The Impact of Gender and Ethnicity on Academic Performance: Insights from Bayesian Models

Anders Havbro Hjulmand
IT-University of Copenhagen
ahju@itu.dk

Andreas F. F. Olsen
IT-University of Copenhagen
frao@itu.dk

Eisuke Okuda
IT-University of Copenhagen
eiok@itu.dk

***Abstract*—Differences in academic achievement have been linked to gender and ethnicity. Understanding the gender- and ethnicity-gaps in academic achievement is of major importance to society. This study examines variations in English Literature achievement among 4.463 American community college students.**

Bayesian inference is used to explore whether the gender- and ethnicity-effects vary among students taught by different instructors. The models comprise complete pooling, no-pooling and multilevel approaches.

This study finds differences in academic performance based on students' ethnicities and genders. These effects are best captured by the multilevel models, indicating that there exists some degree of variability in the effects among instructors. In addition, White students have smaller variation across instructors compared to other ethnicities. It remains unclear how much of this difference can be attributed to unequal sample sizes in ethnicities and among individual instructors.

Future work could incorporate a richer dataset with additional courses, more equal sample sizes among ethnicities, and information on instructors such as their age or demeanour towards students.

I. Introduction

Education and academic achievement are fundamental for human resource development and the progress of society. Recognizing the differences in academic achievement between demographic groups is crucial to promoting equitable opportunities. Gender and ethnicity have been associated with variations in academic performance. Existing research indicates a general female advantage, especially for courses in Literature and Language [1], [2]. In addition, students of Asian ethnicity often achieve better academic results than students of White- and Hispanic ethnicity, while students of Hispanic ethnicity often have lower performance than White students [3]–[5]. Understanding these gender- and ethnicity-gaps in academic achievement is of major concern to individuals and policy-makers.

Beyond the gender- and ethnicity-gaps in academic performance, research suggest that instructors' expectations, attitudes, behavior and speech towards students differ depending on the gender and ethnicity of the student [6]–[8]. It has also been found that congruent instructor and student ethnicity can lead to better performance, for example that Black students score higher on achievement tests when assigned to a Black instructor [9].

In light of these findings, this study examines whether gender- and ethnicity-effects in academic achievement are different between instructors. Consider that instructor variation can be characterized on a continuum, with high variation in one end and low variation in the other end. A high variation among instructors suggest that gender- and ethnicity-effects vary a lot depending on the instructor, for example that the performance of Hispanic students is very high for some instructors and very low for others. From the perspective of individual students, this entails that their assigned instructor impacts the students' performance. On the other hand, a low variation between instructors indicate little difference in gender- and ethnicity-effects among instructors. For instance, female students may consistently have a high performance regardless of their assigned instructor. This suggests that female students would not prefer one instructor over another based on their expected academic performance.

This observational study examines gender- and ethnicity-effects in the academic performance of 4.463 American community college students in a first year English Literature course. Bayesian inference is used to examine overall gender- and ethnicity-gaps, and to examine whether the effects vary among students taught by different instructors. The following hypotheses are evaluated:

- **H1:** There are gender- and ethnicity-gaps in academic performance.
- **H2:** The gender-and ethnicity-effects on academic performance vary among instructors.
- **H3:** The variance in gender- and ethnicity-effects between instructors differs among genders and ethnicities.

II. Dataset

The dataset used in this study, referred to as `writing_center`, was sourced from GitHub [11]. The dataset, as described by its author, comprises community college students' first attempt/enrollment in ENG 1, a first-year college level writing course, over the span of three years. It initially consisted of 4.727 rows, with one observation per student. Each row contains information about a student ranging from demographic independent variables, such as `Gender` and `Ethnicity`, to course-specific independent variables, including `Instructor ID` and `Main Course SuccessFlag`. Additional independent variables of interest consisted of the following: `FinAid`, a binary variable indicating if the student received financial aid, `Term Units Attempted` indicating the total number of units/classes the

student attempted in the current term and Age. For a detailed description on each independent variable, see Figure A1.

A. Exploratory data analysis

1) Preprocessing: To better understand the composition of the dataset, the distribution of key categorical variables: Gender and Ethnicity, were investigated. (See Table I and II.) Undisclosed values were removed in both Gender and Ethnicity. Students whose ethnicity was African American, Pacific Islander, or Native American were infrequent compared to the other ethnicity groups and were removed. This resulted in 4.463 students of ethnicities White, Asian, Mixed, and Hispanic. The categorical variables Gender and Ethnicity were one-hot encoded.

Gender	Count (%)
Male	2403 (51)
Female	2258 (48)
Decline to State	66 (1)

TABLE I
BREAKDOWN OF INDEPENDENT VARIABLE GENDER.

Ethnicity	Count (%)
White	1501 (31)
Asian	1442 (3)
Mixed	972 (2)
Hispanic	605 (13)
Decline to State	102 (2)
African American	93 (2)
Pacific Islander	9 (<1)
Native American	3 (<1)

TABLE II
BREAKDOWN OF INDEPENDENT VARIABLE ETHNICITY.

2) Key statistics: The dataset contained 52 unique instructors with a median (std) number of students of 70 (57). The minimum and maximum number of students in the instructor-groups was 10 and 221.

Main Course SuccessFlag, the outcome variable, had a mean of 0.77 and a standard deviation of 0.42, indicating that 77 % of the students successfully passed the course. The correlation between the outcome variable and independent variables was then examined (See Figure 1). Several independent variables were observed to have some correlation with the outcome variable. Notably, Term Units Attempted showed a positive correlation with the likelihood of a student successfully passing the course, while gender and ethnicity related independent variables had weaker correlations. These required further investigation to evaluate their effects.

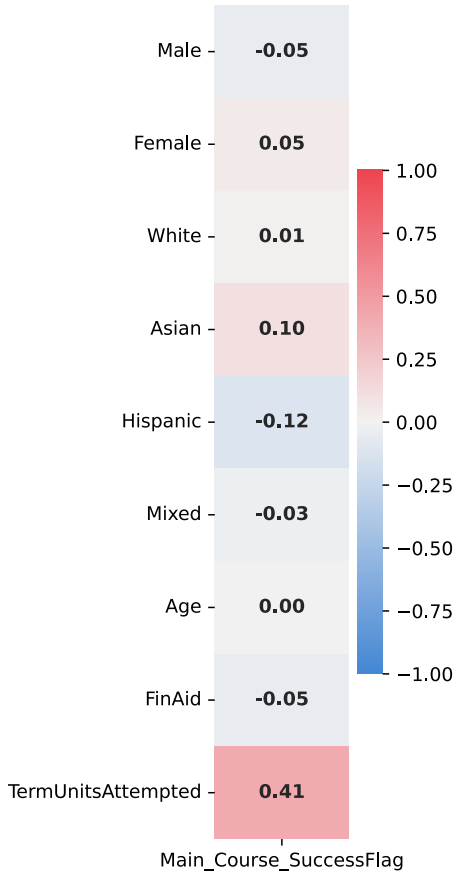

Fig. 1. The correlation of each independent variable to the outcome variable, MainCourseSuccessFlag. Color red indicates a positive correlation to the outcome variable whereas blue suggests negative.

