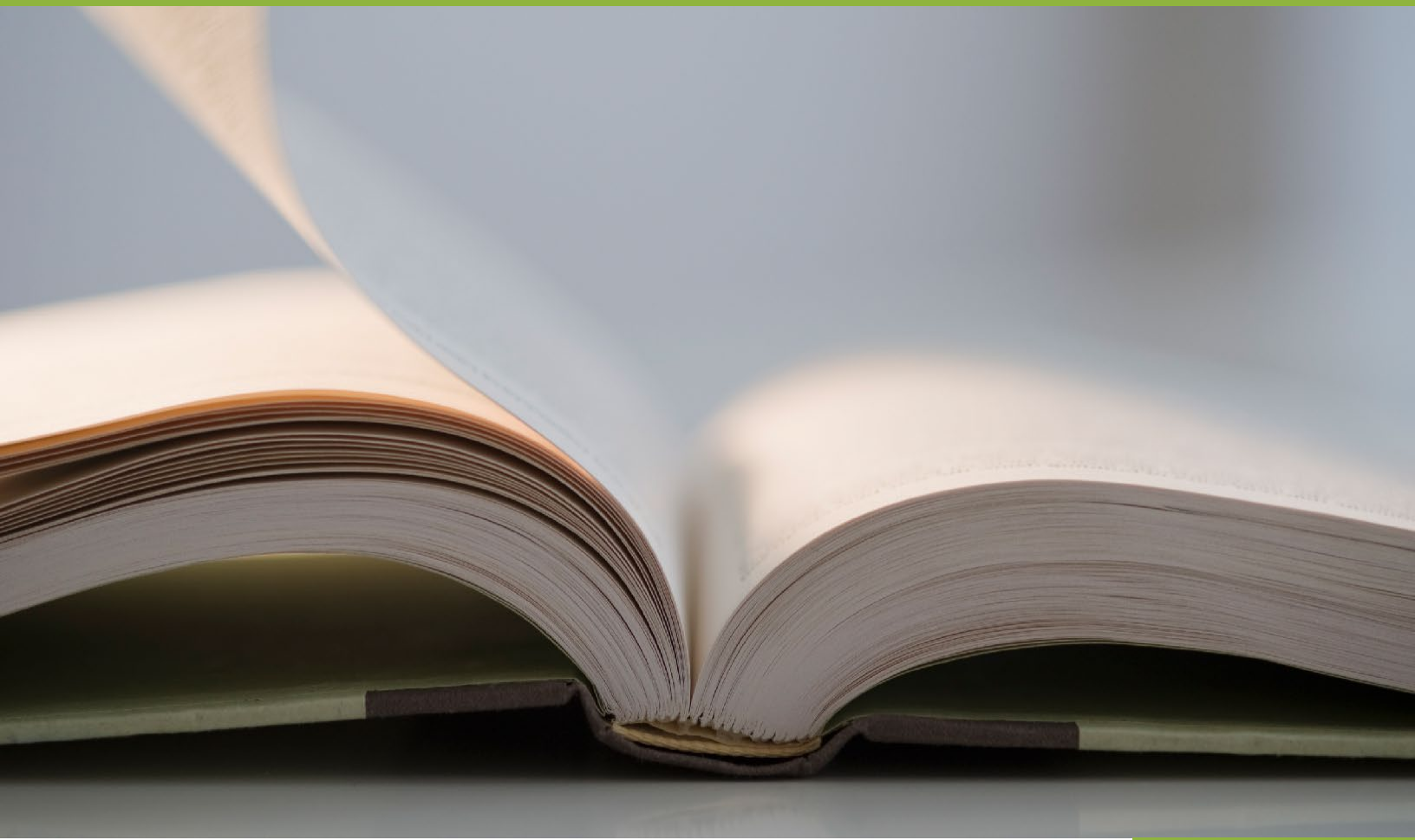
# SAS® EVAAS

## Statewide Year-Over-Year Trends in Achievement: Before, During, and After the Pandemic

**Prepared for the North Carolina Department of Public Instruction**



sas
THE POWER TO KNOW.

