevant potential sources of bias for a given student remains an underexplored area (Belitz et al., 2023). Gender, racial, and economic categories are often analyzed because of longstanding patterns of disparity in the treatment and educational outcomes of these groups (Kao & Thompson, 2003). While these issues remain important, the categories used in these contexts—which are designed to apply to many learners for ease of comparison at scale—are not always the most relevant in a specific situation. For example, populations in South Asia and East Asia have ethnic groups with vast cultural and linguistic diversity, which are often grouped under the same label, even though these groups (and their teachers) likely consider themselves distinct from one another. Efforts to capture identities that could supplement the existing commonly studied categories exist (Crossley et al., 2018), but the ability to do so at scale is quite challenging. One promising approach is the Twenty Statements Test (TST), which elicits a free-response survey to ask students how they would describe themselves (Kuhn, 1954). This technique is designed to surface salient categories that emerge from students' responses, which may not always be captured by traditional approaches.

2.3 Algorithmic Bias in Affect Detection

Existing research on algorithmic bias in education has examined predictive models of dropout (Christie et al., 2019), automated essay scoring (Bridgeman et al., 2009), and help-seeking behaviors (Ogan et al., 2015), but research involving affect detectors remains limited. For observation-based detectors, Ocumpaugh et al. (2014) found detectors failed to work across differences in urbanicity, and Chiu (2020) found gendered affect patterns—female STEM students persisted when less bored and off-task, while male students thrived with higher concentration and lower frustration. More recently, with facial-recognition-based detectors, Ashwin & Biswas (2024) reported higher misclassification rates for students with darker skin tones, underscoring the need for diverse training data. These limited efforts highlight the importance of examining bias in affect detectors across varied subpopulations. Furthermore, since affect is rooted in cultural and linguistic differences (Barrett et al., 2019), it may be important to incorporate variables of this nature in investigating the algorithmic bias that impacts detectors of affect.

3. Methods

3.1 Participants and Study Context

The study was conducted with middle school students across several classes in a small city in the northeastern U.S. with a large immigrant population. These students used Carnegie Learning's MATHia software in 2023-2024 as their regular math instruction. Participation in the study involved two different types of data collection. First, 219 students completed the Twenty Statements Test to determine relevant social categories for the students. Second, field observation of 163 students was conducted to collect training labels for automated detectors. Both parental consent and student assent were acquired.

The labeled affect data from the observers was mapped to action logs in MATHia for each student, as well as the corresponding TST identity labels when present. TST responses from students without affect labels were dropped since the affect labels served as ground truth and were necessary to build the detectors. A final total of 95 students' data were used for analysis. Students who didn't have TST responses for a specific category of interest were noted as NR (not reported) for that category, and evaluated as a separate subgroup for that category (explained in detail below).

3.2 BROMP-based Affect Detectors

3.2.1 BROMP Field Observations

Affect training labels were obtained using the Baker Rodrigo Ocumpaugh Monitoring Protocol (BROMP; Ocumpaugh et al., 2015a), a classroom observation method that has been used to produce detectors for more than two dozen online learning systems (Baker et al., 2020). BROMP observers use a momentary time sampling method, implemented by an open-source Android app (Ocumpaugh et al., 2015b), to label students' affective engagement. As part of the BROMP certification process, the researchers achieved a kappa of greater than 0.6 on all affective states for middle school students' observations, indicating good agreement among the researchers. Following this, coders independently coded over five class periods, observing for boredom, confusion, delight, engaged concentration, and frustration. As advised in the BROMP manual, these coders did not observe the same student simultaneously, maximizing data collection (Ocumpaugh et al., 2015a). In total, 591 observations of affect were obtained from 163 students. Engaged concentration was the most prevalent affective state (N=591; 71.21%), followed by confusion (N=109; 13.13%), frustration (N=42; 5.06%), boredom (N=86; 10.36%), and delight (N=2; 0.24%). This distribution is typical for BROMP research (Karumbaiah et al., 2021). Due to low N, a delight detector was not built.