III. METHODS

A. Model Introduction

To investigate the hypotheses presented in this paper, three different Bayesian regression models were deployed using a generalised linear model (GLM) with link function sigmoid. Each of the models were chosen to compare the variance of academic performance between students based on their genders, and ethnicities. Two of the models respectively only used gender and ethnicity as the independent variables. Additionally, other information of the students such as FinAid, was utilized in an adjusted model to examine the adjusted effects of the independent variables.

To explore hypothesis 2 & 3, all instructors were treated as independent groups in a no-pooling model assuming that the posterior distribution of Main Course SuccessFlag was drawn from independent instructor level model parameters. This assumption allowed for increased variance between model parameters as opposed to a complete pooling model where all instructors were treated as one group that shared model parameters. The complete pooling approach was implemented to examine hypothesis 1 i.e. the overall variance in academic performance based on ethnicity, gender and adjusted independent variables. To explore the spectrum of variance between model parameters, a multilevel model was deployed as a third model to instrument a regularization term on the hyper-parameters defined by the hyper-priors to estimate the model-parameters.

B. Model Specifications

All the models used a linear combination of independent variables $\mathbf{X}$ and coefficients β as linear predictor $\mathbf{X}\beta$ for the

outcome variable $\mathbf{Y}$ called `Main Course SuccessFlag` $\in [0, 1]$. The link function **logit**: $g = ln(\frac{\mu}{1-\mu})$ was then used to find the expected value $\mathbb{E}(\mathbf{Y}|\mathbf{X}) = g^{-1}(\mathbf{X}\beta)$, hence the likelihood of $\mathbf{Y} \in [0, 1]$ was modelled as follows:

$$\hat{y_i} \sim g^{-1}(\mathbf{X}_{ij}\beta_{1,..,j} + \beta_{\mathbf{0}}) \tag{1}$$

As exact optimisations for the marginal likelihoods are required by Bayes theorem but in practice intractable, this paper resolved to the **NUTS** sampler, an adaptive Hamiltonian Monte Carlo sampling algorithm presented by [12].

The three types of Bayesian models used different levels of complexity to estimate the posterior distributions of the β given three different sets of independent variables. The first set of independent variables **Gender** included {`Male`, `Female`}, the second set **Ethnicity** included {`White`, `Asian`, `Hispanic`, `Mixed`}, and the third set **Adjusted** included {`Male`, `Female`, `White`, `Asian`, `Hispanic`, `Mixed`, `Age`, `FinAid`, `Term Units Accepted`}. In an attempt to incorporate extraneous independent variables in the adjusted model, the sampling process failed due to divergence problems. Table III shows abbreviations of the 9 models that the combination of independent variables and approaches resulted in.

	Complete-pooling	**No-pooling**	**Multilevel**
Gender model	$M_{G/CP}$	$M_{G/NP}$	$M_{G/ML}$
Ethnicity model	$M_{E/CP}$	$M_{E/NP}$	$M_{E/ML}$
Adjusted model	$M_{A/CP}$	$M_{A/NP}$	$M_{A/ML}$

TABLE III
OVERVIEW AND ABBREVIATIONS OF THE BAYESIAN MODELS USED IN THE STUDY.

The **Complete Pooling** models assume no differences between instructors thus pooling them together with 2 weakly informative shared prior normal distributions to estimate the β coefficients, where $k \in \{1, \ldots, j\}$:

$$\beta_0 \sim \mathcal{N}(0, 1) \tag{2}$$
$$\beta_k \sim \mathcal{N}(0, 1) \tag{3}$$

The **No Pooling** models assume differences in model parameters depending on instructors, which requires unique β coefficients for each instructor. This is denoted by the additional subscript t on the β coefficients, where $t \in \{1, \ldots, 52\}$:

$$\beta_{0,t} \sim \mathcal{N}(0, 1) \tag{4}$$
$$\beta_{k,t} \sim \mathcal{N}(0, 1) \tag{5}$$

The **Multilevel** models add another level of distributions in the model structure called hyper-parameter distributions. Let these hyper-parameter distributions be denoted by ϕ such that the prior distributions for the β coefficients now are controlled by the hyper parameters: ϕ_μ and ϕ_σ. As in the no pooling model defined above, a beta coefficient is estimated for each instructor, but they now follow a common prior distribution estimated with information from all instructors. The multilevel model is defined with the hyper-priors:

$$\phi_{\beta_0,\mu} \sim \mathcal{N}(0, 1) \tag{6}$$
$$|\phi_{\beta_0,\sigma}| \sim \mathcal{N}(0, 1) \tag{7}$$
$$\phi_{\beta_k,\mu} \sim \mathcal{N}(0, 1) \tag{8}$$
$$|\phi_{\beta_k,\sigma}| \sim \mathcal{N}(0, 1) \tag{9}$$

To increase the effective sampling from the posterior distribution of the NUTS sampler, a non-centered parameterization was implemented on the β parameters to yield less complex posterior geometries as suggested by [13], [14]. This was done by adding auxiliary variables denoted by $\eta_{k,t}$ drawn from a normal distribution for all instructors and β. Then the distributions for the instructor dependent β coefficients with subscripts k can be defined as:

$$\eta_{0,t} \sim \mathcal{N}(0, 1) \tag{10}$$
$$\beta_{0,t} \propto \phi_{\beta_{0,\mu}} + \eta_{0,t}\phi_{\beta_{0,\sigma}} \tag{11}$$
$$\eta_{k,t} \sim \mathcal{N}(0, 1) \tag{12}$$
$$\beta_{k,t} \propto \phi_{\beta_{k,\mu}} + \eta_{k,t}\phi_{\beta_{k,\sigma}} \tag{13}$$

C. Model fitting and convergence diagnostics

The Bayesian models were implemented with the python module `PyMC` [15]. The convergence of the models were assessed using $\hat{R}$, which is a measure of how well the distinct sampling chains have mixed [16].

All the models were fitted with four parallel chains, using 2.000 burn-in-samples and 4.000 sample draws from the posterior distributions. This resulted in a total of 16.000 draws for each parameter. The models had no divergences and $\hat{R}$ values were 1.0 for all posterior distributions. Visual inspection of the Markov chains showed no obvious problems with the sampling process. These observations suggested that the posterior distributions were adequately sampled.

D. Experiment 1 - Andreas

The first experiment aimed to explore the effect of the independent variables X on the outcome variable Y. This experiment solely focused on the complete pooling models ($M_{G/CP}$, $M_{E/CP}$, $M_{A/CP}$) to investigate **H1**, i.e. if there existed gender- and ethnicity-gaps in academic performance.

To address this hypothesis, this experiment compared the posterior distributions of the β coefficients among the adjusted and the non-adjusted models. This is done by visually inspecting the posterior distributions parameterized by: $\mu_{\beta_k}, \sigma_{\beta_k}, \forall k \in \{1 \ldots j\}$ for the adjusted and the non-adjusted models, to inspect the consistency of the effects from the β coefficients when including additional covariates.

Similarly, to compare the effects of two independent variables β_n and β_m when including covariates, the relative difference of the β distributions were sampled and subtracted throughout the inference:

$$\delta(\beta_n, \beta_m) = \sigma_{\beta_n}) - (\mu_{\beta_m}, \sigma_{\beta_m}) \tag{14}$$

This relative difference in the complete pooling models could explain a global impact of the independent variable X_n on

the dependent variable $Y = y_i$. An example could be in the gender model $M_{G/CP}$ where the relative difference between β_{Female} and β_{Male} show a systematic positive impact on y_i towards one of the coefficients.