## Contents

# 1 Executive Summary

The North Carolina Department of Public Instruction (NCDPI) and SAS Institute Inc. (SAS) collaborated to provide educators, policymakers, and other stakeholders with insight into the path or trajectory of student achievement before, during, and after the pandemic. **The purpose of this analysis is to help understand both the initial and remaining impact of the pandemic on student achievement in North Carolina. The analysis focuses on how state achievement changed from one year to the next from 2013 to 2022.**

When discussing the pandemic's impact and students' recovery, a common question is *what constitutes a full recovery?* In other words, how will we know when students have fully recovered from the negative impacts of the pandemic? While there are several ways to measure students' recovery, this analysis focuses on long- and short-term changes in achievement over time. To provide a comprehensive view of student recovery, this analysis includes two different thresholds for evaluating student recovery. Specifically, student achievement in 2022 is compared to a:

1. Continuation of the pre-pandemic trend from the 2013–2019 time period
2. Three-year average of the state's achievement observed in the 2017–2019 time period.

The first threshold considers where the state was going *and continues that path*, whereas the second threshold only considers where the state *was* during the years immediately prior to the pandemic.

As part of understanding students' recovery in North Carolina, the analysis produces three key metrics:

1. **The *Pre-Pandemic Trend* represents the overall trend in achievement for an assessment between 2013 and 2019.** This line smooths out the year-to-year variation in the observed achievement for the state.
2. **The *Pandemic Impact* represents the extent to which actual achievement in 2021 diverged from the pre-pandemic trend, had it continued to 2021.** This line represents a counterfactual, or an estimate of what achievement might have been if the pre-pandemic trend were not disrupted by the pandemic.
3. **The *Distance to a Full Recovery* represents the extent to which actual achievement in 2022 diverged from a full recovery according to two different thresholds.** There are two ways to consider recovery: a continuation of the pre-pandemic trend threshold based on the 2013-2019 timeframe *and* a three-year average threshold representing the more immediate timeframe of 2017-2019.

Collectively, these three metrics provide a comprehensive picture of the long-term trajectory of year-over-year achievement in North Carolina.

This model uses a more sophisticated approach to determine state achievement than simple averages of scale scores or the percentage of student scoring proficient because it takes into account the year-to-year variation and trends in student achievement that existed prior to the pandemic. The model also adjusts for version changes in the assessments that occurred prior to the pandemic as was the case for EOG Math in 2019.

In addition to providing a more robust estimate of achievement trends, a major advantage of this approach is that it is possible to provide measures of the pandemic impact and recovery for earlier grades because it does not require sufficient predictors by 2019, like the cohort model (which was the basis for the earlier reports provided by NCDPI's Office of Learning and Recovery). Where sufficient data

are available, results are provided for EOG Math and Reading in grades 3–8, EOG Science in grades 5 and 8, and EOC Biology, English II, and Math I.

This year-over-year analysis provides the following findings based on statewide trends:

- Prior to the pandemic, achievement was relatively stable for most assessments with small positive or negative effect sizes.
- The pandemic impact was negative for all assessments with the exception of English II, and the effect size ranged from medium to large depending on the assessment.
- One year later, most assessments indicate a recovery that was not sufficient to meet either recovery threshold, though how much pandemic impact remained depended on the threshold.
- As a content area, Math was more negatively impacted by the pandemic than Reading, and Math has a greater distance to full recovery than Reading one year later.
- EOG Reading 3 is the only assessment that meets the extended trend recovery threshold at the state level.
- There is considerable variation among schools within the state in terms of the pandemic impact and recovery thresholds. The extent of variation among schools in achievement trends observed prior to the pandemic was, comparatively, modest.

The following sections provide an overview of the model, the information it provides, and its output.

