POLSCI643: Applied Bayesian Modelling

Final Replication Paper

Kashaf Ali (ka271)

INTRODUCTION

For this course project, I am replicating and extending my previous research from a course project conducted on the impact of the COVID-19 pandemic on academic performance, specifically focusing on gender disparities in educational outcomes. The original study investigated the mean scores in English, Math, and Science across different schools during the pre-pandemic (2018 and 2019) and post-pandemic (2021 and 2022) periods.

I employed a fixed effects model using the frequentist approach at the school level to account for variations both across schools and over time. The initial findings underscored a noteworthy performance gap between genders: prior to the pandemic, females consistently outperformed males by approximately 3 mean score points across the subjects examined. However, when examining the interplay of how the pandemic's effects varied academic performance by gender in the fixed effect model, it was evident that the presence of the pandemic led to a decrease of 1.36 mean score points larger for females when compared to males. This signifies, that although females perform better than their male counterparts in terms of academic performance, the pre-existing gap between them was narrowed due to how females reacted to the pandemic.

In this replication study, the utilisation of a Bayesian framework will facilitate a comprehensive re-examination of the original research's findings on gender-based academic performance disparities amidst the COVID-19 pandemic. By integrating prior knowledge from the initial study into a Bayesian model, I plan to acknowledge the previously observed gender gaps in academic scores, and incorporate it in my prior information to guide this analysis. Moreover, employing Bayesian methods would allow me to understand complex relationships between variables, accommodate nuances in the data and draw more refined conclusions regarding the evolving impact of the pandemic on gender-specific academic performance trends.

To reiterate from the original study, my goal is to answer the following main research question in this replication study through a bayesian approach: *What is the causal effect of the Covid-19 pandemic on student's test scores by gender in New York City?*

DATASET

The replication study will use the same Report Card datasets from the official website of the New York State Education Department for four academic years, namely, 2017-2018, 2018-2019, 2020-2021, and 2021-2022 to understand our problem[1]. These datasets are at the school-level and provide information on mean school results for ELA, Math, and Science assessments at the elementary and intermediate-level, as well as contain mean school assessment results for different groups like gender, race and others. Since this is a panel dataset, it offers information regarding

[1] New York State Education Department. (n.d.). Downloads. Retrieved from https://data.nysed.gov/downloads.php

assessment scores from the same schools before and after the COVID-19 pandemic outbreak. We observed pre-pandemic and post-pandemic mean test scores by gender and subject, and other demographic variables such as race, and whether English is their first language for different schools in New York.

This replication study will use the same data cleaning process used by the original study as the focus of this replication is limited to creating improved versions of the existing model specifications from a bayesian perspective instead of transforming the existing data for the research. The raw report card datasets used in this study included different datasets for English, Math and Science at the school level for different years. Each one of these datasets includes information on the school's district, county, needs assessment indicator, overall status, school type, and information on school teachers, in addition to mean test scores from gender, racial and other groups at the school level. While all this information was pertinent towards understanding the gendered impact of COVID-19 on students' test performance, the following key features were used in the final model. I will continue to use the same variables in the replication study.

Variable Name	**Description**
Gender Mean Score	Mean test scores for the school
Female	1 if the mean school test score was for the female group 0 otherwise
Post Covid	1 years = 2021 or 2022 0 year = 2018 or 2019
Race Gap	1 if mean school test score for white racial group is higher than that of the non-white racial group 0 otherwise
ELL Gap	1 if the mean school test score for English Language Learners (ELL) is higher than that of the non-English Language Learners group 0 otherwise.
Entity ID	School Id
Year	Year
Assessment Name	Assessments for English, Math and Science English: ELA3, ELA4, ELA5, ELA6, ELA7, ELA8 Math: MATH3, MATH4, MATH5, MATH6, MATH7, MATH8 Science: Science4, Science8

It is important to note that these features were engineered from the raw data to be used meaningfully in the model. Before feature engineering, the missing values in the mean test scores data were also imputed using the mean test score of that subject in that specific school in that specific year. More details on the data cleaning process are shared in the original paper.

RESEARCH DESIGN

The design of the original causal inference study used a combination of difference-in-difference (Diff-in-Diff) and fixed effects (FE) to estimate the impact of the COVID-19 pandemic on academic achievement while controlling for unobserved variation across schools. While Diff-in-Diff was used to compare the change in mean test scores for females and males at the school level across different subject assessments before and after the pandemic and provides an estimate of the causal effect of the pandemic on female test scores, it may not account for unobserved variation between different schools. I added fixed effects at the school level to the Diff-in-Diff model as the fixed effects will capture the differences in the unobserved school-level characteristics that are constant over time.

The replication study would contribute to the research design in three essential ways: (i) replicating the existing fixed effects model, (ii) identifying and removing Science Assessment outliers in the data and (iii) introducing a hierarchical model as an innovative statistical approach to answer this question.

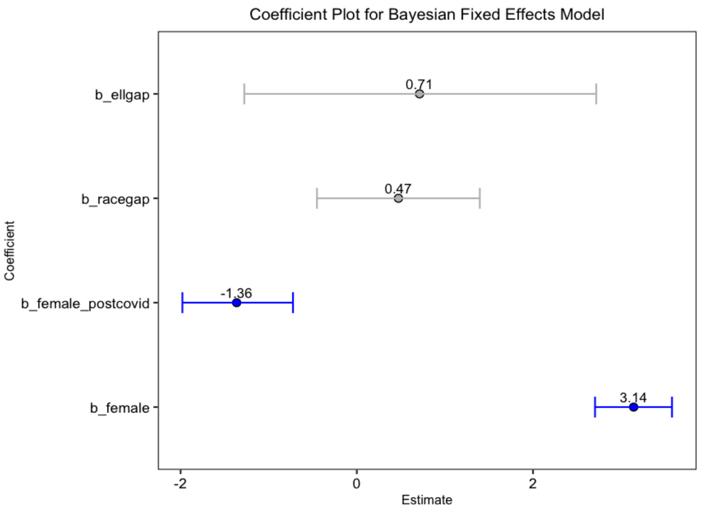
REPLICATION - FIXED EFFECTS MODEL

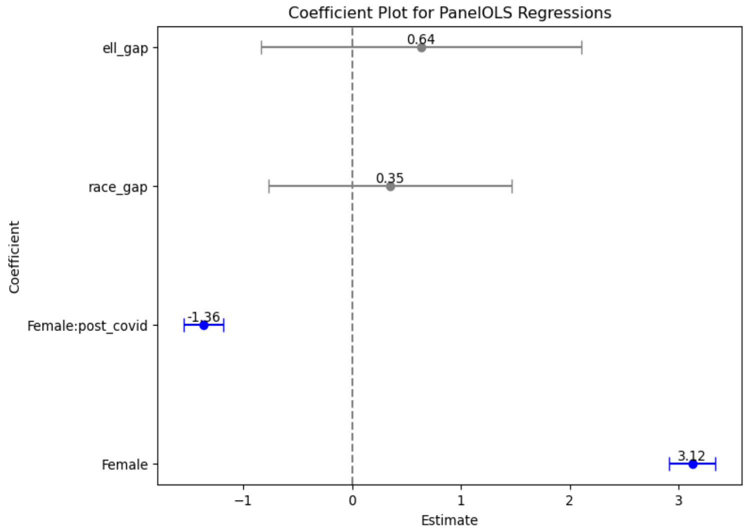
The original paper set up a fixed effects model by using the Entity Effects (school level effects) and Time Effects (year effects) variables in the regression using PanelOLS on python. The inclusion of these fixed effects variables in the regression formula (Entity Effects and Time Effects) indicates that the model will estimate coefficients for other variables while accounting for the effects specific to each school and year. Below is the regression equation I redesigned for the bayesian approach of the fixed effects model. The model was set up in Jags and the code is included in Appendix 1.