E. Experiment 2 - Eisuke

Experiment 2 was designed to explore the variations in academic performance related to gender and ethnicity across different instructors (**H2**).

Comparisons between complete pooling, no-pooling, and multilevel approaches was first conducted for each model (Gender, Ethnicity, Adjusted). It was based on the assumption that if the gender- or ethnicity-effects on academic performance vary among instructors (**H2**), the no-pooling and multilevel models would better capture the overall students' performance than the complete pooling. Leaving-one-out Cross-validation (LOO) was deployed to evaluate the model performance. The LOO was approximated by the Expected Log-pointwise Predictive Density (ELPD) for each model as a metric [17].

Further comparison was made between the results of the parameter estimates of β_{Male}, β_{Female}, β_{White}, β_{Asian}, β_{Mixed}, and $\beta_{Hispanic}$ by instructor-level within the adjusted model. This comparison was used to evaluate if there exist tangible variations in students' academic success based on their attributes across individual instructors, and whether there were any difference in the parameter estimates between no-pooling and multilevel approaches.

F. Experiment 3 - Anders

The aim of experiment 3 was to utilize information from the multilevel models to analyse whether the variance in gender- and ethnicity-effects between instructors differed among genders and ethnicities (**H3**).

In the multilevel models, the distributions of hyper-parameters serve as priors in estimating the parameters of each instructor. The variance hyper-parameter ϕ_{β_σ} signifies the extent to which the parameters for each instructor can vary independently of other instructors, what is also known as the shrinkage effect. A small ϕ_{β_σ} corresponds to an informative prior, and means that the parameter of each instructor is strongly biased towards the grand mean. A large ϕ_{β_σ} corresponds to an uninformative prior and allows the parameter of each instructor to vary independently of other instructors.

Experiment 3 compared the hyper-parameters ϕ_{β_σ} of the independent variables `Male`, `Female`, `White`, `Asian`, `Hispanic`, and `Mixed`. For each pair of independent variables, denoted as (X_n, X_m) the magnitude of differences in $\phi_{\beta_{X_n,\sigma}}$ and $\phi_{\beta_{X_m,\sigma}}$ was assessed by two independent procedures:

First, a distribution of differences was sampled with `PyMC` by subtracting the two distributions: $\phi_{\beta_{X_n,\sigma}} - \phi_{\beta_{X_m,\sigma}}$. This was only conducted for the non-adjusted models $M_{G/ML}$ and $M_{E/ML}$, as the adjusted model $M_{A/ML}$ encountered divergence issues.

Second, a non-parametric bootstrap was used to obtain a distribution of differences in medians. Bootstrapped samples $\phi_{\beta_{X_n,\sigma}}* = (n_1^*, ..., n_{16.000}^*)$ and $\phi_{\beta_{X_m,\sigma}}* = (m_1^*, ..., m_{16.000}^*)$ were obtained by sampling elements with replacement from the original distributions $\phi_{\beta_{X_n,\sigma}} = (n_1, ..., n_{16.000})$ and $\phi_{\beta_{X_m,\sigma}} = (m_1, ..., m_{16.000})$. The difference in medians between $\phi_{\beta_{X_n,\sigma}}*$ and $\phi_{\beta_{X_m,\sigma}}*$ was then calculated. This process was repeated 10.000 times, and performed for all three multilevel models.

IV. Results

A. Experiment 1 - Andreas

The effect of the independent variables X shown by their posterior distributions in figure 2A, appear to align with **H1**. This is supported when inspecting 2B, where the difference between `Asian - Hispanic` throughout the sampling is $\frac{0.9+1.1}{2} = 1$, indicating a higher odds ratio of passing a course for students of Asian ethnicity compared to those of Hispanic.

In the gender model ($M_{G/CP}$) where Y is only dependent $X \in [Female, Male]$ the posterior mean of $\beta_{Female} = 0.53$ and $\beta_{Male} = 0.299$ indicating an odds ratio: $\frac{\exp^\mu}{1+\exp^\mu} = 5.75\%$, meaning females have 5.75% increased chance of a passing the course compared to males. Similarly, from the different effects of β_k in the ethnicity model seen in 2A, it shows that students of different ethnicities have varying odds ratios. As a prevalent example, students of Asian ethnicity with $\mu_{\beta_{Asian}} = 0.71$ versus $\mu_{\beta_{Hispanic}} = -0.32$, have a 25% higher chance of passing the course compared to those of Hispanic ethnicity. 2 A also shows, that when adding terms to the linear predictor in the adjusted, the signal of the gender and ethnicity independent variables generally decreases.

Interestingly looking at figure 2B, it appears, that the gender and ethnicity differences present in the non-adjusted models, are still maintained when including other covariates. This is especially the case for ethnicity variables except `White - Mixed`. This indicates, that the effect of the added independent variables: {`Age`, `FinAid`, `Term Units Accepted`}, seem to encompass the variance in odds ratios of different ethnicities and genders, as seen in 2B. These gender and ethnicity differences in the adjusted- and the non-adjusted models provide evidence for **H1**.

B. Experiment 2 - Eisuke

Figure 3 displays the results of the cross validation for each model (Gender, Ethnicity and Adjusted). It was revealed that the multilevel approach best predicted the students' academic success across all models. This provided evidence for **H2**, indicating that the gender- and ethnicity-effects on academic performance did differ across individual instructors to some extent. Additionally, the results showed that the no pooling approach performed differently for each model. While it predicted better than the complete pooling for the Gender and Ethnicity models, it performed poorly with the Adjusted model. This suggested that it forced the model to estimate parameters on smaller sample size, and that the constraint