# 2 Data

## 2.1 Data Received

The analysis in this report leveraged student-level assessment data, where available, from 2012-13 through the 2021-22 school year in order to compile a longitudinal data set based on the following assessments:

- EOG Mathematics in grades 3–8
- EOG Reading in grades 3–8
- EOG Science in grades 5 and 8
- EOC Biology, English II, Math 1 and Math 3

The state EOG tests are administered in the spring semester whereas the EOC assessments are typically given at the end of the fall and spring semesters with the occasional summer administration. For each administration, SAS used the following student identifiers, assessment data, and student flags:

- Student Identifiers
  - Student Last Name
  - Student First Name
  - Student Middle Initial
  - Student Date of Birth
  - Student Identification Number
- Assessment Information
  - Scale Score
  - Test Taken
  - Tested Grade
  - Test Semester
  - School Number
  - District Number
  - Administration Window

Note that the model adjusted scores for version changes in the assessments that occurred *prior* to the pandemic as was the case for EOG Math in 2019. However, version changes in many of the Reading assessments occurred during or after the pandemic. As a result, the effects of these changes are confounded with the timing of the pandemic and could not be modeled, so there were no adjustments to those scores.

SAS merged the individual student records over time using an algorithm that incorporated all student identifiers to create a longitudinal database that tracks individual students' performance across grade levels on state assessments each year.

## 2.2 Business Rules

In creating the longitudinal database, the following business rules were applied regarding student scores.

### 2.2.1 Missing Grade

In North Carolina, the grade used in the analyses and reporting is the tested grade, not the enrolled grade. If a grade is missing on an early grade or end-of-grade test record, then that record will be excluded from all analyses. The grade is required to include a student's score in the appropriate part of the models.

### 2.2.2 Duplicate (Same) Scores

If a student has a duplicate score for a particular subject and tested grade in a given testing period in a given school, then the extra score will be excluded from the analysis.

### 2.2.3 Students with Missing Districts or Schools for Some Scores but Not Others

If a student has a duplicate score with a missing district or school for a particular subject and grade or course in a given testing period, then the duplicate score that has a district and/or school will be included over the duplicate score that has the missing data.

### 2.2.4 Students with Multiple (Different) Scores in the Same Testing Administration

If a student has multiple scores in the same period for a particular subject and grade or course and the test scores are not the same, then those scores will be excluded from the analysis. If duplicate scores for a particular subject and tested grade in a given testing period are at different schools, then both scores will be excluded from the analysis. Note that if multiple scores are received for grade 3 Reading or Math across years, only the most recent score is used.

### 2.2.5 Students with Multiple Grade Levels in the Same Subject in the Same Year

A student should not have different tested grade levels in the same subject in the same year. If that is the case, then the student's records are checked to see whether the data for two separate students were inadvertently combined. If this is the case, then the student data are adjusted so that each unique student is associated with only the appropriate scores. If the scores appear to all be associated with a single unique student, then scores that appear inconsistent are excluded from the analysis.

### 2.2.6 Students with Records That Have Unexpected Grade Level Changes

If a student skips more than one grade level (e.g., moves from sixth in 2018 to ninth in 2019) or is moved back by one grade or more (i.e., moves from fourth in 2018 to third in 2019) in the same subject, then the student's records are examined to determine whether two separate students were inadvertently combined. If this is the case, then the student data is adjusted so that each unique student is associated with only the appropriate scores. These scores are removed from the analysis if it is the same student. Per NCDPI's decision, the analysis does not remove students with scores that appear to be associated with inconsistent grades. The analysis leaves students in the analysis at the tested grade that EVAAS receives from NCDPI.

### 2.2.7 Students with Records at Multiple Schools in the Same Test Period

If a student is tested at two different schools in a given testing period, then the student's records are examined to determine whether two separate students were inadvertently combined. If this is the case, then the student data is adjusted so that each unique student is associated with only the appropriate scores. When students have valid scores at multiple schools in different subjects, all valid scores are used at the appropriate school.

