Pandemic Impact on School Performance

(In Coordinated Session: Impact of the Pandemic from Multiple Analytic Perspectives)

Dong-In Kim, Marc Julian, Keith Boughton, and Aurore Phenow

Data Recognition Corporation

Paper presented at the annual meeting of
the National Council on Measurement in Education,

San Diego

April 9, 2022

Pandemic Impact on School Performance

Although it was well-established that the Covid-19 pandemic was a widespread public health crisis, minority groups and vulnerable populations were found to be disproportionately impacted. Factors such as a person's employment status, housing situation, and socioeconomic status contributed to an increased risk of contracting and dying from the virus (Center for Disease Control [CDC], 2021). Moreover, this broader social context may have influenced local policies related to school closures and implementation of different modes of instruction.

Pandemic-related policies were typically developed by districts and translated to all schools for implementation. Understanding the degree to which the pandemic impacted school level performance would provide additional perspective for researchers looking to help district and school officials move forward. Identification of schools most impacted could provide valuable information to support pandemic recovery.

Educational data are often hierarchical or multilevel (i.e., students nested within schools). Bryk & Raudenbush (1992) explained that failure to consider their hierarchical nature can lead to unreliable estimation of the effectiveness of school policies and practices. For large scale assessments, it is reasonable to believe that those attending the same school will have scores that are more highly correlated with one another than they are with scores from students attending other schools. Many common factors, such as same teachers, curriculum, administration policies, would lead to high within-school correlation (Finch, Bolin, & Kelly, 2014).

Three regression models were applied in this study. A hierarchical linear design has an inherent advantage in that school effects can be incorporated within its regression model. However, missing longitudinal data resulting from student migration in and out of schools presents a challenge for the implementation of a hierarchical approach. For this study, the hierarchical design requires student scores from both 2019 and 2021 test administration and associated school level information, and some students were dropped from the hierarchical analyses because of missing 2019 school information. As an alternative, conventional multiple regression, which does not require 2019 school information, was included. Also, a multiple regression that included the school mean as one independent variable was included.

The main purpose of this study is to evaluate the pandemic impact at the school level using three different regression approaches. The following three questions that will be addressed in this research are as follows:

1) Is the impact of school on student performance stable across administrations?
2) Which regression method produces the best fitting model?
3) How to identify schools most impacted by the pandemic?

Data

Large scale assessment datasets in ELA and Mathematics grades 5-8 were used in this study. The data included two groups of test-takers for each ELA and Mathematics grade across multiple

years. The reference group included test-takers from the Spring 2017 and 2019 administrations, which were not impacted by the pandemic. The study group was test-takers from the Spring 2019 and 2021 administrations. Note that the Spring 2017, 2019, and 2021 scale scores for ELA and Mathematics are expressed on the same scale of measurement. Student performance was obtained at either the school- or student-level and matched between the two pairs of administrations: Spring 2021 to Spring 2019, and Spring 2019 to Spring 2017. Test-takers were included in the study if they met the following criteria: (a) earned a valid score on the assessment, and (b) had complete demographic characteristics and prior ability scores. The same inclusion criteria were applied to both the reference and study groups. Table 1 shows the list of covariates included in the analysis.

Table 1: Background variable (covariate) description

Demographic Variable	Covariate	Values	Reference Group
Race/Ethnicity	White		White
	Black	0/1	
	Hispanic	0/1	
	Asian/Pacific Islander	0/1	
	Other (Two or more)	0/1	
Gender	Male	0/1	Female
District Location	City		City
	Suburb	0/1	
	Town	0/1	
	Rural	0/1	
Most Often Used Accommodations or Designated Supports	Text-to-Speech	0/1	Students who used this designated support
	Separate Setting	0/1	Students who used this designated support
English Language Proficiency	Yes (LEP)	0/1	Fully English proficient students
Disability Status	Yes (Disability)	0/1	Students without disabilities
Socioeconomic Status	Yes (SES)	0/1	Students not socioeconomically disadvantaged
Scale scores	Individual and/or school mean for ELA and Math	Scale scores	

Method

Student Interdependence within Schools

To quantify the magnitude of the school effects on student performance across administrations, intraclass correlations for each 2019 and 2021 were estimated with a hierarchical linear regression model by including only school identification to predict student performance. For 2019 the ICC is obtained as follows,

$$Y(\text{SS of 2019}) = X\ (\text{school identification of 2019}), \quad (1)$$

where SS represents scale score. Then, intraclass correlation is expressed as,

$$\frac{\text{variance between schools}}{\text{variance between schools+varicance within schools}}. \quad (2)$$

Higher values of the intraclass correlation indicate that students within the same school are more similar to each other in terms of performance than they are with students in other schools. The ICC was also estimated using scale scores and school identification from 2021 using Equation 2.