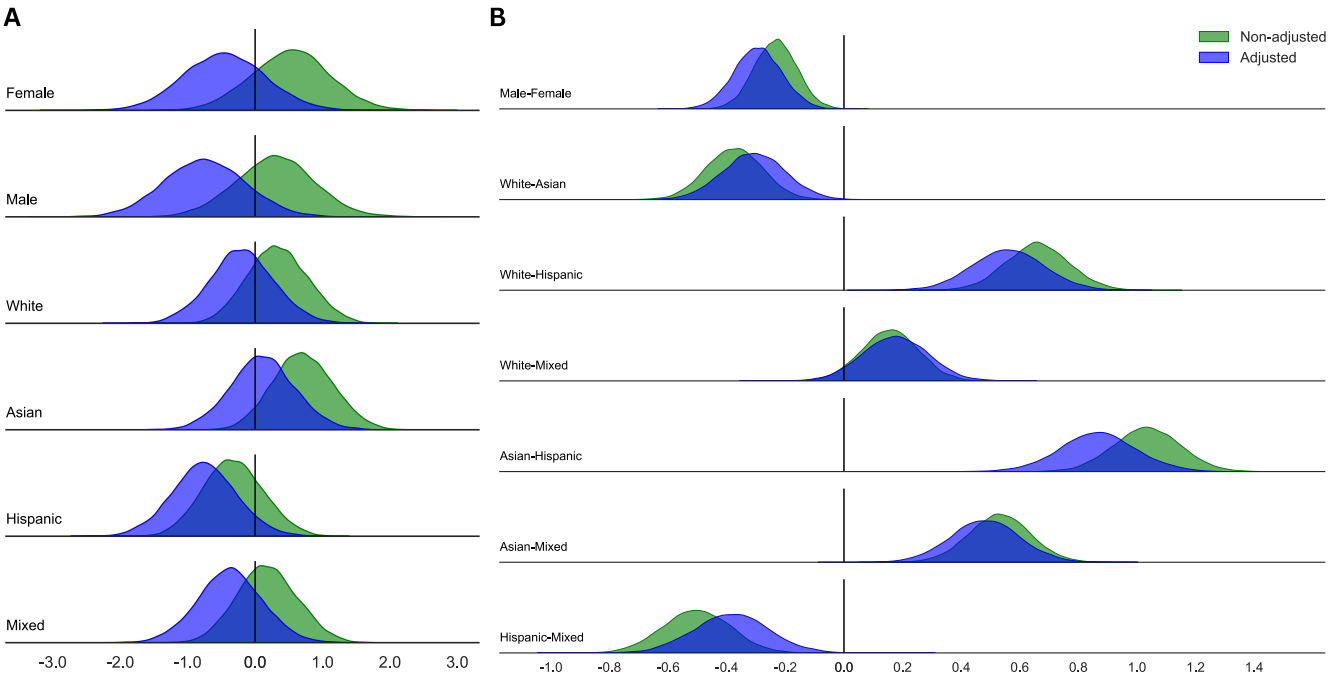

Fig. 2. **A** conveys the posterior distributions of the effects β_k of the independent variables in the non-adjusted models $M_{G/CP}$, $M_{E/CP}$ in **green** compared to the adjusted model $M_{A/CP}$ in **blue**. **B** displays the difference between these effects sampled with $\delta\beta_m, \beta_n$. The x-axis shows the numerical effects of the β coefficients, where the y-axis states the name of the β coefficient. In **B**, the relative difference for each of the coefficient pairs is calculated in the order that the y-axis displays: `Male - Female` $= \delta(\beta_{Male}, \beta_{Female})$.

combined with a larger number of independent variables caused over-fitting to the model.

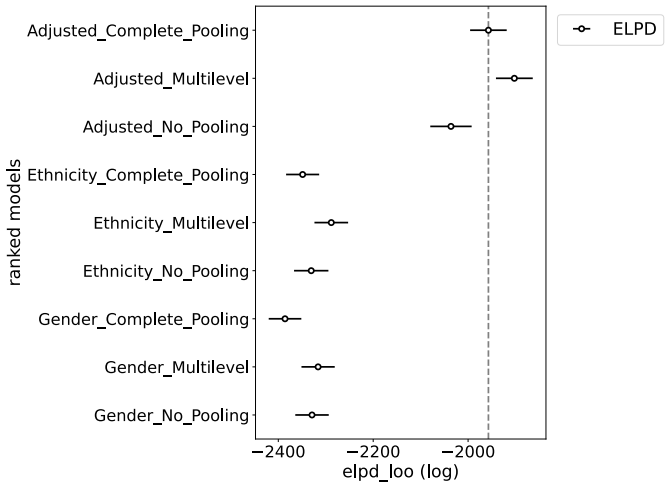

Fig. 3. The results of Leave-one-out Cross-validation (LOO). Expected Log-pointtwise Predictive Density (ELPD) was calculated and visualized with the standard error for each model.

The results for the parameter estimation β_{Male} was shown in Figure 4. While the estimate for the multilevel was characterized by its mean plotted closely to one another and the shorter 95 % High Density Interval (HDI), the counterpart for the no pooling has its mean spread across instructors and the 95 % HDI was wider. Similar patterns were observed for β_{Female}, β_{White}, β_{Asian}, β_{Mixed}, and $\beta_{Hispanic}$ (See Figure A3). It was indicated that the multilevel model provides estimates with a lower variance, hence less susceptible to over-fitting compared to the no pooling model.

Overall, experiment 2 provided some evidence for **H2**. The gender- and ethnicity-effects in academic performance varied among instructors to an extent. The multilevel approach was shown most effective to predict the outcome.

C. Experiment 3 - Anders

Experiment 3 aimed to examine whether the variance in gender- and ethnicity-effects among instructors were different depending on demographic groups. Figure 5 summarises the results. The ϕ_{β_σ} for `Male` and `Female` were generally higher than for ethnicity, suggesting that gender-effects among instructors vary more than ethnicity-effects (Figure 5A+D). There was however, no difference in ϕ_{β_σ} between `Male` and `Female`.

For ethnicity, the bootstrap differences of medians in the non-adjusted models were ordered such that $\phi_{\beta_{White,\sigma}} < \phi_{\beta_{Asian,\sigma}} < \phi_{\beta_{Hispanic,\sigma}} < \phi_{\beta_{Mixed,\sigma}}$ (Figure 5C). In contrast, when examining adjusted models, substantial distinctions were observed primarily between $\phi_{\beta_{White,\sigma}}$ and the remaining ethnicities (5E). The differences among the other ethnicities had diminished, and their effect sizes were deemed negligible, despite the fact that the distributions did not cross zero and had a high confidence. This suggests that the variation in academic performance among instructors was smaller for

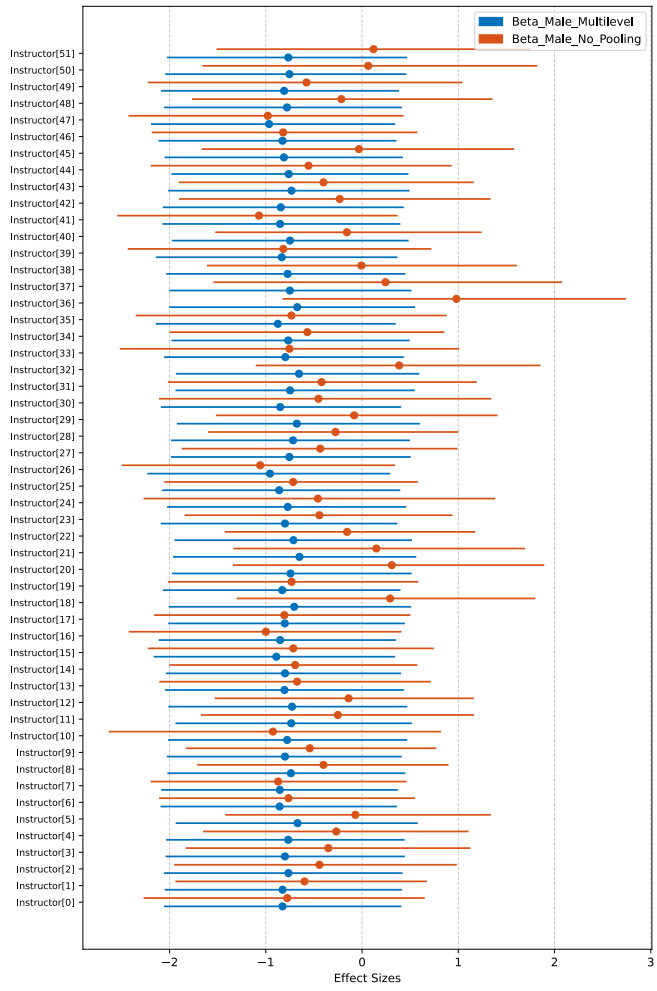

Fig. 4. The parameter estimate results for β_{Male} by the instructor-level with $M_{A/ML}$ (blue) and $M_{A/NP}$ (red). Centered to the mean, the 95 % High Density Interval (HDI) is shown for individual instructors. For the full version, see Figure A3.