### 2.2.8 Outliers

Student assessment scores are checked each year to determine whether they are outliers in context with all the other scores in a reference group of scores from the individual student. These reference scores are weighted differently depending on proximity in time to the score in question. Scores are checked for outliers using related subjects as the reference group. For example, when searching for outliers for EOC Math test scores, all EOG and EOC Math subjects are examined simultaneously, and any scores that appear inconsistent, given the other scores for the student, are flagged.

Scores are flagged in a conservative way to avoid excluding any student scores that should not be excluded. Scores can be flagged as either high or low outliers. It should also be noted that test scores within a year, subject and grade are normalized before checking begins. This helps mitigate any unnecessary flagging of outliers due to a year of assessments shifting across the state as might happen in 2021.

This process is part of a data quality procedure to ensure that no scores are used if they were, in fact, errors in the data, and the approach for flagging a student score as an outlier is fairly conservative. Again, students were expected to score lower in 2021 due to the pandemic, and this process is more about flagging data that might be erroneous.

Considerations included in outlier detection are:

- Is the score in the tails of the distribution of scores? Is the score very high or low achieving?
- Is the score “significantly different” from the other scores as indicated by a statistical analysis that compares each score to the other scores?
- Is the score also “practically different” from the other scores? Statistical significance can sometimes be associated with numerical differences that are too small to be meaningful.
- Are there enough scores to make a meaningful decision?

To decide whether student scores are considered outliers, all student scores are first converted into a standardized normal Z-score. Then each individual score is compared to the weighted combination of all the reference scores described above. The difference of these two scores provides a t-value of each comparison. Using this t-value, the models can flag individual scores as outliers.

There are different business rules for the low outliers and the high outliers, and this approach is more conservative when removing a very high-achieving score.

For low-end outliers, the rules are:

- The percentile of the score must be below 50.
- The t-value must be below -3.5 for EOGs in Math and Reading when determining the difference between the score in question and the weighted combination of reference scores (otherwise known as the comparison score). In other words, the score in question must be at least 3.5 standard deviations below the comparison score. For EOC and EOG Science assessments, the t-value must be below -4.0.
- The percentile of the comparison score must be above a certain value. This value depends on the position of the individual score in question but will range from 10 to 90 with the ranges of the individual percentile score.

For high-end outliers, the rules are:

- The percentile of the score must be above 50.

- The t-value must be above 4.5 for EOGs in Math and Reading when determining the difference between the score in question and the reference group of scores. In other words, the score in question must be at least 4.5 standard deviations above the comparison score. For EOC and EOG Science assessments, the t-value must be above 5.0.
- The percentile of the comparison score must be below a certain value. This value depends on the position of the individual score in question but will need to be at least 30 to 50 percentiles below the individual percentile score.
- There must be at least three scores in the comparison score average.

### 2.2.9 Membership

Students were excluded if they did not meet membership, a designation based on student enrollment at a school and used for accountability purposes.

### 2.2.10 First Year English Learner

Given the research purpose of the analysis and need to create a comparable student population over time, students were not excluded based on first year English Learner designation.

# 3 Methods of Analysis

## 3.1 Overview

This analysis assesses the pre-pandemic trend, pandemic impact, and recovery trends through a model that accounts for the year-to-year variation and trends in student achievement that existed prior to the pandemic. In contrast to the cohort model, this approach evaluates trends in results for specific assessments over time rather than using individual students' prior testing data. A major advantage of this approach is that it is possible to provide measures of the pre-pandemic trend, pandemic impact and recovery for earlier grades that are no longer available in the cohort model.

In comparing student achievement in a given content area (such as fifth-grade Math) over the years prior to the pandemic, the average scale score for the state, a district, or a school might change from one year to the next. At the state level, these fluctuations are typically small, although they might be centered around a positive or negative trend line. On the other hand, schools and districts exhibit varying degrees of stability in average scale scores. Statistically, the average scale scores from small schools and districts are more variable over time than large schools and districts. In addition, prior to the pandemic, some schools and districts might exhibit distinct positive trends in average scale scores while other exhibit flat or negative trends. There are reasonable explanations for any variation: the populations of students in a given subject/grade changes from one year to the next; educators associated with the school or district could change from one year to the next; and there might be curricular and instructional strategies in place that influence achievement over time.