3.2.2 Feature Engineering

To construct affect detectors, we extracted 157 features from the action log data of students. These included features like error counts, number of attempts, and changes in hint levels., based on actions of students. As typical in BROMP-based detectors, features within the 20 seconds (defined as a "clip") before the BROMP label's timestamp were prioritized, but earlier features were used to add contextual information (Baker et al., 2012; Zambrano et al., 2024).

3.2.3 Detector Construction and Validation

We performed stratified 4-fold student-level cross-validation. This ensured that affect labels were balanced across each fold while making sure that each student's labels appeared only in either the training or test set. To address class imbalances, we used a synthetic oversampling technique (SMOTE; Chawla et al., 2002) on training data. Features were selected through a forward feature selection process, evaluating performance using the Area Under the Receiver Operating Characteristic Curve (AUC ROC; AUC for short). Following common practice, each detector was trained as a one vs. rest classification (Baker et al., 2012; Zambrano et al., 2024). We tested 5 machine-learning algorithms from the Scikit-learn library for Python: Logistic Regression (LR), Random Forests (RF), Extreme Gradient Boosting (XGB), Support Vector Machines (SVM), and Decision Trees. We used the default hyperparameters for all models, as the study's goal was not to optimize overall model performance but to assess bias for the novel TST categories.

3.3 Twenty Statements Test (TST)

3.3.1 TST Implementation

This study measures student identity using the Twenty Statements Test (Kuhn, 1954), which asks students to fill in the blank—twenty times—for the question "Who am I?". This free-form survey has been used to produce nuanced identity classifications across urbanity samples (Somech, 2000), age groups (McRae & Costa, 1988), and nationalities (Santamaria et al., 2010). Notably, the TST facilitates identifying locally relevant categories that more fully reflect the identities of those being sampled.

3.3.2 TST Data Categorization

Qualitative coding of TST data was adapted from Gordon and Gergen (1968); these codes were then combined into umbrella categories specific to our sample. Table 1 shows the N for each subgroup across five umbrella categories, with examples given for those that do not align with typical census-style data. Each umbrella category also includes a Not Reported (NR) subgroup, as each umbrella category had students whose answers did not reflect that category. This large number of students being identified as NR is expected for open-ended methods and was included because not reporting a category serves as an indicator of lack of salience of that category in this context.

Table 1. *TST Categories. N≥10 are shown in grayscale and included in the fairness analyses*

Categories	**Subgroups**	**Description/example**	**N**
Gender	B	Boy	13
	G	Girl	21
	NR	not reported	149
Race / Ethnicity / Nationality	Black	-	13
	West African	Deidentified West African Nationality	18
	Caribbean	Deidentified Caribbean Nationality	9
	White	-	4

Categories	Subgroups	Description/example	N
	South Asian	Deidentified South Asian Nationality	2
	East Asian	Deidentified East Asian Nationality	3
	South American	Deidentified South American Nationality	2
	Mixed	-	3
	NR	not reported	142
Learner Identity (LI)	Pos	e.g., “good student” or “good at math”	36
	Neg	e.g. “not great at math” “don’t like school”	10
	Neut	e.g., “student”, or “high schooler”	10
	NR	not reported	127
Interpersonal Style (IPS)	Pos	e.g., “ nice” or “kind” or ”respectful”	33
	Neg	e.g., “rude”	3
	Neut	e.g., “shy”	2
	NR	not reported	145
Sense of Competence (SoC)	Pos	e.g., “smart” “leader”, “talented”	37
	Neg	e.g., “idiot”	1
	NR	not reported	145

Traditional Demographic Categories. Two categories aligned with traditional census measures emerged in this data: race and gender. Gender produced two named subgroups from the data: boy and girl, with NR included as a third subgroup. Race (a merger of race, ethnicity, and nationality) has many named subgroups but only two large enough (N≥10) for analysis: Black and West African (a pseudonym for a specific nationality that students provided, which we use to prevent re-identification of this school district). Because many students identified as both groups, we focus only on the West African subgroup. This decision reflects the higher specificity of the national label, and the concern that the Black subgroup would be more heterogenous. Students who reported being Black but were not in the West African subgroup had N<10 and were thus not analyzed. For analyses, we compare to “Other” labels and “not reported” (NR).