White students compared to the other students. White students performed more similar regardless of their instructor, whereas the performance of other students was more dependant on the instructor. The relative magnitudes of the bootstrap differences in medians were up to $\sim 90\%$ (White - Mixed), signifying that the variance among teachers with respect to the effect of `Mixed` was almost twice the size as for the effect of `White`. It is however difficult to interpret whether the absolute magnitude of the bootstrap differences are large enough to have real-world implications.

While the bootstrap results found differences in medians of $\phi_{\beta_{White,\sigma}}$ and the remaining ethnicities, all the distributions of differences greatly overlapped with zero, signifying that neither of the ϕ_{β_σ} were different from each other (Figure 5B). This suggests that the variation in academic performance among instructors was not substantially different between genders or ethnicities. Stronger effects are likely needed to push the distributions of differences substantially away from zero.

In summary, experiment 3 found partial support for **H3**. While the differences in ϕ_{β_σ} showed no demographic variations among instructors, the bootstrap difference of medians found evidence in favor of differences between $\phi_{\beta_{White,\sigma}}$ and other ethnicities.

V. Discussion

A. Main findings and implications

This study investigated gender- and ethnicity-gaps in the academic performance of 4.463 students, and further examined whether the gender- and ethnicity effects varied among students taught by different instructors.

Moderate support was found for **H1**. Most notably, the gaps in academic performance was highest between students of Asian and Hispanic ethnicity, whereas the gender-gap was of less magnitude.

Some evidence was provided for **H2**. While the gender- and ethnicity-effects on academic performance did not vary a great deal among instructors, there existed an instructor-level difference to a certain degree. The difference should be interpreted with caution as the number of students for each instructor largely differed from 10 to 221, which could potentially lead to an apparent difference.

Partial support was found for **H3**. While there was no difference in the variation of academic performance between males and females among instructors, White students had smaller variation compared to other ethnicities. This implies that the academic performance of Asian, Hispanic, and Mixed students relies more on the assigned instructor, whereas the performance of White students remains less affected by their instructor. It is however important to note that this observational study did not examine *why* White students remained less affected by their instructor. Nonetheless, the results suggest that there exist differences in the fairness of opportunities between ethnicities in this dataset.

B. Limitations and future research

There were several limitations in this study. Although the total size of the population was 4.463, there were some instructors with few students (10 instructors had fewer than 30 students). The small sample sizes were associated with larger standard deviations in the parameter estimations in the no-pooling models (Figure 6A). This is especially apparent for independent variables with small sample sizes such as `Hispanic` (n=605). On the other hand, the standard deviations were not associated with instructor sample size in the multilevel models, because parameter estimates were shrunk towards a grand mean (Figure 6B). Consequently, the limited sample size for certain instructors is not an issue in the context of multilevel models. Similar results were found in the adjusted model (See Figure A4).

Another limitation was the unequal number of students across different ethnicities. While there were 1.501 students of White ethnicity and 1.442 of Asian ethnicity, there were

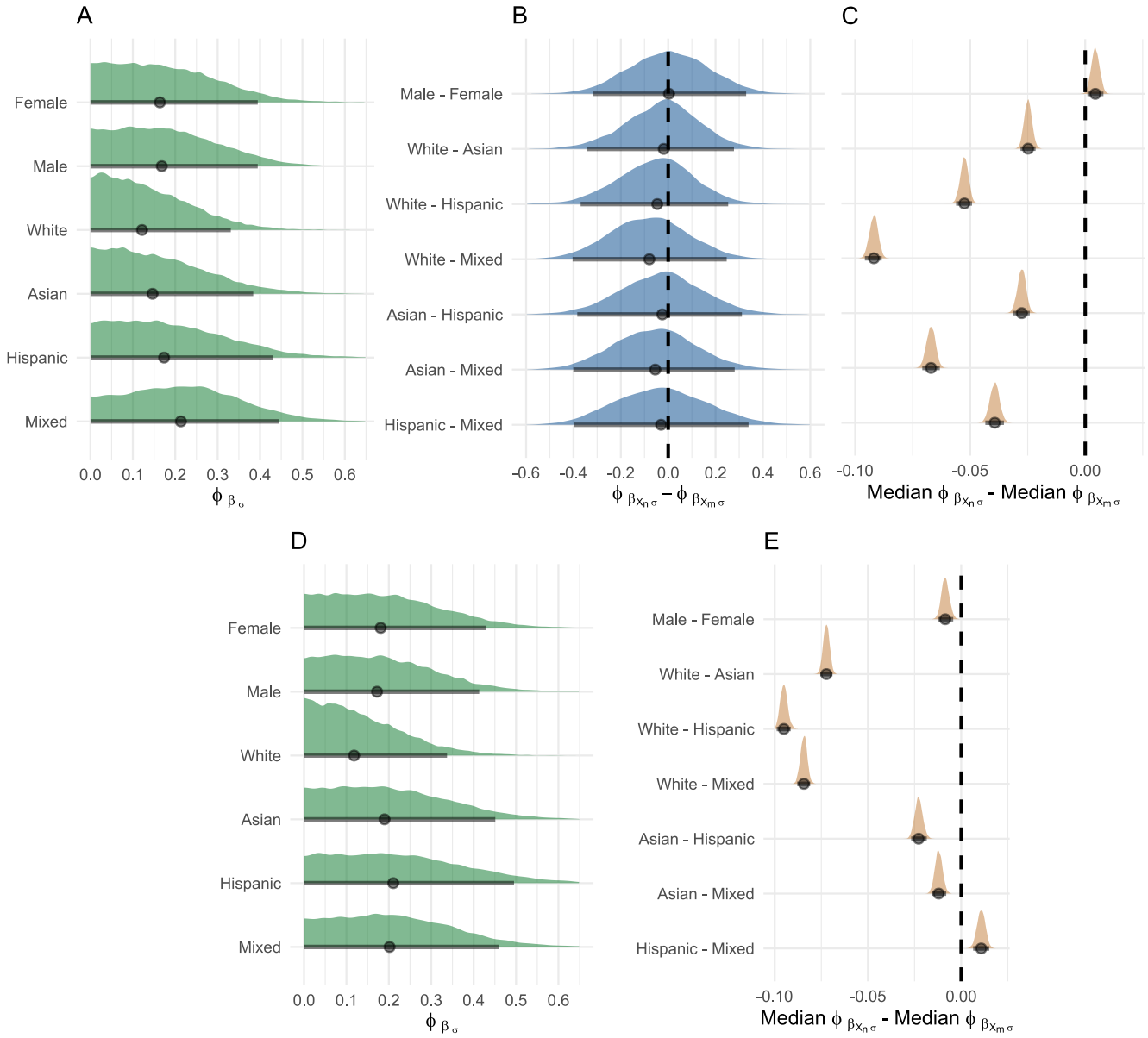

Fig. 5. Distribution of hyper-parameter variance ϕ_{β_σ} for the independent variables in non-adjusted models $M_{G/ML}$, $M_{E/ML}$ (A) and the adjusted model $M_{A/ML}$ (D). Difference between $\phi_{\beta_{X_n,\sigma}}$ and $\phi_{\beta_{X_m,\sigma}}$ for pairs of independent variables (X_n, X_m) in non-adjusted models (B). Difference of medians between $\phi_{\beta_{X_n,\sigma}}$ and $\phi_{\beta_{X_m,\sigma}}$ for pairs of independent variables (X_n, X_m) estimated with non-parametric bootstrap in non-adjusted models (C) and the adjusted model (E). The dots and the line show the median and 95% highest density interval (HDI).

only 972 of Mixed ethnicity and 605 of Hispanic ethnicity. Reduced sample sizes can lead to increased noise in parameter estimations for each instructor in the no-pooling models. Consequently, this makes the instructor effects appear more heterogeneous than the underlying "true" effect. The imbalance reduces the confidence in the results of experiment 3. Although it was found that the hyper-parameter variances for White students were smaller than for Asian, Hispanic and Mixed students, it remains unclear how much of this difference can be attributed to different sample sizes.