$$Gender\ Mean\ Scores\ =\ b_0\ +\ b_{female} \times Female\ +\ b_{postcovid} \times Post\ Covid\ +b_{ellgap} \times Ell\ Gap\ +$$
$$b_{racegap} \times Race\ Gap\ +\ b_{female-postcovid} \times\ Female\ \times\ Post\ Covid\ +$$
$$b_{school} \times Entity\ CD[i]\ +b_{assess} \times Assessment\ Name[i]\ +\ b_{year} \times\ Year[i]$$

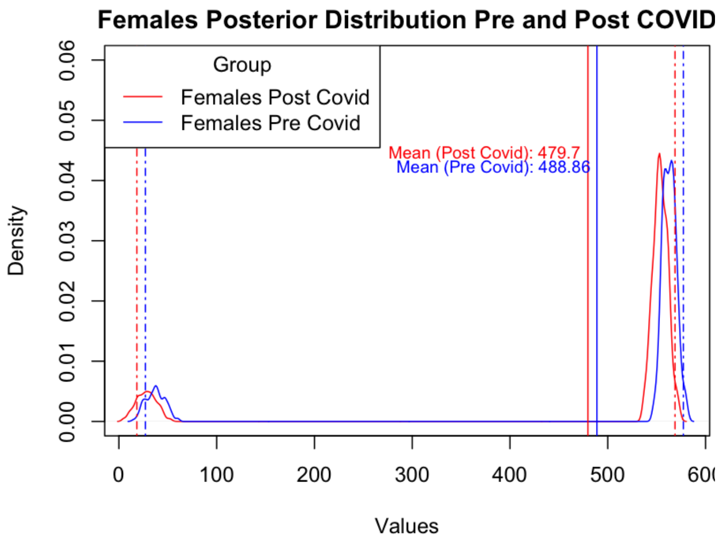
This regression equation includes dummies for each school, year and assessment to account for school, assessment and year effects on the gender mean score. Moreover, I also added an interaction term between female and post_covid to this regression because it was used in the original study as a difference-in-differences (Diff-in-Diff) estimator to evaluate whether the change in Gender Mean Score before and after the post-Covid period varies for females compared to males.

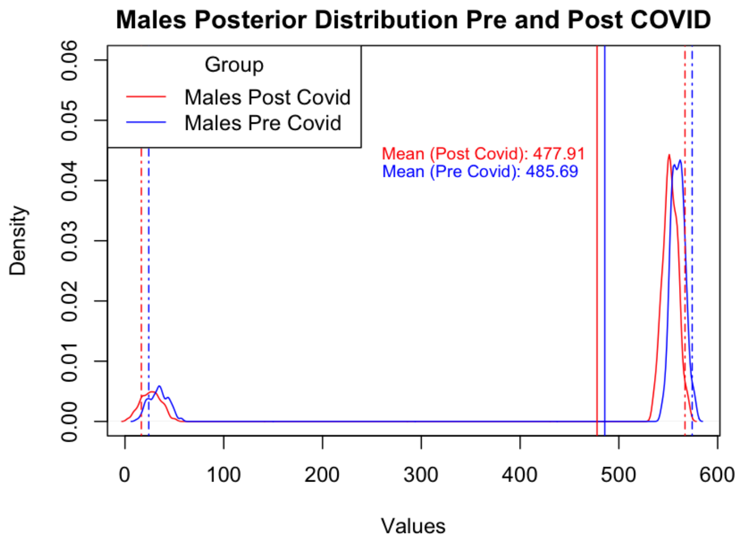
It is important to note that the mean of the beta distribution generated by the Bayesian fixed effects model was very close to the coefficient results from the original study's PanelOLS regression. Results from the original study showed that females scored about 3 mean score points higher than their male counterparts without COVID-19, and the presence of the COVID-19 produced a decrease larger for females than for males, and the size of that difference in decrease was 1.36 mean score points. Based on the frequentist approach, both these results were statistically significant at 95% confidence level. In comparison, the results from the Bayesian fixed effects model also showed that on average females scored about 3 mean score points higher than their male counterparts without COVID-19 and the decrease caused for girls compared to boys because of the pandemic was 1.36 mean score points higher. Hence, it is evident that the replication of the regression results through the Bayesian approach was successful. The coefficient plots for some coefficients from the replicated and original study are shown below (Appendix 2 shows a more detailed summary of all the results from the Bayesian model).



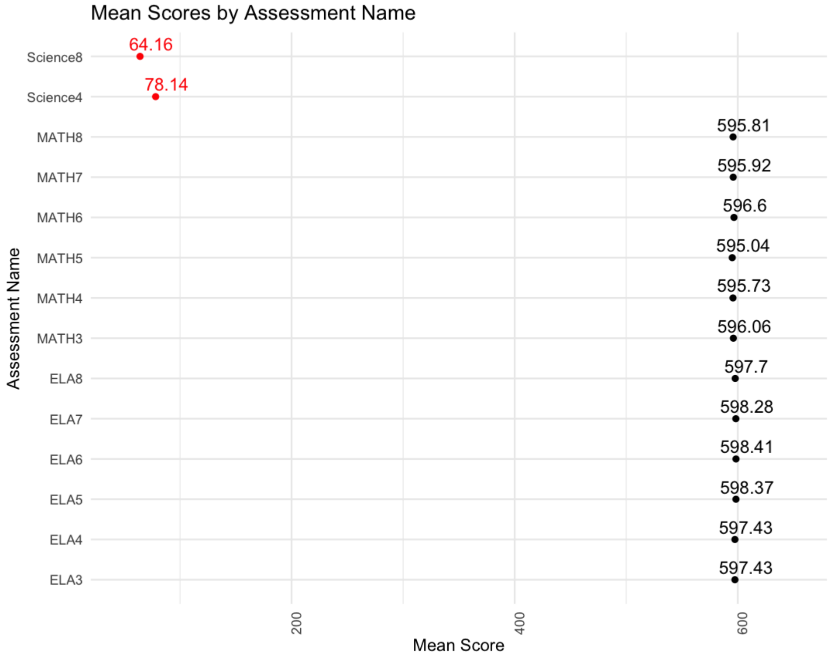


In addition to this, the Bayesian fixed effects model further allowed me to explore the posterior distribution of males and females before and after the pandemic to understand if the gender mean scores changed. The two plots below show the pre and post covid mean test score posterior distributions for females and males. It is evident that both females and male students were affected adversely due to the pandemic, with females being impacted more as their mean test scores (values) declined more compared to what they were before the pandemic in comparison to their male counterparts.