Novel Social and Contextual Categories. Three named categories emerged that did not reflect typical census data. Learner identity (LI) emerged with students reporting to be “a good student” (positive LI), “not good at math” (negative LI), or simply “a student” (neutral LI). Likewise, Interpersonal Style (IPS) also emerged with positive, negative and neutral subgroups, while Sense of Competence (SoC) had positive and negative but no neutral subgroups. An NR subgroup is included for all categories.

To facilitate fairness analysis, final subgroups within categories are mutually exclusive. In cases where TST identity categories are not mutually exclusive (e.g., valence subgroups, where students sometimes provided multiple, contradictory responses), a resolution process totaled the number of positive and negative responses and applied the majority label (e.g., if a student provided four neutral responses and one positive, the final code was neutral). Since no student had equal numbers of contradictory codes, no further resolutions were necessary.

3.4 Testing for Algorithmic Bias

We used three complementary fairness metrics—difference in AUC and weighted F1 (Wtd F1), and Mean Absolute Distribution Difference (MADD)—to test bias against TST subgroups with N≥10. AUC and Wtd F1 are standard predictive-performance metrics, while MADD is a novel fairness metric (Verger et al., 2023). AUC and Wtd F1 were calculated separately for each subgroup. MADD was calculated by comparing one subgroup to all others.

For AUC (range: 0-1; chance = 0.5), a threshold of ΔAUC≥0.1 was used to label bias. AUC reflects the trade-off between true positive and false positive rates, and it can be interpreted as a measure of how well an algorithm ranks individual cases. For the weighted F1 score (range: 0-1), the same threshold was chosen to label bias (ΔwtdF1≥0.1). This score combines precision and recall, while also taking class imbalance into account. Finally, MADD measures the divergence between the predicted probability distributions of a model across

subgroups, with scores ranging from 0 (most fair/identical distributions) to 2 (most biased/completely divergent distributions). For this study, we apply a conservative threshold of MADD>1 for labeling bias (Verger et al., 2023).

As small differences could arise stochastically (Kohavi et al., 2022), we chose conservative thresholds for our metrics. Our decision aligns with calls for precautionary fairness checks that are needed in all contexts (Passi & Barocas, 2019) but perhaps especially in educational AI to proactively mitigate harm (Holstein & Doroudi, 2021). These differences should not be treated as definitive evidence of bias but as signals to prioritize subgroups for validation in larger, representative samples (Mitchell et al., 2021).

4. Results

4.1 Affect Detector Models

The performance of different algorithms was compared for each binary affect classifier, and we selected the model for investigation of algorithmic bias based on the highest AUC. Table 2 shows performances for the final detectors that are comparable to previously published interaction-based affect detectors (Baker et al., 2014; Tabanao & Rodrigo, 2018): Random Forest for boredom (AUC=0.68), Logistic Regression for confusion and engaged concentration (AUC=0.65 and 0.66, respectively), and XGBoost for frustration (AUC=0.74).

Table 2. *Performance of detectors across algorithms with AUC values*

Algo.	**BOR AUC**	**CONF AUC**	**Eng CONC AUC**	**FRU AUC**
Decision Tree	0.62	0.56	0.60	0.63
SVM	0.62	0.61	0.59	0.65
LR	0.67	**0.65**	**0.66**	0.71
RF	**0.68**	0.64	0.64	0.69
XGBoost	0.67	0.64	0.64	**0.74**

4.2 Algorithmic Bias

We next evaluate bias in these models across the four affect detectors. Although we report bias evaluation both for categories similar to traditional, census-style demographics (i.e., gender and race) and for novel categories (i.e., Learner Identity, Interpersonal Style, and Sense of Competence), all data have emerged from students' free form responses in the TST.