Regardless of the source, when assessing the pandemic's impact and recovery trends, it is important to consider this natural variation and the trends in achievement that existed in the state, a district, or a school in the years prior to the pandemic. Doing so provides a more robust estimate of the pandemic's impact and recovery.

The following sections provide a more detailed review of the model and its output.

## 3.2 Determining Average Achievement

The analysis uses the student assessment data and business rules described in Section 2 to build the set of student data. The data follow successive cohorts of students within a specific subject and grade as they pass through the schools over the multiple years, such as 2017 fifth-grade students, 2018 fifth-grade students, 2019 fifth grade-students, etc. For each subject and grade analyzed, the model breaks down variation in student achievement over time into state, school, and student contributions. Specifically, the model estimates the following:

- Statewide pre-pandemic trend in achievement.
- Statewide version effects (when applicable), which is operationalized as an abrupt and persistent change in achievement that occurs the first year a new version of an assessment is administered.
- Statewide pandemic impact, which is operationalized as a deflection of the pre-pandemic trend in 2021.
- Statewide recovery, which is operationalized as a change in achievement from 2021 to 2022.
- Stochastic cohort-to-cohort variation in achievement observed within all schools included in the analysis. This component is only estimated in the years prior to the pandemic.
- School-specific pre-pandemic trend parameters, version effect parameters, pandemic impact parameters, and recovery parameters which are operationalized as random effects.

- A residual, student-level error term that represents the deviation of a student's scale score value from an expectation derived from all of the components described above.

Generally speaking, the analysis uses a piecewise, hierarchical linear growth model that is estimated separately by subject and grade.[1]

In more technical terms, this analysis lets $y_{i,j,t}$ equal the test score (achievement) in a specific subject and grade for student $i$ in school $j$ at time $t$, where $t$ is an index on a set of years that span the pre- and post-pandemic period e.g., $t \in (2017,2018, \ldots 2022)$. The model can be expressed as:

$$y_{i,j,t} = \beta_{0,j} + \beta_{1,j}T + \beta_{2,j}P + \beta_{3,j}R + \gamma_{j,t} + \epsilon_{i,j,t} \quad (1)$$

With respect to the covariates:

$$\text{For } t \leq 2021 \text{ let } T = t - 2020\text{; else } T = 1\text{. Thus, } T \in (-3,-2,-1,0,1,1). \quad (2)$$

$$\text{For } t \leq 2020\text{, let } P = 0\text{; else } P = 1\text{. Thus, } P \in (0,0,0,0,1,1). \quad (3)$$

$$\text{For } t \leq 2021\text{, let } R = 0\text{; else } R = 1\text{. Thus, } R \in (0,0,0,0,0,1) \quad (4)$$

$\beta_{1,j}T$ captures school-specific trends in scale scores observed prior to the pandemic. $\beta_{2,j}P$ represent a deflection from the prior trend associated with the pandemic, and $\beta_{3,j}R$ represents a change in achievement from 2021 to 2022. The coding of the $T$, $P$, and $R$ variables result in an intercept term, $\beta_{0,j}$ that can be interpreted as the "average" scale score of a student attending school $j$ in 2020. Since there was no testing data collected in the spring of the 2019-2020 school year in North Carolina, $\beta_{0,j}$ is an extrapolation from each schools' prior trend ($\beta_{1,j}$).