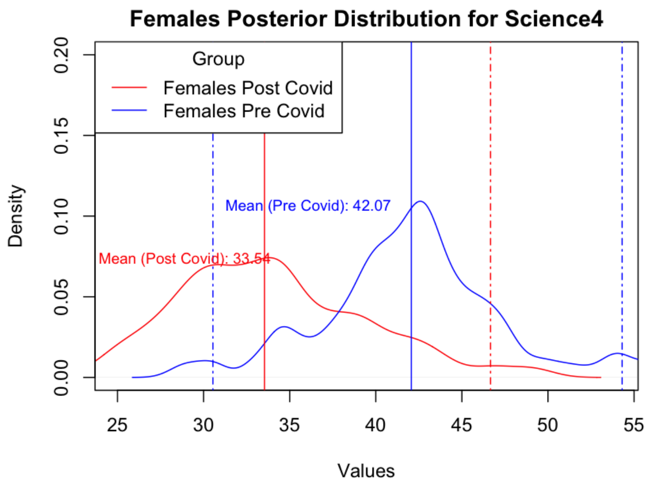


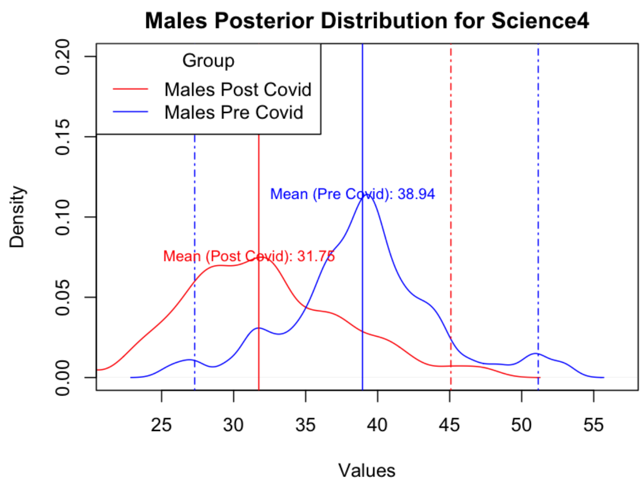


Although we are getting the expected results, I found the distribution in the plots above to be skewed, where it was estimating the mean test scores to be close to zero. Hence, I decided to investigate the data further to understand what the mean test scores look like across different assessments in the actual data. Based on the plot below the mean scores for science assessments were on a different scale and hence, were causing a skewed distribution.



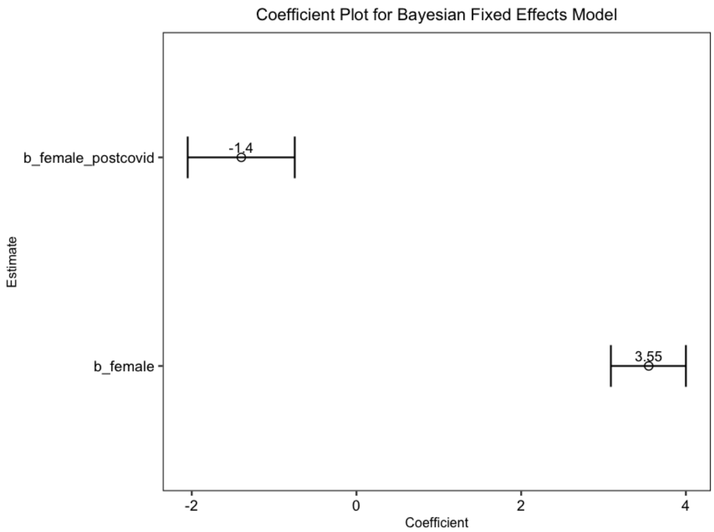
To further investigate how females and males were performing in Science, I zoomed in on the skewed portion of the distribution and looked at female and male scores pre-covid and post-covid for Science4 assessment as an example. The plots below show the posterior distribution for Science4 shows the same trend where females mean test scores decline more post-covid than their male counterparts, it is important to note the x-axis here which represents very low test score values, indicating that the science assessment scores were skewing our overall distribution.



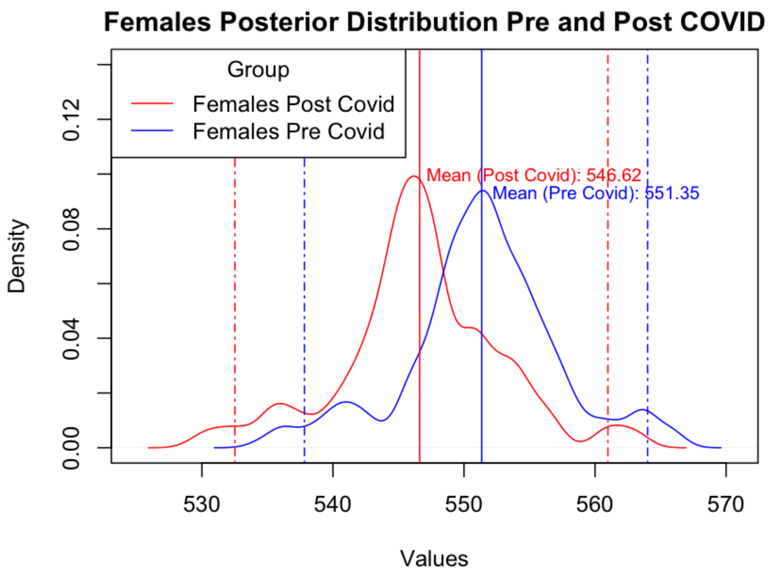


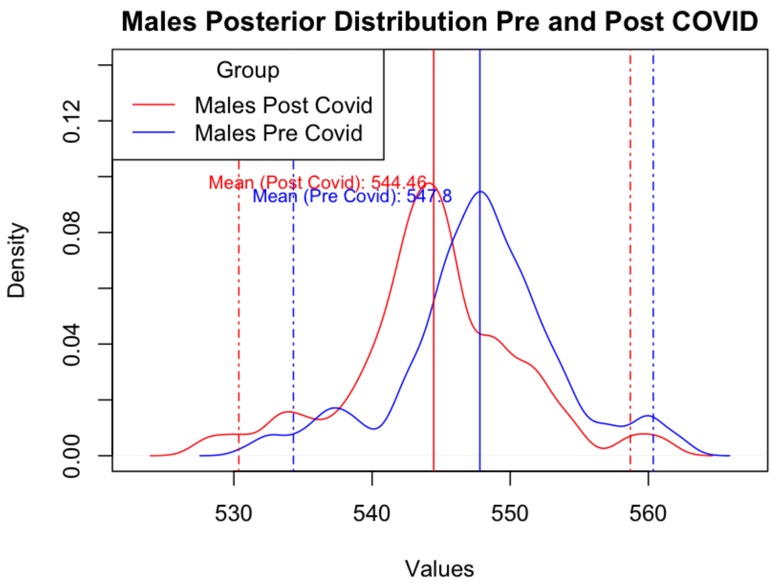
As the Science assessment scores were outliers in the dataset, and the original data source did not share more information on scaling them to match the range for English and Math, I decided to exclude them from my analysis and conducted the rest of the replication excluding Science test scores.

After running the fixed effects model again without the Science assessments test scores, our coefficients for the impact on female and the female and post_covid interaction term increased as shown in the coefficient plot below because the lower science scores were skewing the results towards a lower mean. The distance between 2.5 and 97.5 quantiles also became slightly narrower because most of the scores were in a closer range to each other after the exclusion of Science assessments (Appendix 3 shows a more detailed summary of all the results from the Bayesian model).



I also plotted the mean of the posterior distribution for females and males for this updated model as shown below. The distribution is no longer skewed, but the results are still consistent with the original study, where females suffer more from the pandemic as their mean test scores decline more than their male counterparts post-pandemic.