Model Fit

This study includes a series of regression analyses designed to potentially estimate the impact of the pandemic at the school level. Therefore, it is important to examine how well the linear regression model fits to the observed data. The statistics that were used to evaluate the regression model were(a) variance explained, R^2, (b) Root mean square residual, (c) Akaike's Information Criteria (AIC), and (d) Bayesian information criteria (BIC). These four criteria were applied to compare the three regression models in this study.

R-squared, R^2

R-squared is the proportion of variation in the outcome that is explained by the predictor variables. In multiple regression models, R^2 is the squared correlation between the observed outcome values and the predicted values by the model. The higher the R-squared, the better the model. In hierarchical linear models, however, there are no accepted standards for measures of multiple R^2 (i.e., total variance accounted for in the outcome), although several have been proposed (See LaHuis, Hartman, Hakoyama, & Clark, 2014; Roberts, Monaco, Stovall, & Foster, 2011). In this study, An R^2 statistic, R2c, of Nakagawa and Schielzeth (2013) was applied. R2c is the conditional R squared value associated with fixed effects and random effects.

Residuals from the model, RMSE

Root mean squared error (RMSE) was included to examine the residuals from the model:

$$\text{RMSE}=\sqrt{(observed\ scale\ score\ of\ 2019 - predicted\ 2019\ score)^2} \quad (3)$$

The lower the RMSE, the better the model.

AIC & BIC

Akaike's Information Criteria (AIC) is a commonly used measure of model/data fit used to compare different models for a common set of data. The AIC penalizes the inclusion of additional variables to a model, prioritizing parsimony rather than overfitting the data with more complex models. The Bayesian information criteria (BIC) is a variant of AIC with a stronger penalty for including additional variables to the model. For both statistics, the lower the statistics, the better the model.

Identifying Pandemic Impact with Three Regression Models

The first step to estimate the pandemic impact was to perform a regression from Spring 2017 scale onto the Spring 2019 scale:

$$Y(\text{SS of 2019}) = X\ (\text{SS of 2017} + \text{bio-demographic data of 2019}) \quad (4)$$

It is important to note that this first regression analysis does not reflect any pandemic impact.

Coefficients from this initial regression based on the prediction of 2019 from 2017 were then applied to the study group, which were Spring 2021 to Spring 2019. The residual for a spring 2021 student is represented by:

$$\text{Residual} = \text{2021 SS} - E(2021|2019) \text{ with Equation 1 coefficients,} \quad (5)$$

where E(2021|2019) represents a score of Spring 2021 based on Spring 2019 scale score and 2021 covariates.

The residual computed in Equation (5) was considered the pandemic impact for individual students in this study. If there were no pandemic impact, residuals will be 0 or close to 0 when summarized at state level. Given this relationship articulated in the regression, negative residual values reflect a learning impact for Spring 2021 students during the pandemic, while positive residuals indicate Spring 2021 students had positive academic growth. To calculate the pandemic impact of a school, residual values for all students in the school were averaged.

Three versions of this regression approach were implemented within this study:

- Model 1 – simple multiple regression as expressed in equations 4 and 5.
  Equations 4 and 5 for Model 1 do not include any school information. School information is only required when residual values of Equation 5 are averaged at each school.

  Model 2 – incorporates the school mean of 2019 scores as an additional independent variable in equation 4 and is expressed as follows:

  $$Y(\text{SS of 2019}) = X\ (\text{SS of 2017} + \text{bio-demo of 2019} + \text{School mean of 2019 SS}) \quad (6)$$

  Therefore, in this Model 2, E(2021|2019) in Equation 5 includes one 2021 school coefficient, which is the same for all schools.