4.2.1 Traditional Demographic Categories

Table 3 reports detector performance across TST subgroups and Table 4 shows AUC and F1 differences between subgroups for traditional demographic categories (gender and race).

Table 3. *Model evaluation results for all categories.*

		Gender			Race			Learner Identity				IPS		SoC	
Affect	Metrics	boy	girl	NR	W. Afr.	Other	NR	NR	pos	neg	neut	NR	pos	NR	pos
BOR (RF)	AUC	0.58	0.64	0.70	0.63	0.61	0.69	0.72	0.56	0.67	0.74	0.70	0.70	0.69	0.65
	Wtd F1	0.73	0.73	0.75	0.76	0.69	0.74	0.75	0.73	0.66	0.82	0.76	0.74	0.76	0.71
	MADD	0.45	0.49	0.40	0.49	0.40	0.49	0.31	0.36	0.64	0.66	0.34	0.37	0.35	0.35
CONF (LR)	AUC	0.72	0.62	0.62	0.55	0.66	0.65	0.64	0.64	0.70	0.70	0.58	0.78	0.62	0.67
	Wtd F1	0.73	0.68	0.65	0.64	0.66	0.67	0.67	0.64	0.64	0.79	0.64	0.77	0.64	0.72
	MADD	0.53	0.40	0.41	0.50	0.39	0.50	0.29	0.29	0.66	0.65	0.36	0.41	0.42	0.43
ENG (LR)	AUC	0.63	0.60	0.67	0.65	0.60	0.65	0.66	0.62	0.72	0.60	0.67	0.63	0.68	0.60
	Wtd F1	0.64	0.56	0.61	0.65	0.54	0.60	0.61	0.59	0.63	0.68	0.61	0.63	0.62	0.60

		Gender			Race			Learner Identity				IPS		SoC	
Affect	Metrics	boy	girl	NR	W. Afr.	Other	NR	NR	pos	neg	neut	NR	pos	NR	pos
	MADD	0.44	0.44	0.30	0.38	0.36	0.38	0.28	0.30	0.61	0.73	0.28	0.32	0.29	0.29
FRU (XGB)	AUC	0.71	0.64	0.73	0.79	0.76	0.71	0.70	0.74	0.94	0.80	0.75	0.56	0.72	0.75
	Wtd F1	0.82	0.93	0.90	0.87	0.85	0.89	0.90	0.87	0.92	0.83	0.88	0.92	0.89	0.88
	MADD	0.41	0.27	0.24	0.35	0.26	0.35	0.23	0.28	0.39	0.63	0.17	0.22	0.20	0.20

Table 4. *Differences in fairness metrics for all categories. Diff ≥ |0.1| are highlighted in gray.*