EXTENSION - HIERARCHICAL MODEL

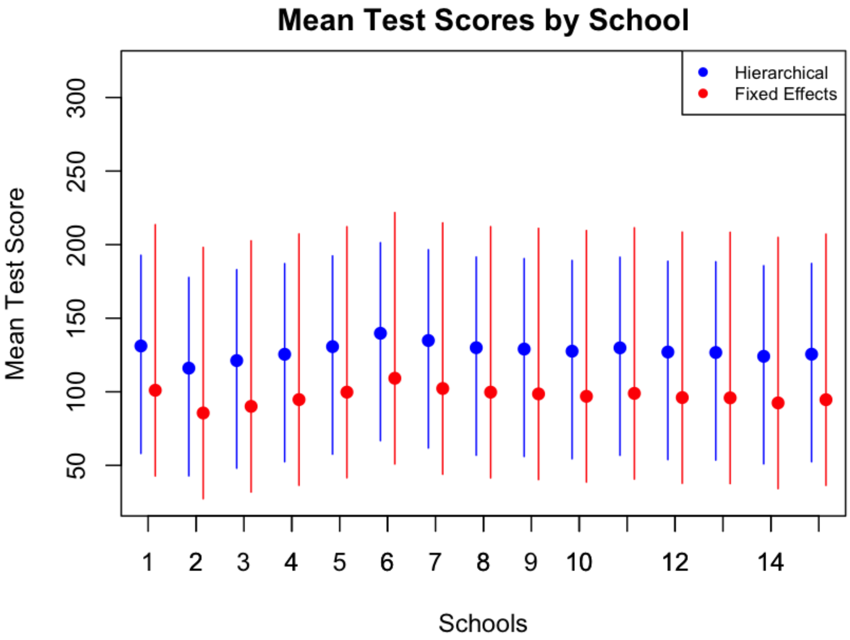
Although the original study was limited to a fixed effects model with difference-in-difference estimator, I extended the replication study to conduct the same analysis with a hierarchical model as well. While a fixed effects model assumes that each individual unit (e.g. schools) has its own unique effect or intercept that remains constant across observations, a hierarchical models assumes a hierarchical or nested structure in the data, where lower-level units (e.g., assessment scores) are nested within higher-level units (e.g., schools), and variability exists at multiple levels. Employing a Bayesian approach to build a hierarchical model for this problem will also allow for the incorporation of prior distributions at different levels, capturing uncertainties not only at the individual level but also at higher group or cluster levels. This approach enables the estimation of both fixed effects (representing specific group differences) and random effects (representing variability between groups).

This hierarchical model equation below is designed to explore the relationships between various predictors and the Gender Mean Scores variable, representing educational performance, through a Bayesian framework using JAGS (code is shown in Appendix 4). The model defines a linear relationship between the Gender Mean Scores and predictors such as gender, post-COVID status, race and English language learner gaps, school, year, and assessment influences. The coefficients associated with these predictors are assigned prior distributions, allowing for estimation and inference of their effects on educational performance. However, it also incorporates hierarchical structures for school-specific effects by assuming that the coefficients related to schools (e.g., "b_school," "r_female," "r_postcovid," "r_femalepostcovid") vary across schools. The parameters governing these school-specific effects are themselves given prior distributions based on a multivariate normal distribution (dmnorm), allowing for the estimation of variability between schools while considering potential relationships among these effects.

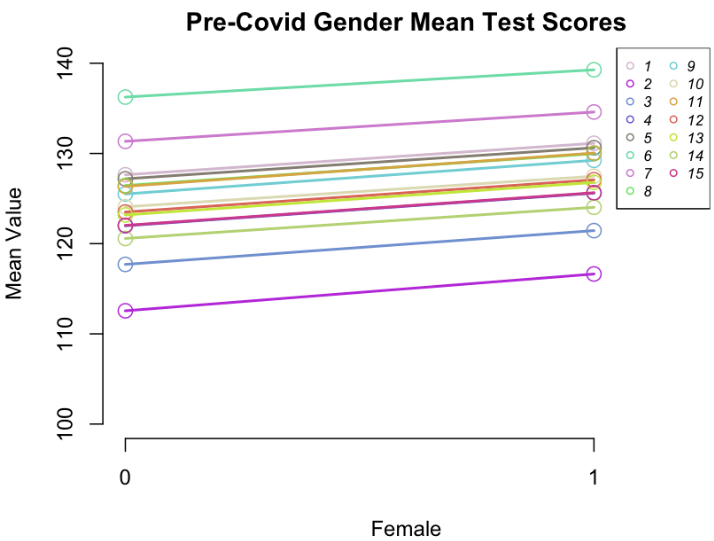
$$
\begin{aligned}
Gender\ Mean\ Scores \ = \ & b_0 \ + \ r_{female}[ENTITY\_CD[i]] \times Female \ + \\
& r_{postcovid}[ENTITY\_CD[i]] \times Post\ Covid \ + \\
& r_{female-postcovid}[ENTITY\_CD[i]] \times Female \times Post\ Covid \ + \\
& b_{ellgap} \times Ell\ Gap \ + \ b_{racegap} \times Race\ Gap \ + \\
& b_{school} \times Entity\ CD[i] \ + b_{assess} \times Assessment\ Name[i] \ + \ b_{year} \times Year[i]
\end{aligned}
$$

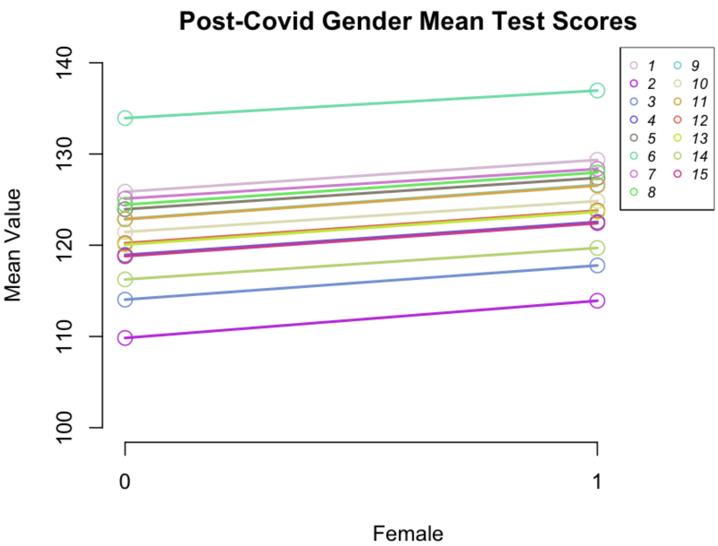
As shown in the equation above, the model incorporates random intercepts and slopes at the school level, which allows us to account for the heterogeneity among schools in terms of their baseline performance (mu.school). The random intercept for each school in this case indicates that each school has its own baseline or starting point for the Gender Mean Score, and captures the variability among schools that cannot be explained by the fixed effects in the model. On the other hand, random slopes for the variables related to female (r_female), post_covid (r_postcovid), and the interaction term between female and post_covid (r_femalepostcovid) also show how the effects of gender, post-COVID status, and their interaction differ across schools.

The plot below shows the mean test scores by school from both the fixed and hierarchical models. As the hierarchical model induces "shrinkage" of the individual parameter estimates towards the population mean, the mean test scores by school are different in the hierarchical model and slightly closer to each other.