Another constraint was that the dataset did not contain any information about the instructors, such as gender, ethnicity, age or demeanour towards their students. As a result, this study was restricted to *observe* differences in gender- and ethnicity-effects among instructors rather than *explain* which factors cause the instructor-differences.

Lastly, the results found in this study was based on a population of American community college students, and may not generalise to other educational institutions such as university, other age-groups, or to other countries. For example, it has been found that gender-gaps in academic performance is smaller in Scandinavian countries compared to North American countries [1].

Future work could incorporate several elements: an

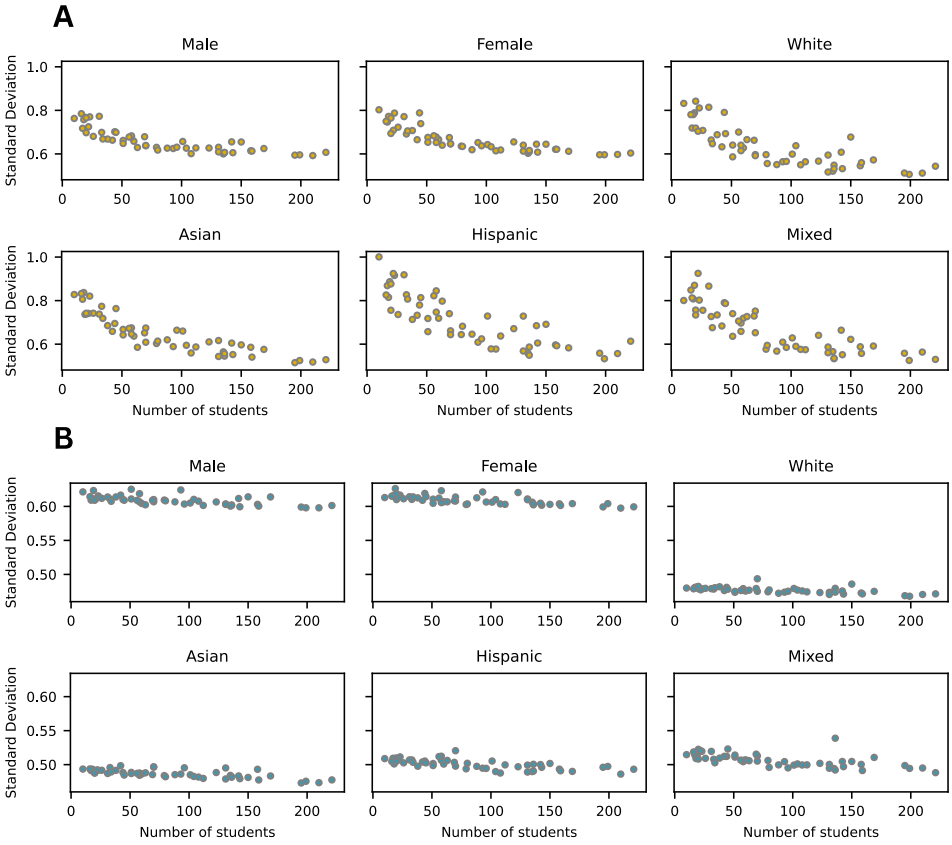

Fig. 6. Standard deviation of parameter estimations according to number of students belonging to each instructor of no-pooling models $M_{G/NP}$, $M_{E/NP}$ (A) and multilevel models $M_{G/ML}$, $M_{E/ML}$ (B).

increased sample size across ethnicities and instructors, instructor-level variables, and the scope of the study expanded to a more general population.

VI. Conclusion

This observational study examined gender- and ethnicity-effects in the academic performance of 4.463 American community college students in a first year English Literature course. Bayesian inference was used to explore these effects with three different approaches: complete pooling, no-pooling, and multilevel. The complete pooling model was deployed to investigate overall gender- and ethnicity-gaps in academic performance. The no-pooling and multilevel approach was utilized to examine whether gender- and ethnicity-effects varied among different instructors.

This study found differences in academic performance based on the effect of students' ethnicities and genders. These effects were best modelled by the multilevel models, indicating that there existed some degree of variability in the effects among instructors. In addition, White students had smaller variation across instructors compared to other ethnicities.

References

[1] Voyer, Daniel, and Susan D. Voyer. "Gender differences in scholastic achievement: a meta-analysis." Psychological bulletin 140.4 (2014): 1174.

[2] Hedges, Larry V., and Amy Nowell. "Sex differences in mental test scores, variability, and numbers of high-scoring individuals." Science 269.5220 (1995): 41-45.

[3] Lee, Jaekyung. "Racial and ethnic achievement gap trends: Reversing the progress toward equity?." Educational researcher 31.1 (2002): 3-12.

[4] Bali, Valentina A., and R. Michael Alvarez. "Schools and educational outcomes: What causes the "race gap" in student test scores?." Social Science Quarterly 84.3 (2003): 485-507.

[5] Hsin, Amy, and Yu Xie. "Explaining Asian Americans' academic advantage over whites." Proceedings of the National Academy of Sciences 111.23 (2014): 8416-8421.

[6] Sansone, Dario. "Why does teacher gender matter?." Economics of Education Review 61 (2017): 9-18.

[7] Li, Qing. "Teachers' beliefs and gender differences in mathematics: A review." Educational Research 41.1 (1999): 63-76.

[8] Tenenbaum, Harriet R., and Martin D. Ruck. "Are teachers' expectations different for racial minority than for European American students? A meta-analysis." Journal of educational psychology 99.2 (2007): 253.

[9] Redding, Christopher. "A teacher like me: A review of the effect of student–teacher racial/ethnic matching on teacher perceptions of students and student academic and behavioral outcomes." Review of educational research 89.4 (2019): 499-535.

[10] Huang, Chiungjung. "Gender differences in academic self-efficacy: A meta-analysis." European journal of psychology of education 28 (2013): 1-35.

[11] Nguyen, Vinh. "Collection of Practical Institutional Research Examples and Tutorials." (2023). GitHub repository.