		Gender			Race			Learner Identity						IPS	SoC
Affect	Metrics	B-G	B-NR	G-NR	WA-Ot	WA-NR	Ot-NR	NR-Pos	NR-Neg	NR-Neut	Pos-Neg	Pos-Neut	Neg-Neut	NR-pos	NR-pos
BOR	AUC	-0.06	-0.12	-0.06	0.02	-0.06	-0.08	0.16	0.05	-0.02	-0.11	-0.18	-0.07	0.00	0.04
(RF)	WtdF1	0.00	-0.02	-0.02	0.07	0.02	-0.05	0.02	0.09	-0.07	0.07	-0.09	-0.16	0.02	0.05
CONF	AUC	0.10	0.10	0.00	-0.11	-0.10	0.01	0.00	-0.06	-0.06	-0.06	-0.06	0.00	-0.20	-0.05
(LR)	WtdF1	0.05	0.08	0.03	-0.02	-0.03	-0.01	0.03	0.03	-0.12	0.00	-0.15	-0.15	-0.13	-0.08
ENG	AUC	0.03	-0.04	-0.07	0.05	0.00	-0.05	0.04	-0.06	0.06	-0.10	0.02	0.12	0.04	0.08
(LR)	WtdF1	0.08	0.03	-0.05	0.11	0.05	-0.06	0.02	-0.02	-0.07	-0.04	-0.09	-0.05	-0.02	0.02
FRU	AUC	0.07	-0.02	-0.09	0.03	0.08	0.05	-0.04	-0.24	-0.10	-0.20	-0.06	0.14	0.19	-0.03
(XGB)	WtdF1	-0.11	-0.08	0.03	0.02	-0.02	-0.04	0.03	-0.02	0.07	-0.05	0.04	0.09	-0.04	0.01

Gender. The performance of each affect detector was tested across three gender subgroups: boy, girl, and not reported (NR). Boredom detection produced ΔAUC>0.1 when comparing the NR group (0.70) to boys (0.58). In contrast, confusion's AUCs are higher for boys (0.72) than for both girls and NR (each 0.62). For engaged concentration and frustration, no major disparities are found. Wtd F1 scores are largely comparable across subgroups—with the only notable exception for frustration (better for girls than for boys). MADD values ranged from 0.40 to 0.53, suggesting fair affective models.

Race. Next, we examine the performance of these detectors across three racial subgroups: West African, Others, and NR. For boredom, AUC values are similar across subgroups, but confusion detection does not perform as well for West African students (0.55) as for Others (0.66) and NR (0.65). For engaged concentration and frustration, AUC values showed minimal differences across subgroups. Wtd F1 scores also show few differences, though engaged concentration detection performs better for West African students than for Others (ΔWtd F1=0.11, see Table 3). All MADD values are relatively small (≤0.5).

4.2.2 Novel Identity Categories

Detector performance was tested across three novel TST categories: Learner Identity (LI), Interpersonal Style (IPS), and Sense of Competence (SoC). Among the LI codes, four subgroups are compared (positive, negative, neutral, and NR), and biases were evenly distributed across the affect detectors. Among the IPS and SoC codes, only two subgroups are compared (NR, positive), as only these subgroups had N>10 (Table 1). More differences in performance between the subgroups are seen in these novel categories than for the traditional ones. The specific results are discussed below.

Learner Identity (LI). Testing across the four LI subgroups shows that the boredom detector underperforms for students with a positive learner identity (AUC=0.56) compared to other students (NR: 0.72, negative: 0.67, neutral: 0.74), but for Wtd F1 scores, the detectors performed worse for students with negative LI than students with neutral LI (0.66 vs 0.82). For confusion, AUC values are comparable across subgroups, but Wtd F1 scores show better performance for the neutral subgroup (0.79) compared to the other three subgroups (ΔAUC≥0.12). For concentration, performance is better for the negative subgroup (0.72) than the positive (0.62) and neutral (0.60) subgroups (ΔAUC≥0.10). For frustration, the negative subgroup again performed best on AUC (negative: 0.94 vs. NR: 0.70, positive: 0.74, neutral:

0.80; ΔAUC=0.14 to 0.24). MADD values are highest when comparing the neutral subgroup across affective states (avg=0.67) but remain below 1 in all cases.

Interpersonal Style (IPS). For boredom, AUC values of both subgroups were equivalent (0.70), but for confusion, performance was better for the positive subgroup than the NR subgroup (0.78 vs 0.58; ΔAUC≥0.20). Similar differences were observed in Wtd F1, with the positive subgroup performing better than NR (0.77 vs 0.64; ΔwtdF1=0.13). For engaged concentration, there were no differences (.63 vs 0.67), but frustration detection performed better for the NR subgroup than the positive subgroup (0.75 vs 0.56; ΔAUC≥0.19). Low MADD values (<0.45) were seen across all subgroup comparisons.