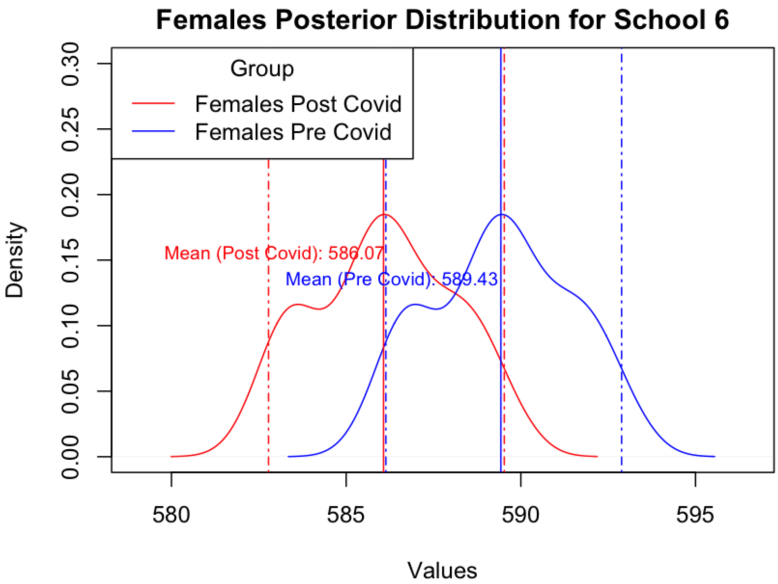


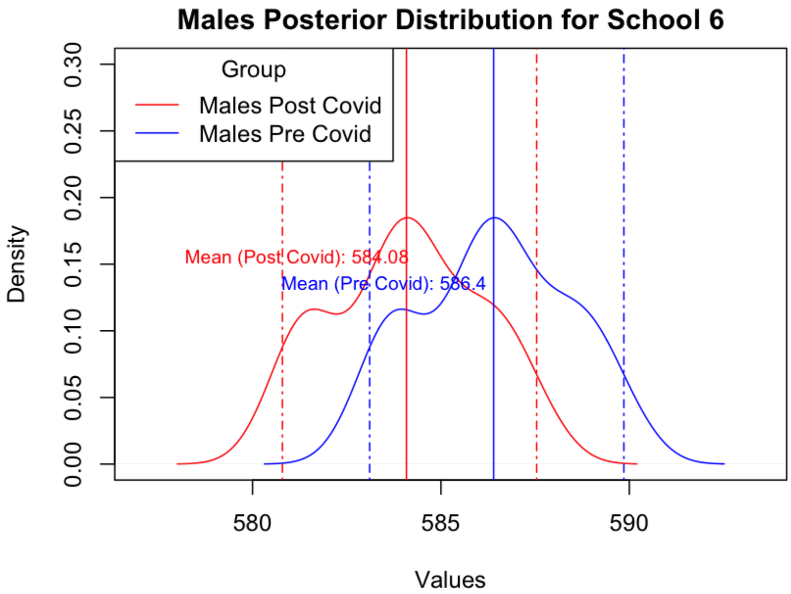
The plot below shows the pre-covid and post-covid Gender Mean Test Scores for female and male students. It is evident that for all schools, females had higher test scores than their male counterparts both before and after covid (females are represented by 1 and males are represented by 0 in the plots below). Moreover, the hierarchical model also highlights that there are differences in the baseline scores between different schools. For instance, school 6 (light green) and school 2 (purple) had higher and lower test scores on average for both females and males in the pre-covid period, respectively. Although the mean test scores decreased post-covid, these two schools continued to be the highest and lowest performing schools, respectively.

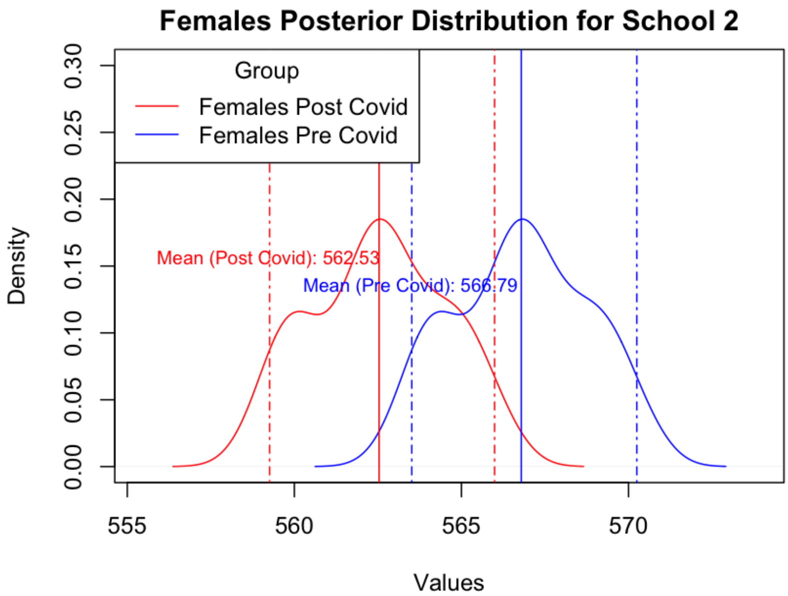


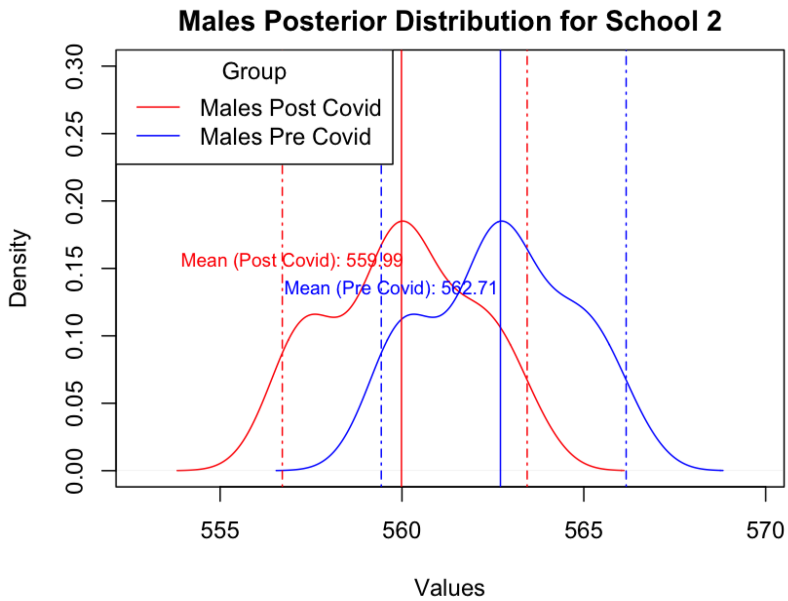


The hierarchical model also allows us to look at the post distributions of mean test scores separately for each school to understand how school-level factors would impact the mean test scores of female and male students. Below are four plots showing the posterior distributions for schools 2 and 6 showing separate mean posterior distributions for female and males both pre-covid and post-covid. It is evident that for school 2, which had the lowest baseline test scores for both males and females, the impact of the pandemic was worse than it was for school 6, which had the highest baseline test scores for both males and females. The mean test scores also decreased more for females than they did for males, and for school 2, this difference was also higher. Hence, this indicates that schools that were worse off before the pandemic were affected more from it, and females in these low performing schools were affected the most.









It is also important to note that we saw a greater impact on the mean test scores in the fixed effects model, however, the hierarchical model allows us to explore how that impact might not be the same for different schools more specifically. Hence, we can say that there was an impact of the pandemic and females suffered more than males despite having higher pre-covid mean test scores. However, it is difficult to say if the magnitude of this impact in itself was meaningful at an overall level or not. While the frequentist approach relies on statistical significance, in this case, the magnitude might be different for each school and it would be important to investigate that further.

LIMITATIONS

This replication study inherits some of the existing limitations of the original study, and also introduces new challenges that need to be considered before using this study for policy implementation.

- **Missing Values:** it is important to highlight that the original dataset included a lot of missing values, and the original study had to limit the analysis to schools that completed the survey in all four years. To address this concern, the original study used mean imputation based on the same school, year, and subject (original study's appendix section provides more detail on this process). In the replication study, I decided to use the same dataset for further analysis from a bayesian approach. However, it is important to acknowledge that while mean imputation is a common approach used for dealing with missing values, this data cleaning and imputation process might introduce bias in our causal inference estimates.
- **External Validity**: It is important to note that this study solely focuses on elementary and middle school districts in New York and the number of final schools used in this analysis was limited due to the missing data in the original dataset. Hence, these results may not extend to all kinds of schools everywhere in the world, especially other schools in other less developed or less gender-equal regions in the United States compared to New York. In addition to this, my final analysis is based on English and Math assessments because I excluded unscaled Science assessment results from our model. While my initial model showed similar trends for Science Assessments, perhaps, incorporating more data will offer more generalizable results for this subject.
- **Targeted Focus on Gendered Impact:** Since, this study is focused on the gendered dimension of the pandemic's effects, other social, economic, racial, language-related factors were not actively explored in this study. Since these factors might also contribute to different effects of the pandemic for different schools and students, they might be worth exploring in future studies. Although this limits the generalizability of our results across all groups of students, a more targeted focus on the gendered impact in this study allowed us to highlight the potential disproportionate impact on females and highlight challenges females may encounter in the education system.