[12] Hoffman, Matthew D., and Andrew Gelman. "The No-U-Turn sampler: adaptively setting path lengths in Hamiltonian Monte Carlo." J. Mach. Learn. Res. 15.1 (2014): 1593-1623.
[13] Betancourt, Michael, and Mark Girolami. "Hamiltonian Monte Carlo for hierarchical models." Current trends in Bayesian methodology with applications 79.30 (2015): 2-4.
[14] Papaspiliopoulos, Omiros, Gareth O. Roberts, and Martin Sköld. "A general framework for the parametrization of hierarchical models." Statistical Science (2007): 59-73.
[15] Patil, Anand, David Huard, and Christopher J. Fonnesbeck. "PyMC: Bayesian stochastic modelling in Python." Journal of statistical software 35.4 (2010): 1.
[16] Vehtari, Aki, et al. "Rank-normalization, folding, and localization: An improved for assessing convergence of MCMC (with discussion)." Bayesian analysis 16.2 (2021): 667-718.
[17] Vehtari, Aki, Andrew Gelman, and Jonah Gabry. "Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC." Statistics and computing 27 (2017): 1413-1432.

VII. APPENDIX

Student_ID a unique identifier of the student.

Term Describes the year and term (fall, spring, summer) in which the student was enrolled in ENG 1.

Section_ID an identifier that describes the ENG 1 section for which the student was enrolled in. Students with the same `Section_ID` are considered to be in the same class.

Instructor_ID an identifier that describes the instructor that taught the section. Students with the same `Instructor_ID` were taught by the same instructor.

Main_CourseID The course for which the student was enrolled in. For this sample data set, this field should be `ENG 1` for all rows.

Gender gender (one of: `Male`, `Female`, `Decline`).

Age the age of the student when they enrolled in ENG 1.

Ethnicity ethnicity (one of: `African American`, `Asian`, `Decline to State`, `Hispanic / Latino`, `Mixed Ethnicity`, `Native American`, `Pacific Islander`, or `White, Non-Hispanic`).

FirstGen 1 (yes) or 0 (no) indicating whether or not a student is a first generation college student.

Military 1 (yes) or 0 (no) indicating whether or not a student is using military veteran benefits (ie, a veteran).

FosterYouth 1 (yes) or 0 (no) indicating whether or not a student was a foster youth.

DSPS 1 (yes) or 0 (no) indicating whether or not a student has ever leveraged the Disabled Student Programs and Services (eg, has a learning disability and requires extended test time).

FinAid 1 (yes) or 0 (no) indicating whether or not a student has ever received federal or state financial aid at the district.

Units_Attempted_Beg_Of_Term the total number of units the student attempted at the district prior to the start of the term for which they enrolled in the current course.

GPA_Beg_Of_Term the student's cumulative grade point average (GPA) at the district prior to the start of the term for which they enrolled in the current course. Students that have not attempted any courses previously will have a GPA of 0.

TermUnitsAttempted the total number of units attempted in the current term.

K12_Student 1 (yes) or 0 (no) indicating whether or not a student is K-12 special admit student taking college courses.

First_Time_College_Student 1 (yes) or 0 (no) indicating whether or not a student is a first time college student (not a university student taking courses at a community college, not a continuing student, not a returning student, etc.).

Nonresident_Tuition_Exempt 1 (yes) or 0 (no) indicating whether or not a student is exempt from nonresident tuition based on AB 540.

International 1 (yes) or 0 (no) indicating whether or not a student is on an international student visa.

Nonresident 1 (yes) or 0 (no) indicating whether or not a student is not a resident of the college's state.

WR_Center 1 (yes) or 0 (no) indicating whether or not a student is enrolled in the writing center support section.

Main_Course_SuccessFlag 1 (yes) or 0 (no) indicating whether or not a student was successful (grades of A, B, C, or P) in the main writing course, `Main_CourseID`.

Main_Course_GradePoints a value of 4 (A), 3 (B), 2 (C), 1 (D), or 0 (F) indicating the student's grade in `Main_CourseID`.

HS_GPA student's high school GPA, if available.

Online 1 (yes) or 0 (no) indicating whether or not the student's `Main_CourseID` section was an online or in-person.

N_Center_Visits the student's number of writing center visits in the term.

Center_Attendance_Hours the total amount of time the student spent at the writing center in the term, measured in hours.

N_Conf the number of formal conferences the student had with an on-duty instructor at the writing center.

Fig. A1. Detailed description of the variables in the dataset.

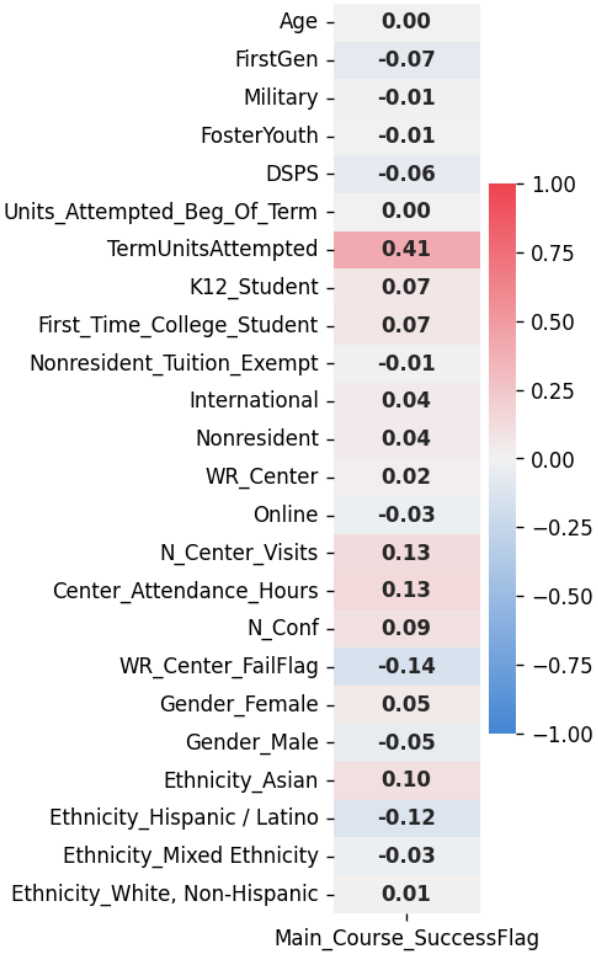

Fig. A2. The correlation of each independent variable from the dataset to the outcome variable, `MainCourseSuccessFlag`. Color red indicates a positive correlation to the outcome whereas blue indicates negative.