- Model 3 - hierarchical linear regression with school as a nested design

  Model 3 includes a nested design, 2019 students nested within their 2019 schools (nested 2019 school), in Equation 4:

  Y(SS of 2019) = X (SS of 2017 + bio-demo of 2019 + nested 2019 school) (7)

  Note that E(2021|2019) in Equation 5 includes one school coefficient for each school.

All three models require spring 2021 school identification to summarize 2021 individual residuals at school level. The first model does not require spring 2019 school information, so reference dataset of the first model included more students when compared to the other two models.

Software

Multiple regression was performed using the lm R package and hierarchical linear regression was conducted using the lme4 R package in R (Bates, Mächler, Bolker & Walker, 2015).

Flagging Criteria for Impacted Schools

Two statistical criteria were applied to flag schools with pandemic impact: 1) effect size (ES) with 0.5 and 0.8 and 2) differences of three and four standard deviations (SD) from state mean. The first flagging criterion does not reflect the residual values of state residual means, while the second flagging criterion considers that information.

Effect Size Flagging

The residuals in Equation 5 consists of two parts, the first term uses the 2021 scale score, while the second term using E(2021|2019) with Equation 1 coefficients. Effect size at school *i* was calculated as,

$$\text{Effect size at school } i = \frac{Mean_{School_i}(2021\ scale\ score) - Mean_{School_i}(\text{E}(2021|2019)\text{with Equation 1coefficients})}{\sqrt{SD_{School_i}(F2021\ scale\ scor)^2 - SD_{School_i}(\text{E}(2021|2019)\text{with Equation 1 coefficients})^2}}. \quad (8)$$

This study used Sawilowsky's 2009 rules of thumb for effect sizes, which list a medium effect size as 0.5 and a large effect size as 0.8. Note that a negative effect size indicates a pandemic impact and thus instead of 0.5 and 08, -0.5 and -0.8 were used to flag schools in this regard.

Significant Difference from the State Mean

The regression analyses include a statistical test of the null hypothesis (H_0) that the mean residuals of a test administration group, such as school, constitute a random sample from the state distribution of residuals. The hypothesis is tested against the left-sided alternative (H_1), that the mean school residual value is too low to be explained by random sampling. Schools for which H_0 is rejected are flagged. The central limit theorem in statistics indicates that the sampling distribution of mean residuals for class *i* (m_i) is asymptotically normal with the mean and standard deviation expressed as follows;

$$\text{Mean}(m_i) = \mu \qquad (9)$$

$$\text{Standard deviation } (m_i) = \frac{\sigma}{\sqrt{n_i}}, \qquad (10)$$

where n_i denotes the size of the school and m_i denotes the mean residual value for a school *i*, respectively. In addition, the population mean and standard deviation constitute the distribution of the residuals for all individual students. The denominator in the formula for the state standard deviation (Equation 8) indicates that the flagging criterion for each school is adjusted for the number of test takers in a school with valid test scores.

Schools were flagged if their m_i was smaller than

$$\mu \text{ - 3 (or 4) } \times \frac{\sigma}{\sqrt{n_i}} \qquad (11)$$

This flagging criterion of four standard deviations above the state mean provides a statistically conservative test. The standard normal table shows that under random sampling the asymptotic probability of observing a sample mean more than four standard deviations above the population mean is around 0.0001, or one in ten thousand. Even with this conservative test, rejection of H_0 tells us only that the observed residuals in a school are unlikely to be the result of random sampling, and nothing beyond that with any type of certainty.

Results

Table 2 presents intraclass correlations for each Spring 2019 and Spring 2021 administration. The intraclass correlation for Spring 2019 administration ranged from 0.20 to 0.26, while the Spring 2021 administration ranged from 0.17 to 0.28. Similar intraclass correlations for Spring 2021 have been reported in other large-scale assessments. The largest difference of intraclass correlation was 0.05, which was found in English grade 6 between Spring 2019 and Spring 2021. For the matching Spring 2019 and 2021 grades, ICCs for Math were larger than those for ELA. This indicates that students' performances for Mathematics were more impacted by the

belonging of certain schools than those for ELA. As can be seen in Table 2, there were more Spring 2021 schools compared to Spring 2019, although there was a smaller number of students in spring 2021. This was because many new schools were opened during the pandemic in order to accommodate remote learning.