Sense of Competence (SoC). Across all affect**,** AUC and Wtd F1 differences were minimal (ΔAUC < 0.1), indicating no significant disparities in model performance, while MADD values were also low, showing no substantial differences in probability distribution across subgroups.

5. Discussion and Conclusion

5.1 Summary of Results

Overall, results show performance differences across groups, emphasizing the importance of novel categories of Learner Identity and Interpersonal Style, in addition to demographic categories of gender and race for analyzing bias in affect detectors.

5.1.1 Performance Across Subgroups

Within gender, four group comparisons showed performance differences. Boys' boredom was not as accurately modeled (vs the NR group), nor was their frustration (vs the girls). However, the confusion detectors outperformed for boys compared to the other two groups. Within race categories, only two groups and two affective states experienced bias. The confusion detectors underperformed for the West African subgroup (compared to either subgroup), while concentration detection underperformed for the Others (vs. the West African subgroup) based on weighted F1.

More differences emerge within the novel categories. As Table 4 shows, no strong trends appear in terms of which detector is underperforming on which subgroup. Detectors underperformed for the positive LI subgroup in six comparisons to other groups, three of which involve boredom. Both the negative LI subgroup and neutral LI are underpredicted for two affective states (negative: boredom and confusion; neutral: frustration and concentration). Finally, the NR group is underpredicted for confusion (one comparison) and frustration (two comparisons). For the Interpersonal Style and Sense of Competence categories, differences only emerge for the positive and NR Interpersonal subgroups (in frustration and confusion).

Overall, the demographics typically studied in algorithmic bias research (gender, race/ethnicity) do not seem to be the categories showing the most bias in affect detectors. Instead, learner identities show the most variability in model performance, suggesting that non-traditional self-categorizations might help us to better understand algorithmic bias and the development of context-aware detectors more generally.

One limitation of the current method is the possibility of self-presentation effects. Combining the TST with validated surveys (e.g., self-efficacy) could disentangle some self-presentation effects from genuine identity differences.

5.1.2 Interpreting Not Reported (NR) Groups

The most common subgroup across all categories was NR, the group that had not provided a TST response relevant to that category. The NR groups raise important questions about our data, since identity salience—shaped by context, self-presentation effects, and societal norms—may influence reporting patterns. For instance, students in sensitive contexts (e.g., immigration status concerns) might underreport that identity, and other marginalized students

may avoid labels to avoid stigmatization. Likewise, cultural norms could make some groups less comfortable reporting accomplishments, which could influence categories like learner identity (e.g., “I’m a good student!”).

Although these issues complicate the interpretation of NR subgroups, additional surveys and efforts to cross-reference this label with the records maintained by a school could help us better understand these patterns. These efforts could be useful if we want to understand why, for example, a frustration detector is underperforming for NR group.

5.2 Conclusion

This study evaluates algorithmic bias within affect detectors that were developed for the MATHia online learning system. Unlike many previous studies of algorithmic bias, which have typically relied on census-style demographic data, we check model performance among social identity categories extracted from the Twenty Statements Test (TST). Evidence of algorithmic bias demonstrate that the bias may occur for characteristics beyond traditional demographic categories (cf. Baker & Hawn, 2024). The inclusion of TST categories enabled us to detect differences that might have been missed using more traditional categories. In particular, learner identity showed multiple differences across affect detectors and subgroups.

These findings offer evidence for the need to evaluate bias across contextually relevant categories (e.g., Learner Identity and Interpersonal Style), but further research is necessary to generalize these results to a broader population. In addition, future work should investigate how self-reported identity categories interact with presentation effects and other dynamics related to students’ identity reporting behaviors. More research on under-studied dimensions of identity, including those shaped by specific regional and ethnic norms, could allow us to better mitigate potential biases against small, contextually situated subgroups. Such efforts could enhance the fairness of models across a broader range of socially and contextually grounded student identities.