CONCLUSION

In conclusion, this replication study allowed me to replicate a fixed effects model from a Bayesian approach to highlight how females, who were performing better than their male counterparts in school before the pandemic, were affected more by the pandemic than males. While these results remained consistent in a Bayesian fixed effects model, an extension to the study also allowed me to further dive into school-level effects using a hierarchical model. Results from the hierarchical model further corroborated the original study's results, but also added that schools which were performing worse than average before the pandemic were affected more by the pandemic than schools which were performing better. This trend continues to hold for both females and males, and females in low performance schools were impacted more than females in high performance school post-pandemic.

APPENDIX

Appendix 1

Bayesian Fixed Effects Model Setup in JAGS

```
# Model specification in JAGS
model_FE <- "model {
  for (i in 1:N) {
    Gender_MEAN_SCORE[i] ~ dnorm(mu[i], tau)
    mu[i] <- b_0+ b_female*Female[i] + b_postcovid*post_covid[i] +
             b_racegap*race_gap[i] + b_ellgap*ell_gap[i] +
             b_school[ENTITY_CD[i]] + b_year[YEAR[i]]  +
             b_assess[ASSESSMENT_NAME[i]] +
             b_female_postcovid * Female[i] * post_covid[i]
  }

  # Priors for coefficients
  b_0 ~ dnorm(0, 0.001)
  b_female ~ dnorm(0, 0.001)
  b_postcovid ~ dnorm(0, 0.001)
  b_racegap ~ dnorm(0, 0.001)
  b_ellgap ~ dnorm(0, 0.001)
  b_female_postcovid ~ dnorm(0, 0.001)

   # School effects
  for (i in 1:unique_schools) {
    b_school[i] ~ dnorm(0, 0.001)
  }

  # Year effects
  for (i in 1:unique_years) {
    b_year[i] ~ dnorm(0, 0.001)
  }

  # Assessment effects
  for (i in 1:unique_assessments) {
    b_assess[i] ~ dnorm(0, 0.001)
  }

  tau ~ dgamma(0.001, 0.001) ## (inverse) Variance
  sigma2 <- 1 / tau
}"

write(model_FE,file="educov_FE.jags")
```

Appendix 2

Results from Bayesian Fixed Effects Model

Displaying Mean, Standard Deviation, and Quantiles (2.5%, 25%, 50%, 75% and 97.5%)

	Mean	SD	2.5%	25%	50%	75%	97.5%
b_0	76.3418	32.4402	33.9564	41.86498	75.8500	104.1278	130.0554
b_assess[1]	147.1704	16.8117	106.1768	139.77073	147.1418	162.0308	170.6249
b_assess[2]	147.1724	16.8099	106.1636	139.83414	147.1484	162.0513	170.6348
b_assess[3]	148.0884	16.8139	107.0958	140.60760	148.0848	162.8372	171.5790
b_assess[4]	148.1430	16.8123	107.1410	140.66941	148.1354	162.8863	171.6129
b_assess[5]	148.0177	16.8095	107.0030	140.62308	147.9941	162.8820	171.4903
b_assess[6]	147.4467	16.8132	106.4163	140.00910	147.4117	162.1982	170.9680
b_assess[7]	145.7844	16.8103	104.7410	138.30305	145.7688	160.5648	169.2704
b_assess[8]	145.4446	16.8138	104.4158	137.99240	145.4259	160.2354	168.9430
b_assess[9]	144.7910	16.8106	103.7666	137.36418	144.7908	159.5330	168.2755
b_assess[10]	146.2858	16.8119	105.2787	138.77314	146.2729	161.0309	169.7470
b_assess[11]	145.5985	16.8156	104.5945	138.14038	145.5700	160.3422	169.0785
b_assess[12]	145.4923	16.8058	104.4740	138.09832	145.4817	160.2504	168.9468
b_assess[13]	-372.1717	16.8131	-413.2168	-379.61701	-372.2036	-357.3367	-348.7327
b_assess[14]	-386.1582	16.8087	-427.1880	-393.52701	-386.1834	-371.3206	-362.6770
b_ellgap	0.7122	1.0132	-1.2772	0.02511	0.7135	1.3846	2.7180
b_female	3.1422	0.2262	2.7038	2.98763	3.1424	3.2972	3.5778
b_female_postcovid	-1.3640	0.3240	-1.9796	-1.58426	-1.3657	-1.1460	-0.7243
b_postcovid	-9.3151	4.8310	-19.1337	-12.23981	-9.3022	-5.2968	-1.7347
b_racegap	0.4725	0.4624	-0.4522	0.17465	0.4675	0.7756	1.3962
b_school[1]	88.0780	45.3413	34.4431	47.60590	69.4101	125.0392	179.6264
b_school[2]	72.8276	45.3463	19.1817	32.33780	54.0949	109.7204	164.4346
b_school[3]	77.8994	45.3437	24.3103	37.45860	59.2157	114.8431	169.5585
b_school[4]	82.8270	45.3412	29.2262	42.30311	64.1497	119.6861	174.4366
b_school[5]	87.5579	45.3419	33.9629	47.08480	68.8061	124.4394	179.1748
b_school[6]	96.3799	45.3460	42.7633	55.81497	77.6779	133.3004	188.1612
b_school[7]	89.2481	45.3455	35.6143	48.74312	70.5324	126.1662	180.8921
b_school[8]	87.2881	45.3431	33.6850	46.78089	68.5392	124.1633	179.0041
b_school[9]	86.6949	45.3471	33.0842	46.18998	67.9675	123.6549	178.3756
b_school[10]	84.9123	45.3456	31.2842	44.40188	66.1465	121.8344	176.5939
b_school[11]	86.8125	45.3415	33.2387	46.32502	68.1317	123.7262	178.5207
b_school[12]	84.1457	45.3423	30.4962	43.64709	65.4365	121.0301	175.8804
b_school[13]	84.0306	45.3455	30.4257	43.48029	65.3294	120.9520	175.6595
b_school[14]	80.4924	45.3497	26.8443	40.02703	61.7724	117.3498	172.2130
b_school[15]	83.1474	45.3461	29.6057	42.66017	64.4017	120.0598	174.9104
b_year[1]	287.9358	11.8558	267.0124	279.29853	288.9590	297.0584	307.6971
b_year[2]	287.8001	11.8579	266.8581	279.17749	288.8410	296.9461	307.5608
b_year[3]	296.5451	9.3725	277.2538	287.74338	299.2094	303.8879	310.0079
b_year[4]	295.8317	9.3739	276.5302	287.03587	298.5028	303.1740	309.2950
sigma2	10.9893	0.3890	10.2592	10.71992	10.9812	11.2469	11.7771

Appendix 3

Results from Bayesian Fixed Effects Model (without Science Assessments)

Displaying Mean, Standard Deviation, and Quantiles (2.5%, 25%, 50%, 75% and 97.5%)