Table 2: Intra Class Correlation for school

		2019 Administration			2021 Administration		
Content	Grade	N_Total	N_School	ICC	N_Total	N_School	ICC
EL	5	64531	1193	0.23	53793	1218	0.21
	6	65212	853	0.23	55198	873	0.18
	7	63683	810	0.22	56013	840	0.17
	8	62817	812	0.20	56411	840	0.17
MA	5	64631	1194	0.24	53711	1216	0.28
	6	65355	853	0.26	55181	871	0.26
	7	63830	811	0.24	56053	841	0.23
	8	62963	813	0.25	56455	840	0.24

Table 3 represents the R-squared and RMSE values across grades and models for the different regression models used to predict 2019 performance from 2017. Across content and grade, R-squared values for Model 1 ranged from 0.60 to 0.63, those for Model 2 ranged from 0.61 to 0.64, those for Model 3 ranged from 0.62 to 0.66. Therefore, Model 3, hierarchical model with nested design, produced the largest R-squared values and the multiple regression model without school effect produced the smallest R-squared values. RMSE ranged from 29.25 to 38.10 for Model 1, 28.52 to 37.42 for Model 2, and 27.63 to 36.59 for Model 3. RMSE values for Model 3 were smaller than those for Model 1, which were the largest.

Table 3: R-squared and RMSE across Models and Grades for Regression from 2017 to 2019

			R-Squared			RMSE		
Content	Grade	Year	M1	M2	M3	M1	M2	M3
EL	5	2019	0.63	0.65	0.66	29.25	28.52	27.63
EL	6	2019	0.63	0.65	0.66	30.22	29.47	28.65
EL	7	2019	0.63	0.64	0.65	33.19	32.57	31.70
EL	8	2019	0.62	0.64	0.65	36.55	35.81	34.75
MA	5	2019	0.60	0.63	0.64	33.27	32.02	30.91
MA	6	2019	0.61	0.63	0.65	36.12	34.97	33.88
MA	7	2019	0.60	0.61	0.62	38.10	37.42	36.59
MA	8	2019	0.60	0.62	0.63	36.17	35.33	34.36

* Model 1 (M1): multiple regression without school mean
* Model 2 (M2): multiple regression with spring 2019 school mean
* Model 3 (M3): hierarchical linear regression with spring 2019 school as nested design

Table 4: represents AIC and BIC across models for regression from 2017 to 2019. For AIC, Model 3 value was the smallest, and Model 1 value was the largest. The same pattern was found for BIC.

Table 4: AIC and BIC Across Models for Regression from 2017 to 2019

		AIC			BIC		
Content	Grade	Model 1	Model 2	Model 3	Model 1	Model 2	Model 3
EL	5	580057	576986	576171	580201	577139	576324
	6	587568	584509	583367	587712	584662	583520
	7	584544	582299	581153	584688	582452	581306
	8	592536	590128	588676	592680	590280	588829
MA	5	595174	590560	589662	595318	590713	589815
	6	608826	604918	603483	608970	605071	603637
	7	600641	598523	597782	600785	598676	597935
	8	590602	587834	586655	590746	587986	586807

Tables 5 and 6, present summary statistics for 2021 scale scores, 2021 predicted scores, and residuals using Equation 5 for both ELA and Mathematics. 2021 predicted scores are scores with 2017 to 2019 coefficients. Residuals are values subtracting spring 2021 scale scores from 2021 predicted scores. Therefore, negative values mean there was a learning impact, while positive values indicate positive learning growth. Except for ELA grades 7 and 8, all residuals were negative values. Model 3 produced the smallest residuals (i.e., largest negative values for most cases), while Model 2 produced the largest residuals (i.e., smallest negative values for most cases). Absolute values of Mathematics Residuals were larger than those of corresponding ELA tests.

Figure 1 shows the distributions of residuals for individual students and schools for Mathematics Grade 6, which produced largest negative residuals. The x-axis is the residual for each individual student or school, while the y-axis is the frequency of the residual values. Both residuals for individual students and schools showed a slightly skewed normal distribution for all three models. Compared to Models 1 and 3, Model 2 produced more school residuals around the residual mean.