References

Ashwin, T. S., & Biswas, G. (2024). Identifying and mitigating algorithmic bias in student emotional analysis. *Int. Conf. Art. Intelligence in Ed., 89-103.* Cham: Springer Nature Switzerland.

Baker, R.S., Gowda, S., Wixon, M., Kalka, J., Wagner, A., Salvi, A., Aleven, V., Kusbit, G., Ocumpaugh, J. Rossi, L. (2012). Towards Sensor-Free Affect Detection in Cognitive Tutor Algebra. *International Conference on Educational Data Mining (2012), 126-133.*

Baker, R. S., & Hawn, A. (2022). Algorithmic bias in education. *International Journal of Artificial Intelligence in Education, 1-41*.

Baker, R. S., Ocumpaugh, J. L., & Andres, J. M. A. L. (2020). BROMP quantitative field observations: A review. Learning science: Theory, research, and practice, 127-156.

Baker, R.S., Ocumpaugh, J., Gowda, S.M., Kamarainen, A.M. and Metcalf, S.J. (2014). Extending log-based affect detection to a multi-user virtual environment for science.

Barrett, L. F. (2019). How emotions are made. Providence Book Festival, May, 25, 2019.

Belitz, C., Ocumpaugh, J., Ritter, S., Baker, R. S., Fancsali, S. E., & Bosch, N. (2023). Constructing categories: Moving beyond protected classes in algorithmic fairness. *Journal of the Association for Information Science and Technology, 74(6), 663-668.*

Bridgeman, B., Trapani, C., & Attali, Y. (2009). Considering fairness and validity in evaluating automated scoring*. In annual meeting of the National Council on Measurement in Education, San Diego, CA.*

Chawla et al. (2002). SMOTE: synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research, 16, 321-357.*

Chiu, M. S. (2020). Gender Differences in Predicting STEM Choice by Affective States and Behaviors in Online Mathematical Problem Solving: Positive-Affect-to-Success Hypothesis. *Journal of Educational Data Mining, 12(2), 48-77.*

Christie, S. T., Jarratt, D. C., Olson, L. A., & Taijala, T. T. (2019). Machine-Learned School Dropout Early Warning at Scale. *International Educational Data Mining Society.*

Cribbs, J.D., Hazari, Z., Sonnert, G., Sadler, P.M. (2015). Establishing an explanatory model for mathematics identity. *Child Development 86(4), 1048-1062.*

Crossley, S., Ocumpaugh, J., Labrum, M., Bradfield, F., Dascalu, M., Baker, R.S. (2018). Modeling Math Identity and Math Success through Sentiment Analysis and Linguistic Features. *International Educational Data Mining Society.*

Gordon, C., and Gergen, K.J. (1968). Self-conceptions: Configurations of content.

Holstein, K., & Doroudi, S. (2021). Equity and artificial intelligence in education: Will aied amplify or alleviate inequities in education?. arXiv preprint arXiv:2104.12920.

Hu, Q., & Rangwala, H. (2020). Towards Fair Educational Data Mining: A Case Study on Detecting At-Risk Students. International Educational Data Mining Society.

Kao, G., & Thompson, J. S. (2003). Racial and ethnic stratification in educational achievement and attainment. Annual Review of Sociology, 29(1), 417-442.

Karumbaiah, S., Baker, R. S., Ocumpaugh, J., & Andres, J. M. A. L. (2021). A re-analysis and synthesis of data on affect dynamics in learning. *IEEE Transactions on Affective Computing*, *14*(2), 1696-1710.

Karumbaiah, S., Ocumpaugh, J., & Baker, R. S. (2022). Context matters: Differing implications of motivation and help-seeking in educational technology. Int. J. Artificial Intel. in Ed., 32(3), 685-724.