	Mean	SD	2.5%	25%	50%	75%	97.5%
b_0	70.6348	28.2310	22.9348	43.9123	73.2532	91.355	120.0151
b_assess[1]	63.0284	19.6963	33.4517	49.2644	60.4844	79.933	97.5380
b_assess[2]	63.0276	19.6967	33.4548	49.2589	60.4262	79.939	97.5288
b_assess[3]	63.9719	19.6928	34.4305	50.2120	61.3783	80.870	98.4801
b_assess[4]	64.0013	19.6953	34.4528	50.2022	61.3760	80.942	98.5304
b_assess[5]	63.8712	19.6951	34.3185	50.0644	61.3129	80.824	98.4281
b_assess[6]	63.2961	19.6924	33.7001	49.5479	60.6692	80.250	97.8496
b_assess[7]	61.6593	19.6890	32.0771	47.9079	59.0793	78.574	96.2323
b_assess[8]	61.3311	19.6926	31.7908	47.5112	58.6991	78.234	95.8212
b_assess[9]	60.6841	19.6894	31.1484	46.8734	58.0751	77.608	95.2267
b_assess[10]	62.1601	19.6939	32.6112	48.3748	59.5238	79.088	96.6703
b_assess[11]	61.4822	19.6935	31.9361	47.7206	58.8413	78.389	96.0417
b_assess[12]	61.3701	19.6947	31.8535	47.6024	58.7661	78.283	95.9052
b_ellgap	-0.1449	31.3831	-61.2426	-21.5494	-0.0633	21.113	62.0927
b_female	3.5457	0.2321	3.0871	3.3918	3.5442	3.701	4.0013
b_female_postcovid	-1.3933	0.3303	-2.0412	-1.6130	-1.3962	-1.171	-0.7493
b_postcovid	48.4805	14.4156	27.7888	34.5771	45.8664	63.581	70.0901
b_racegap	-0.1326	0.4317	-0.9468	-0.4315	-0.1421	0.158	0.7316
b_school[1]	95.2494	52.7970	35.5460	47.9931	82.6380	125.361	212.8802
b_school[2]	79.8199	52.7934	20.1545	32.5771	67.1998	109.988	197.4604
b_school[3]	84.3099	52.7921	24.6255	37.0577	71.7385	114.480	201.9587
b_school[4]	88.9170	52.7940	29.2211	41.7295	76.3275	119.106	206.7349
b_school[5]	93.9745	52.7942	34.2775	46.7395	81.4038	124.036	211.5671
b_school[6]	103.4096	52.7945	43.7287	56.1631	90.8614	133.528	220.9813
b_school[7]	96.4535	52.7960	36.7843	49.1949	83.8990	126.571	214.1589
b_school[8]	93.9685	52.7971	34.2996	46.7346	81.4198	124.155	211.5999
b_school[9]	92.7974	52.7865	33.0982	45.5556	80.2368	122.862	210.5473
b_school[10]	91.1330	52.7985	31.4458	43.8843	78.5151	121.312	208.7493
b_school[11]	93.1416	52.7963	33.4482	45.8647	80.5415	123.216	210.7928
b_school[12]	90.2950	52.7947	30.5938	43.0494	77.7000	120.377	207.9853
b_school[13]	90.0909	52.7972	30.3632	42.9313	77.5264	120.173	207.6629
b_school[14]	86.6676	52.7891	26.9825	39.3942	74.0408	116.721	204.2787
b_school[15]	88.8538	52.7867	29.1637	41.6293	76.2996	118.874	206.5250
b_year[1]	371.2802	16.4837	327.2893	363.8174	378.9167	382.804	388.1408
b_year[2]	371.3344	16.4831	327.3433	363.8924	378.9742	382.876	388.1947
b_year[3]	323.0568	17.6768	294.9792	309.0826	321.8723	340.296	349.8846
b_year[4]	321.9179	17.6762	293.8571	307.9532	320.7373	339.168	348.7623
sigma2	9.7096	0.3698	9.0138	9.4547	9.6978	9.956	10.4615

Appendix 4

Bayesian Hierarchical Model Setup in JAGS

```
# Model specification in JAGS
model_H <- "model {
  for (i in 1:N) {
    Gender_MEAN_SCORE[i] ~ dnorm(mu[i], tau.y)
    mu[i] <- b_0 + r_female[ENTITY_CD[i]]*Female[i] +
               r_postcovid[ENTITY_CD[i]]*post_covid[i] +
               b_racegap*race_gap[i] + b_ellgap*ell_gap[i] +
               b_school[ENTITY_CD[i]] + b_year[YEAR[i]]  +
               b_assess[ASSESSMENT_NAME[i]] +
               r_femalepostcovid[ENTITY_CD[i]] * Female[i] * post_covid[i]
  }

  # Priors for coefficients
  b_0 ~ dnorm(0, 0.001)
  b_racegap ~ dnorm(0, 0.001)
  b_ellgap ~ dnorm(0, 0.001)

  tau.y <- 1 / (sigma.y^2)
  sigma.y ~ dunif(0, 100)


  # Year effects
  for (i in 1:unique_years) {
    b_year[i] ~ dnorm(0, 0.001)

  }

  # Assessment effects
  for (i in 1:unique_assessments) {
    b_assess[i] ~ dnorm(0, 0.001)
  }

  for (n in 1:unique_schools) {
    b_school[n] <- B[n,1]
    r_female[n] <- B[n,2]
    r_postcovid[n] <- B[n,3]
    r_femalepostcovid[n] <- B[n,4]

    B[n,1:4] ~ dmnorm(Mu.B[n,], Tau.B[,])
    Mu.B[n,1] <- mu.school
    Mu.B[n,2] <- mu.female
    Mu.B[n,3] <- mu.postcovid
    Mu.B[n,4] <- mu.femalepostcovid
  }

  mu.school ~ dnorm(0, 0.001)
  mu.female ~ dnorm(0, 0.001)
  mu.postcovid ~ dnorm(0, 0.001)
  mu.femalepostcovid ~ dnorm(0, 0.001)


  Tau.B[1:4,1:4] ~ dwish(V[,], df)
  Sigma.B <- inverse(Tau.B)

  sigma.school <- sqrt(Sigma.B[1,1])
  sigma.female <- sqrt(Sigma.B[2,2])
  sigma.postcovid <- sqrt(Sigma.B[3,3])
  sigma.femalepostcovid <- sqrt(Sigma.B[4,4])

}"

write(model_H,file="educov_H.jags")
```

Appendix 5

Results from Bayesian Hierarchical Model (without Science Assessments)

Displaying Mean, Standard Deviation, and Quantiles (2.5%, 25%, 50%, 75% and 97.5%)