Table 7 presents percent of schools flagged with 3 and 4 standard deviations from state mean. Model 3 requires the same schools appear in Spring 2017, 2019, and 2021 administrations. Therefore, there were some missing schools for Model 3 compared to Models 1 and 2., Only the common schools among these 3 models were used to compare the models (i.e., Model 3 schools) for flagging. When 3 standard deviations were applied, the percent of schools flagged ranged from 7.42 to 14.61 for Model 1, from 1.61 to 6.58 for Model 2, and from 8.67 to 13.23 for Model 3 across contents and grades. When 4 standard deviations were applied, the percent of schools flagged ranged from 2.68 to 8.95 for Model 1, from 0.45 to 3.68 for Model 2, and from 3.57 to 7.89 for Model 3 across contents and grades. Among the three models, Model 2 flagged the smallest percent of schools, and Models 1 and 3 flagged similar percent of schools. When the

Table 5: Summary Statistics for ELA 2021 Scale Scores, 2021 Predicted Scores, and Residuals

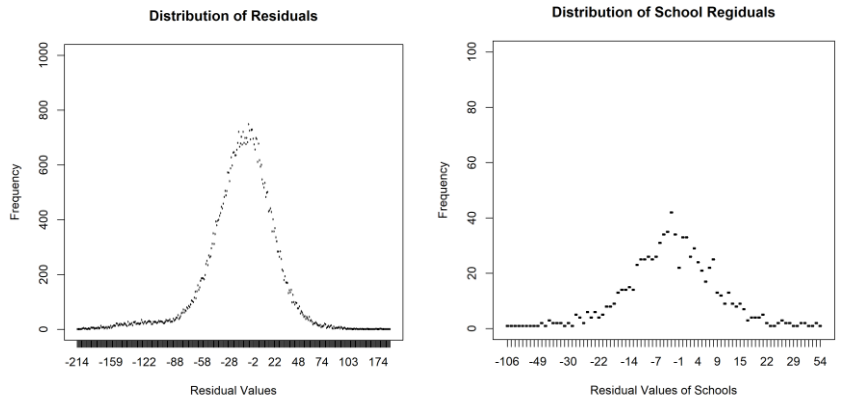
Grade	Method	N	Score	Mean	SD	Min	Max
5	1	50735	2021.SS	593.52	48.75	350	940
			Predicted	595.84	36.22	409	841.63
			Residual	-2.32	30.51	-235.86	298.66
	2	50735	2021.SS	593.52	48.75	350	940
			Predicted	594.25	36.74	408.63	831.78
			Residual	-0.72	29.95	-231.09	306.39
	3	49023	2021.SS	593.45	48.75	350	940
			Predicted	595.92	37.26	394.93	846.92
			Residual	-2.47	30.74	-228.52	295.5
6	1	52167	2021.SS	604.43	49.79	360	950
			Predicted	608.47	37.19	415.33	844.34
			Residual	-4.04	31.39	-274.56	299.41
	2	52167	2021.SS	604.43	49.79	360	950
			Predicted	606.6	37.41	414.18	828.6
			Residual	-2.17	30.72	-272.42	289.65
	3	49939	2021.SS	604.25	49.7	360	950
			Predicted	608.59	37.9	414.37	841.43
			Residual	-4.34	31.62	-268.71	297.1
7	1	52880	2021.SS	625.7	54.97	370	960
			Predicted	625.54	40.15	411.67	887.94
			Residual	0.16	33.75	-294.68	290.54
	2	52880	2021.SS	625.7	54.97	370	960
			Predicted	624.23	40.36	422.19	877.55
			Residual	1.47	33.42	-294.06	290.87
	3	52293	2021.SS	625.78	54.92	370	960
			Predicted	625.93	41.08	406.06	895.09
			Residual	-0.15	33.88	-285.51	293.8
8	1	53542	2021.SS	628.63	58.44	380	970
			Predicted	626.36	45.59	385.85	922.69
			Residual	2.27	36.65	-319.23	295.92
	2	53542	2021.SS	628.63	58.44	380	970
			Predicted	625.72	45.68	370.66	917.97
			Residual	2.91	36.25	-310.01	307.4
	3	52681	2021.SS	628.85	58.35	380	970
			Predicted	626.84	47.02	347.65	925.92
			Residual	2.01	37.3	-317.78	308.57