Kizilcec, R. F., & Lee, H. (2022). Algorithmic fairness in education. The ethics of artificial intelligence in education, 174-202. Routledge.

Kohavi, R., Deng, A., & Vermeer, L. (2022). A/B testing intuition busters: Common misunderstandings in online controlled experiments. Proc. 28th ACM SIGKDD Conf. on Knowledge Discovery & Data Mining, 3168-3177.

Kuhn, M. & McPartland, T. (1954). An empirical investigation of self-attitudes. American Sociological Review 19, 68-76.

Lee, H., Belitz, C., Nasiar, N., & Bosch, N. (2025). XAI Reveals the Causes of Attention Deficit Hyperactivity Disorder (ADHD) Bias in Student Performance Prediction.

McCrae, R. R., & Costa Jr, P. T. (1988). Age, personality, and the spontaneous self-concept. *Journal of Gerontology, 43(6), S177-S185.*

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.*

Ocumpaugh, J., Baker, R., Gowda, S., Heffernan, N., & Heffernan, C. (2014). Population validity for educational data mining models: A case study in affect detection. British Journal of Educational Technology, 45(3), 487-501.

Ocumpaugh, J., Baker, R.S. and Rodrigo, M.M.T. (2015). Baker Rodrigo Ocumpaugh Monitoring Protocol (BROMP) 2.0 Technical and Training Manual.

Ogan, A., Walker, E., Baker, R., Rodrigo, M. M. T., Soriano, J. C., & Castro, M. J. (2015). Towards understanding how to assess help-seeking behavior across cultures. *International Journal of Artificial Intelligence in Education, 25, 229-248.*

Passi, S., & Barocas, S. (2019). Problem formulation and fairness. *In Proceedings of the conference on fairness, accountability, and transparency, 39-48.*

Rai, R. D., & Murthy, S. (2022). Evidence Over Intuition: Identifying Factors That Influence the Effectiveness of Large Scale Edtech Initiatives. In *International Conference on Computers in Education, 261-267*.

Santamaría, A., de la Mata, M. L., Hansen, T. G., & Ruiz, L. (2010). Cultural self-construals of Mexican, Spanish, and Danish college students: Beyond independent and interdependent self. Journal of Cross-Cultural Psychology, 41(3), 471-477.

Somech, A. (2000). The independent and the interdependent selves: Different meanings in different cultures. *International Journal of Intercultural Relations, 24(2), 161-172.*

Švábenský, V., Verger, M., Rodrigo, M. M. T., Monterozo, C. J. G., Baker, R. S., Saavedra, M. Z. N. L., & Shimada, A. (2024). Evaluating algorithmic bias in models for predicting academic performance of filipino students. EDM, 2024.

Tabanao, E., & Rodrigo, M. M. (2018). Investigating the Generalizability of Affect Detectors from Facial Expressions. In *International Conference on Computers in Education*.

Verger, M., Lall, S., Bouchet, F., & Luengo, V. (2023). Is Your Model "MADD"? A Novel Metric to Evaluate Algorithmic Fairness for Predictive Student Models. arXiv preprint arXiv:2305.15342.

Wang, A., Ramaswamy, V. V., & Russakovsky, O. (2022). Towards intersectionality in machine learning: Including more identities, handling underrepresentation, and performing evaluation. *In Proc. of the 2022 ACM Conference on Fairness, Accountability, and Transparency, 336-349.*

Yu, R., Lee, H., & Kizilcec, R. F. (2021). Should college dropout prediction models include protected attributes?. In Proc. of the 8th ACM conference on learning@ scale (pp. 91-100).

Zambrano, A. F., Nasiar, N., Ocumpaugh, J., Goslen, A., Zhang, J., Rowe, J., ... & Hutt, S. (2024). Says Who? How different ground truth measures of emotion impact student affective modeling. *In Proceedings of the 17th Int. Conference on Educational Data Mining, 211-223.*