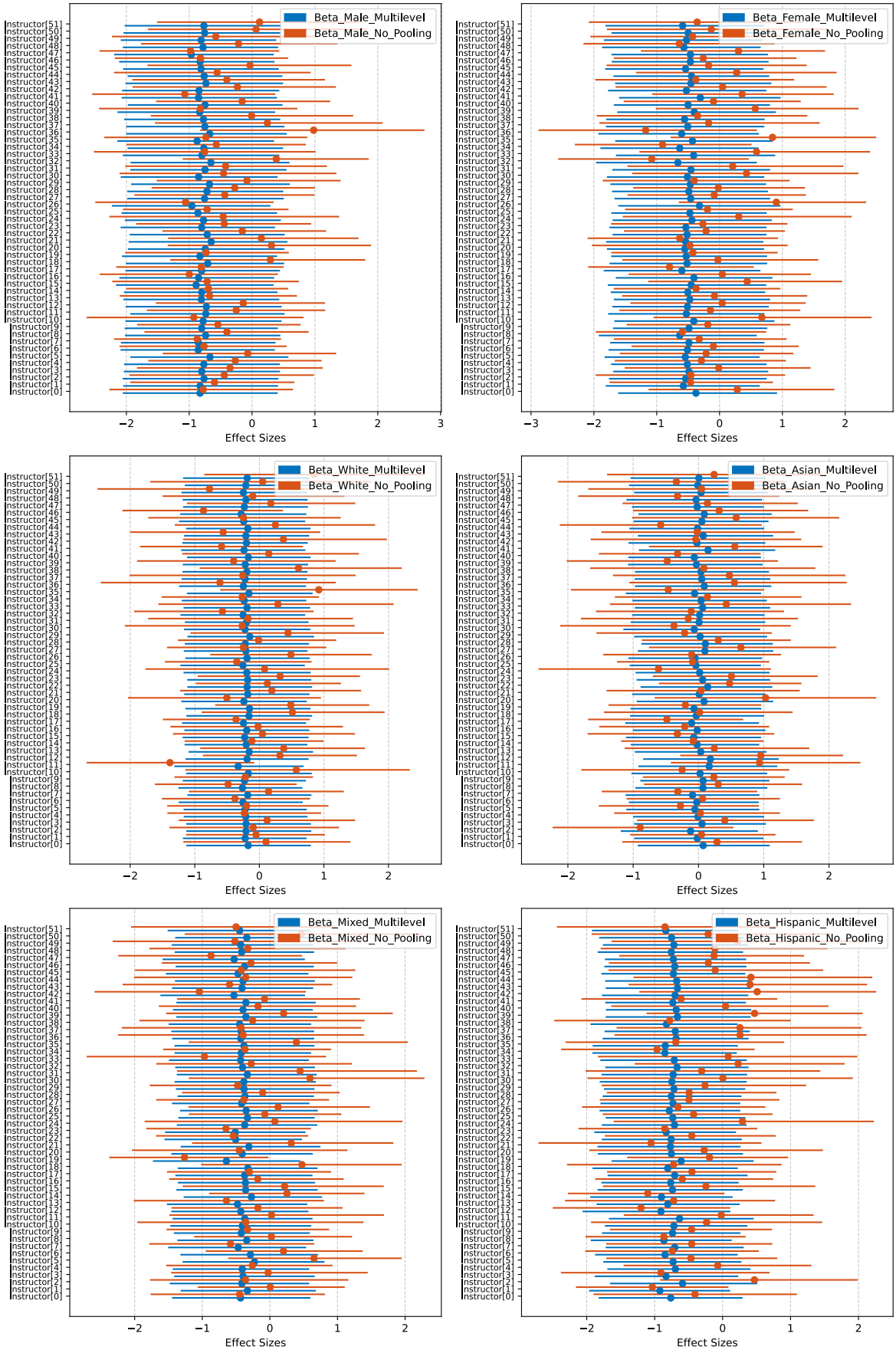

Fig. A3. The parameter estimate results for β_{Male}, β_{Female}, β_{White}, β_{Asian}, β_{Mixed}, and $\beta_{Hispanic}$ respectively by the instructor-level with $M_{A/ML}$ (blue) and $M_{A/NP}$ (red). Centered to the mean, the 95 % High Density Interval (HDI) is shown for individual instructors.

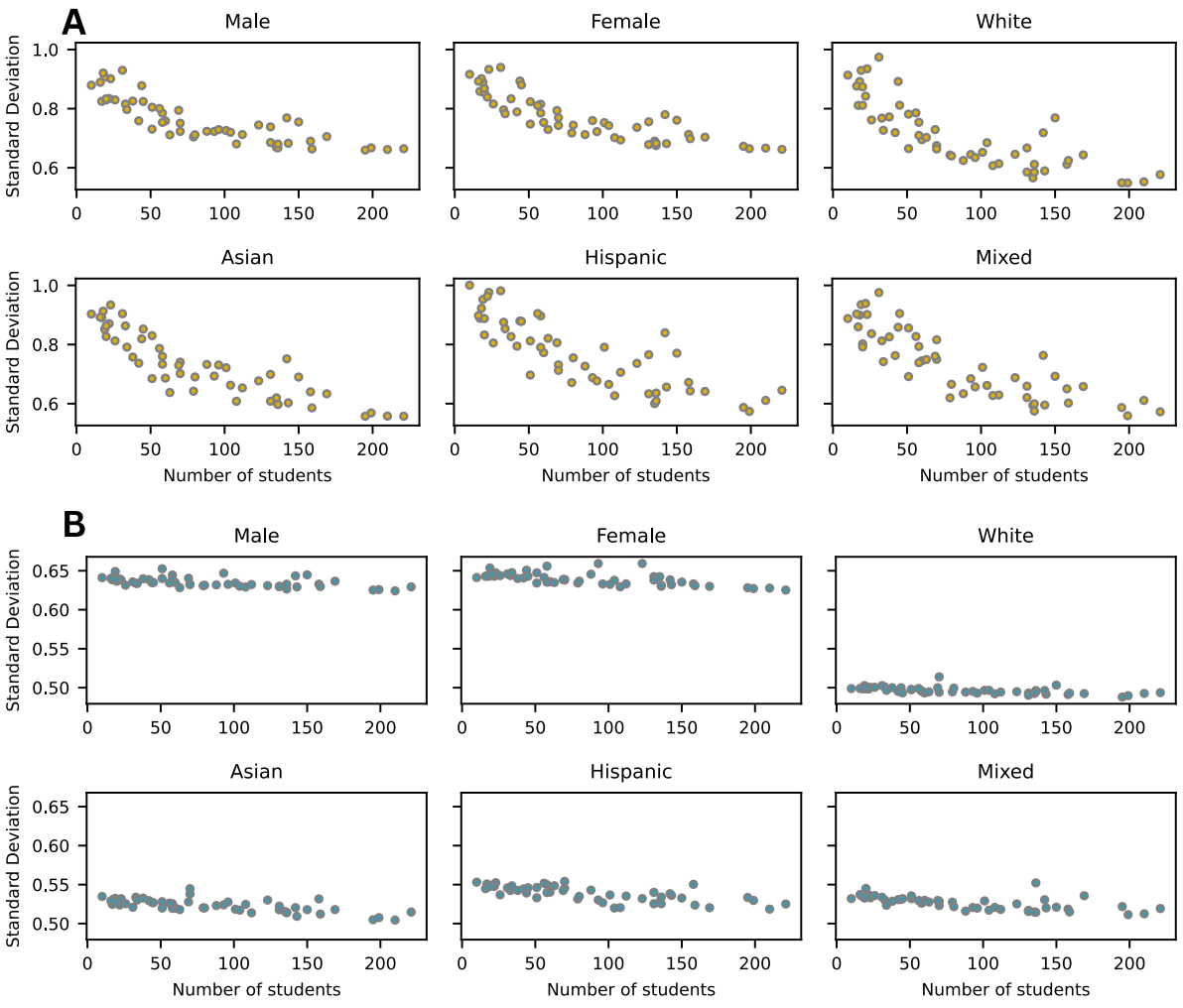

Fig. A4. Standard deviation of parameter estimations according to the number of students belonging to each teacher of no-pooling model $M_{A/NP}$ (A) and multilevel model $M_{A/ML}$ (B) in the adjusted models.