Table 6: Summary Statistics for Mathematics 2021 Scale Scores, 2021 Predicted Scores, and Residuals

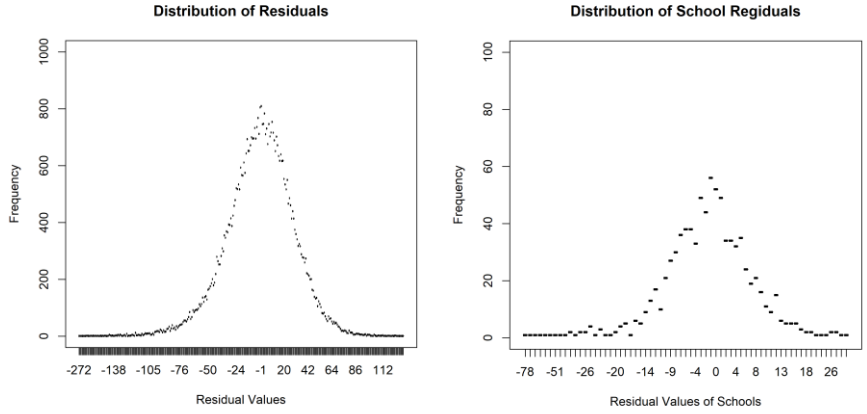
Grade	Method	N	Score	Mean	SD	Min	Max
5	1	50616	2021.SS	594.93	55.71	430	830
			Predicted	606.37	41.81	425.67	761.99
			Residual	-11.43	36.41	-224.87	194.02
	2	50616	2021.SS	594.93	55.71	430	830
			Predicted	601.83	43.55	394.52	778.4
			Residual	-6.9	34.51	-216.05	194.6
	3	48904	2021.SS	595.03	55.74	430	830
			Predicted	606.64	43.47	408.11	785.96
			Residual	-11.61	36.11	-213.33	189.34
6	1	52094	2021.SS	602.56	57.39	440	870
			Predicted	617.49	40.77	450.46	786.47
			Residual	-14.93	36.21	-213.55	226.08
	2	52094	2021.SS	602.56	57.39	440	870
			Predicted	612.44	41.92	409.02	783.82
			Residual	-9.88	34.73	-209.41	225.62
	3	49804	2021.SS	602.24	57.28	440	870
			Predicted	617.55	42.32	415.86	800.86
			Residual	-15.31	35.89	-218.31	214.55
7	1	52791	2021.SS	620.54	59.68	450	880
			Predicted	631.56	45.73	441.65	820.39
			Residual	-11.02	39	-221.51	198.66
	2	52791	2021.SS	620.54	59.68	450	880
			Predicted	628.65	46	426.17	815.47
			Residual	-8.11	38.16	-225.92	203.8
	3	52212	2021.SS	620.68	59.6	450	880
			Predicted	631.85	46.61	425.29	831.92
			Residual	-11.17	38.7	-236.43	197.49
8	1	53456	2021.SS	638.95	56.77	470	890
			Predicted	647.56	44.79	489.35	854.92
			Residual	-8.61	37.24	-222.9	194.37
	2	53456	2021.SS	638.95	56.77	470	890
			Predicted	644.45	45.2	453.07	864.22
			Residual	-5.5	36.2	-224.19	199.95
	3	52604	2021.SS	639.26	56.65	470	890
			Predicted	648.08	45.88	460.74	871.61
			Residual	-8.83	37.18	-228.82	202.19

Figure 1: Distributions of Residuals for Individual Students and Schools for Math Grade 6