	Mean	SD	2.5%	25%	50%	75%	97.5%
b_0	49.67091	12.65884	25.0175	41.9850	51.13351	58.1598	69.85058
b_assess[1]	101.88807	3.85268	95.3038	99.1834	101.90211	104.1663	110.94117
b_assess[2]	101.88733	3.85080	95.3180	99.1653	101.91668	104.1708	110.96257
b_assess[3]	102.83363	3.85353	96.2339	100.1316	102.86802	105.1120	111.94084
b_assess[4]	102.86614	3.84987	96.2668	100.1551	102.89254	105.1522	111.91096
b_assess[5]	102.73826	3.85409	96.1524	100.0303	102.78508	105.0014	111.83547
b_assess[6]	102.16296	3.85028	95.5744	99.4468	102.19888	104.4418	111.30885
b_assess[7]	100.52106	3.85748	93.9240	97.8023	100.54608	102.7944	109.58293
b_assess[8]	100.19322	3.85321	93.6007	97.4953	100.22232	102.4729	109.31359
b_assess[9]	99.54482	3.85230	92.9486	96.8217	99.56461	101.8182	108.65058
b_assess[10]	101.02428	3.85237	94.4269	98.3075	101.04764	103.3026	110.13297
b_assess[11]	100.34075	3.85269	93.7479	97.6484	100.36206	102.6128	109.37883
b_assess[12]	100.22709	3.85506	93.6335	97.5115	100.26381	102.5106	109.30887
b_ellgap	0.06463	31.25895	-61.8839	-21.0578	0.21130	21.4375	60.38590
b_racegap	-0.07288	0.43736	-0.9044	-0.3731	-0.07405	0.2213	0.79113
b_school[1]	127.63500	32.17029	57.0007	106.8914	141.02891	148.2437	175.45283
b_school[2]	112.55688	32.17148	41.9342	91.9871	125.91167	133.2457	160.43678
b_school[3]	117.70068	32.17199	47.1515	97.0434	131.06436	138.3647	165.58904
b_school[4]	121.98233	32.17050	51.4565	101.3057	135.36557	142.6148	169.78583
b_school[5]	127.18603	32.18157	56.6940	106.3824	140.58473	147.8144	174.98356
b_school[6]	136.24820	32.17771	65.6597	115.5617	149.61948	156.9051	184.05180
b_school[7]	131.34155	32.19373	60.6887	110.7033	144.69152	152.0295	179.05614
b_school[8]	126.44277	32.17147	55.8326	105.7008	139.82423	147.0539	174.31931
b_school[9]	125.50681	32.17583	54.9087	104.7706	138.88124	146.2260	173.29479
b_school[10]	124.08184	32.17598	53.5056	103.4582	137.47034	144.7267	171.87961
b_school[11]	126.35186	32.18122	55.7984	105.7051	139.74025	147.0047	174.19814
b_school[12]	123.50645	32.17204	52.9410	102.7688	136.86825	144.1894	171.31243
b_school[13]	123.17907	32.17580	52.6371	102.5648	136.55122	143.8518	171.02447
b_school[14]	120.57726	32.18601	49.9805	99.9091	133.90690	141.2892	168.40505
b_school[15]	122.02603	32.17726	51.5043	101.3433	135.38952	142.7129	169.81533
b_year[1]	320.17404	43.74195	255.5180	292.9225	307.67519	346.4518	416.73426
b_year[2]	320.22691	43.74345	255.5641	292.9542	307.71682	346.4874	416.76829
b_year[3]	323.61743	58.67697	230.2051	276.7666	320.24147	377.8710	426.26726
b_year[4]	322.47312	58.67770	229.0514	275.6905	319.11933	376.7236	425.13203
mu.female	3.56326	0.26844	3.0318	3.3861	3.56523	3.7422	4.09287
mu.femalepostcovid	-1.40783	0.35697	-2.1156	-1.6426	-1.40755	-1.1710	-0.70032
mu.postcovid	-3.17449	20.46503	-31.7091	-19.8251	-9.05786	17.9396	32.13436
mu.school	124.18590	32.15093	53.4370	103.6492	137.08462	144.7936	172.16556
r_female[1]	3.49408	0.47634	2.5656	3.1753	3.49090	3.8090	4.43220
r_female[2]	4.08248	0.62171	2.8750	3.6629	4.07981	4.4966	5.30906
r_female[3]	3.74409	0.49282	2.7843	3.4091	3.74085	4.0733	4.70860
r_female[4]	3.62732	0.42005	2.8031	3.3462	3.62492	3.9017	4.47165
r_female[5]	3.44273	0.42013	2.6053	3.1663	3.44289	3.7201	4.27349
r_female[6]	3.02837	0.63205	1.7842	2.6047	3.02689	3.4455	4.28152
r_female[7]	3.25349	0.64506	1.9584	2.8290	3.25737	3.6901	4.50372
r_female[8]	3.53214	0.45437	2.6344	3.2328	3.53442	3.8398	4.42632
r_female[9]	3.73219	0.42830	2.8872	3.4447	3.72682	4.0158	4.58408
r_female[10]	3.39618	0.41506	2.5628	3.1248	3.40245	3.6755	4.18783
r_female[11]	3.69894	0.41887	2.8907	3.4149	3.69488	3.9678	4.54914
r_female[12]	3.57549	0.40382	2.7791	3.3085	3.57766	3.8398	4.37624
r_female[13]	3.61605	0.41013	2.8101	3.3442	3.61724	3.8882	4.42212
r_female[14]	3.45296	0.47227	2.5067	3.1407	3.46195	3.7713	4.36649
r_female[15]	3.62834	0.42956	2.7981	3.3411	3.62375	3.9100	4.48786

r_femalepostcovid[1]	-1.12438	0.59947	-2.3324	-1.5146	-1.12238	-0.7301	0.05534
r_femalepostcovid[2]	-1.53375	0.84316	-3.1682	-2.1092	-1.53241	-0.9726	0.13612
r_femalepostcovid[3]	-1.68155	0.63574	-2.9330	-2.1073	-1.67748	-1.2563	-0.43653
r_femalepostcovid[4]	-1.45020	0.51020	-2.4513	-1.7849	-1.45283	-1.1152	-0.44444
r_femalepostcovid[5]	-1.35877	0.50383	-2.3428	-1.6930	-1.35926	-1.0303	-0.35414
r_femalepostcovid[6]	-1.04311	0.84255	-2.7475	-1.6002	-1.02975	-0.4808	0.59686
r_femalepostcovid[7]	-1.78525	0.85924	-3.4440	-2.3653	-1.79859	-1.2263	-0.05813
r_femalepostcovid[8]	-1.09491	0.55787	-2.1883	-1.4729	-1.10463	-0.7217	0.01073
r_femalepostcovid[9]	-1.11117	0.51478	-2.1213	-1.4536	-1.11513	-0.7782	-0.07092
r_femalepostcovid[10]	-1.36701	0.50667	-2.3604	-1.7025	-1.37142	-1.0367	-0.36163
r_femalepostcovid[11]	-1.32839	0.50602	-2.3173	-1.6645	-1.33127	-0.9946	-0.32478
r_femalepostcovid[12]	-1.50287	0.49313	-2.4800	-1.8293	-1.50434	-1.1797	-0.53253
r_femalepostcovid[13]	-1.39668	0.49308	-2.3712	-1.7248	-1.40017	-1.0661	-0.43594
r_femalepostcovid[14]	-1.83112	0.59490	-3.0095	-2.2278	-1.82713	-1.4370	-0.66444
r_femalepostcovid[15]	-1.45937	0.54134	-2.5275	-1.8158	-1.46273	-1.1014	-0.39250
r_postcovid[1]	-1.77089	20.48533	-30.3588	-18.5378	-7.66191	19.2194	33.64823
r_postcovid[2]	-2.72179	20.46530	-31.1813	-19.3788	-8.60462	18.4544	32.60062
r_postcovid[3]	-3.66333	20.46765	-32.1416	-20.3390	-9.61069	17.3496	31.69475
r_postcovid[4]	-3.04880	20.47944	-31.6281	-19.7060	-8.95619	18.1231	32.42193
r_postcovid[5]	-3.24045	20.48032	-31.8046	-19.9435	-9.11752	17.8284	32.15192
r_postcovid[6]	-2.31479	20.49070	-30.9286	-19.1083	-8.12004	18.8546	33.06772
r_postcovid[7]	-6.23191	20.48704	-34.8849	-22.9789	-12.00375	14.4671	29.20990
r_postcovid[8]	-1.99786	20.48573	-30.5853	-18.7014	-7.92239	19.0466	33.42573
r_postcovid[9]	-2.62239	20.48028	-31.1857	-19.3023	-8.53705	18.4107	32.71731
r_postcovid[10]	-2.63024	20.47397	-31.1980	-19.3379	-8.50715	18.5668	32.83333
r_postcovid[11]	-3.51091	20.47881	-32.0637	-20.2412	-9.37608	17.5250	31.86062
r_postcovid[12]	-3.27870	20.47141	-31.8409	-19.9302	-9.13530	17.6587	32.01166
r_postcovid[13]	-3.11920	20.47207	-31.6484	-19.8378	-8.99974	17.9545	32.23506
r_postcovid[14]	-4.32909	20.47116	-32.8816	-21.0089	-10.15494	16.6835	31.07526
r_postcovid[15]	-3.23334	20.47537	-31.7915	-19.8905	-9.11025	17.8178	32.15538
sigma.femalepostcovid	0.62267	0.20456	0.3280	0.4747	0.58840	0.7328	1.12193
sigma.postcovid	1.21325	0.31477	0.7082	0.9907	1.17442	1.3934	1.93143
sigma.school	5.46187	1.05866	3.8535	4.7224	5.30762	6.0307	8.00211
sigma.y	3.04427	0.05836	2.9320	3.0054	3.04345	3.0828	3.16053