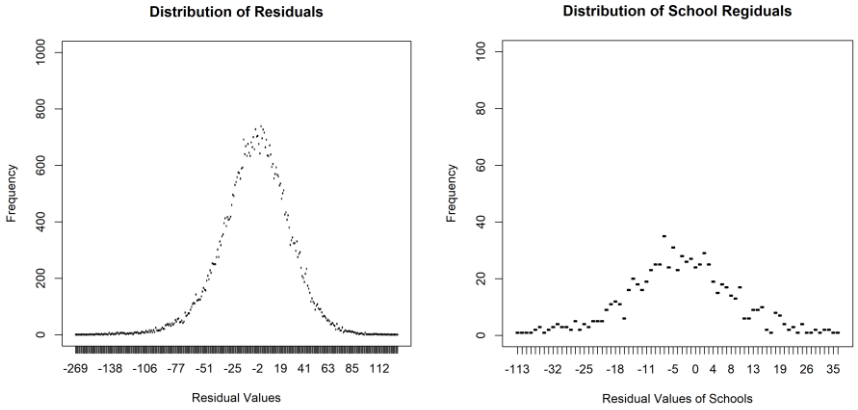
Model 1



Model 2



Model 3



same grades were compared by content area, ELA and Mathematics, there were more flagged schools in Mathematics.

Table 7: Percent of Schools Flagged with 3 or 4 SD from State Mean

		N of Schools		SD > 3			SD > 4		
Content	Grade	M1 & M2	M3	M1	M2	M3	M1	M2	M3
EL	5	1200	1119	7.42	1.61	8.67	2.68	0.45	3.57
	6	858	760	8.55	2.76	11.18	4.34	0.92	5.13
	7	820	746	8.18	3.08	10.19	3.62	1.47	4.56
	8	813	742	7.95	4.72	11.99	4.18	2.16	6.33
MA	5	1200	1119	12.33	4.02	13.23	6.43	1.7	7.51
	6	859	760	14.61	6.58	13.03	8.95	3.68	7.89
	7	821	747	10.44	4.69	9.64	6.02	2.14	5.49
	8	812	741	12.15	6.21	12.82	6.34	2.29	6.61

Table 8 shows the percentage of schools flagged with an effect size of 0.5 and 0.8. Samples sizes of SD flags and effect size flags were different. SD flags reflect the number of students in each school, while effect size does not. For the effect size flag, schools with at least 10 students were included due to the stability of statistics.

When ES < -0.5 was applied, the percent of schools flagged ranged from 1.55 to 31.21 for Model 1, from 0 to 11.39 for Model 2, and from 4.35 to 31.66 for Model 3 across contents and grades. When ES < -0.8 was applied, the percent of schools flagged ranged from 0.31 to 8.14 for Model 1, from 0 to 1.48 for Model 2, and from 0.75 to 9.91 for Model 3 across contents and grades. For SD flag, among three models, Model 2 flagged the smallest percent of schools, and Models 1 and 3 flagged similar percent of schools. Again, when the same grades were compared for ELA and Mathematics, there were more flagged schools in Mathematics.

Table 8: Percentage of Schools Flagged with Effect Size

		N of Schools		ES < -0.5			ES < -0.8		
Content	Grade	M1 & M2	M3	M1	M2	M3	M1	M2	M3
EL	5	1058	1031	6.01	0.97	8.05	1.16	0.1	1.36
	6	709	677	5.17	1.33	9.31	0.74	0.3	1.77
	7	684	667	1.95	0	4.35	0.45	0	0.75
	8	664	646	1.55	0.46	4.95	0.31	0	1.39
MA	5	1058	1031	24.29	8.88	25.66	7.8	0.78	8.88
	6	709	677	31.21	11.39	31.36	8.14	1.48	9.91
	7	684	667	15.27	5.99	14.82	2.99	0.9	3.44
	8	664	646	13.82	4.97	15.99	2.8	0.78	4.81

Summary and Discussion

To examine the pandemic impact on schools, three regression models were applied to 2017, 2019, and 2021 datasets. Residual values were estimated by subtracting predicted 2021 scores with the regression coefficients of 2017 to 2019 from 2021 scale scores, with the residuals being considered as the pandemic impact. A school residual was calculated by averaging residuals of the students in the schools.

There were three research questions. The first question was about the stability of school effects on students' performances across 2019 and 2021. The school effects across 2019 and 2021 were examined using ICCs. The largest ICC value was 0.26 for spring 2019, and 0.28 for spring 2021. For both 2019 and 2021, there were some school effects on students' performance. For ELA, ICC values for spring 2021 were lower than those for spring 2019. That is, school effects on students' performance for spring 2021 were lower than those for spring 2019. There was no clear pattern found for Mathematics.

The second question was regarding which regression method produced the best fit model. Four goodness of fit statistics, which were R-squared, RMSE, AIC, and BIC, were applied to the regression of 2017 to 2019. Model 3 produced the largest R-squared and smallest RMSE, AIC, and BIC, and Model 1 produced the smallest R-squared and largest RMSE, AIC, and BIC. Therefore, Model 3 was the best fitting model, and Model 1 was the worst fitting model. It is natural to find that Model 3, which considers each school effect using nest design, showed the best fit. However, the differences among the three models were not large in most cases. Therefore, in practice, when many schools are dropped with Model 3, which requires that schools exist in all spring 2017, 2019, and 2021 data sets, the other two models could be considered.

The last question was how to flag schools impacted by the pandemic. In this study, the two flagging criteria utilized were the SD and effect size flags. Because the SD flag compares school residuals to state residuals, it can be considered as a relative criterion. These effect size flags compared expected performance under a normal environment and 2021 observed performance of a school, and thus can be considered an absolute criterion. When pandemic impacts of the same grade were compared for ELA and Mathematics, it was clear that there were more flagged schools in Mathematics, especially with effect size flag. Because the residuals of Mathematics were much larger than those of ELA, as a kind of absolute criterion, the effect size criterion flagged more schools in Mathematics than schools in ELA. The SD criterion also flagged more schools in Math, but not as many schools compared to effect size criterion. In practice, schools flagged in both criteria also can be considered. In this study, the 3 and 4 SD, as well as the effect sizes of 0.5 and 0.8 were applied. However, different flag values could also have been considered.

It is important to note that the ICC values showed that school effects exist in students' performance. Model 3 considered the school effect as the second level in a nested design. Educational policies, including pandemic policies, are typically developed by districts and then applied to all schools within the district. Therefore, a future study needs to examine a nested design of the district effect, as well as the school effect.

Note that this study did not examine the students who did not take spring 2021 tests. Therefore, if the presence of these students in a school is not random, then the residuals of the school can lead to an incorrect result in the pandemic effect at the school level.

Lastly, it is important to note that the pandemic impacted all students, parents, teachers, school officials, and district officials. This study simply attempts to capture and understand some of variability associated with the pandemic using regression models at a specific level of analysis (i.e., the school).

References

Bosker, R.J. (2012). *Multilevel analysis: An introduction to basic and advanced multilevel modeling* (2nd Edition). London: Sage.

Bryk, A. S., & Raudenbush, S. W. (1992). *Hierarchical linear models: Applications and data analysis methods.* Sage Publications, Inc.

Center for Disease Control (April 19, 2021). Health Equity Considerations and Racial and Ethnic Minority Groups. Retrieved from https://www.cdc.gov/coronavirus/2019-ncov/community/health-equity/race-ethnicity.html.

Finch, W.H, Bolin, J.E, Kelley, K (2014). *Multilevel Modeling Using R*. Chapman and Hall/CRC,Boca Raton, Fl.

Hox, J. J. (2010). *Quantitative methodology series. Multilevel analysis: Techniques and applications (2nd ed.)*. New York, NY, US: Routledge/Taylor & Francis

In J.J. Hox & J.K. Roberts (Eds.), *Handbook of Advanced Multilevel Analysis*. New York: Routledge.

LaHuis, D. M., Hartman, M. J., Hakoyama, S., & Clark, P. C. (2014). Explained variance measures for multilevel models. *Organizational Research Methods*, 17(4), 433-451.

Nakagawa, S., & Schielzeth, H. (2013). A general and simple method for obtaining R2 from generalized linear mixed-effects models. *Methods in Ecology and Evolution*, 4(2), 133-142. doi:10.1111/j.2041-210x.2012.00261.x

Roberts, J. K., Monaco, J. P., Stovall, H., & Foster, V. (2011). Explained variance in multilevel models. In J. J. Hox & J. K. Roberts (Eds.), *Handbook for advanced multilevel analysis* (pp. 219–230). Routledge/Taylor & Francis Group.