The school-specific growth parameters are random effects centered around state-wide averages. Specifically,

$$\beta_{0,j} = \tau_{00} + \mu_{0,j}, \quad (5)$$

$$\beta_{1,j} = \tau_{10} + \mu_{1,j}, \quad (6)$$

$$\beta_{2,j} = \tau_{20} + \mu_{2,j}\text{, and} \quad (7)$$

$$\beta_{3,j} = \tau_{30} + \mu_{3,j} \quad (8)$$

A final school-level random effect, $\gamma_{j,t}$, represents idiosyncratic variation of the school means around the school-specific growth trajectories captured by the random growth parameters from above. In the remainder of this document, "cohort error" refers to the $\gamma_{j,t}$ term and the model assumes that it is sampled from a univariate normal distribution with time-constant variance.

$$\gamma_{j,t} \sim N(0, \omega^2) \quad (9)$$

[1] See, for example: Willett, J. B., Singer, J. D., & Martin, N. C. 1998. "The Design and Analysis of Longitudinal Studies of Development and Psychopathology in Context: Statistical Models and Methodological Recommendations." *Development and Psychopathology* 10(2), 395-426.

While the random growth parameters capture systematic changes in a school's average scale score as a function of the prior trend, pandemic impact, and recovery parameters, $\gamma_{j,t}$ captures the influence of unobserved and time-varying factors that are unique to that year and entirely independent of $T$, $P$, and $R$.

Theoretically, the inclusion of $\gamma_{j,t}$ allows the model to control for idiosyncratic, and unobserved events impacting achievement that would have taken place, even in the absence of other modeled influences on achievement (e.g., the impact of the pandemic). However, such an adjustment is based on the very strong assumption of independent cohort error. It would also lead to a very conservative estimate of school-level pandemic impact and recovery estimates since such effects would be confounded with any cohort error that coincided with pandemic and recovery periods. Thus, for $t \in 2021{,}2022$ the model assumes $\omega^2 = 0$. Effectively, this assigns all 2021 and 2022 school deviations to the statewide pandemic setback and achievement target respectively. Although this risks overstating school variation in the magnitude of pandemic setbacks and achievement targets, it makes the analysis more sensitive to detecting school-specific effects. However, by allowing for non-zero $\gamma_{j,t}$ variance during the period prior to the pandemic, the risk is reduced of over adjusting for school-specific pre-pandemic trends and version effects.

All school-level random effects are assumed to be sampled from a multivariate normal distribution:

$$\left[\mu_{0,j}, \mu_{1,j}, \mu_{2,j}, \mu_{4,j}, \gamma_{j,2017}, \gamma_{j,2018}, \gamma_{j,2019}, \gamma_{j,2021}, \gamma_{j,2022}\right] \sim N(\mathbf{0}, \mathbf{G}) \tag{10}$$

where,

$$\mathbf{G} = \begin{bmatrix} \lambda_0^2 & \lambda_{0,1} & \lambda_{0,2} & \lambda_{0,3} & 0 & 0 & 0 & 0 & 0 \\ \lambda_{1,0} & \lambda_1^2 & \lambda_{1,2} & \lambda_{1,3} & 0 & 0 & 0 & 0 & 0 \\ \lambda_{2,0} & \lambda_{2,1} & \lambda_2^2 & \lambda_{2,3} & 0 & 0 & 0 & 0 & 0 \\ \lambda_{3,0} & \lambda_{3,1} & \lambda_{3,2}^2 & \lambda_3^2 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & \omega^2 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & \omega^2 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & \omega^2 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix} \tag{11}$$

The last two rows and columns of the $\mathbf{G}$ matrix is included to emphasize the assumption of zero variance in the cohort error term discussed previously.

$\epsilon_{i,j,t}$ is a within-school residual. To account for changes in the within school variances over time and school, the model estimates $t \times j$ specific variances. Let $N$ equal the total sample size. Conditional on the random effects above, the model assumes that the within-school random effect is sampled from a $N$-dimensional multivariate normal distribution.

$$\epsilon_{i,j,t} \sim N(\mathbf{0}, \mathbf{R}) \sim N\left(\begin{bmatrix}0\\0\\\vdots\\0\end{bmatrix}, \begin{bmatrix}\sigma^2_{j,t} & 0 & \dots & 0\\0 & \sigma^2_{j,t} & 0 & \vdots\\\vdots & 0 & \ddots & 0\\0 & \dots & 0 & \sigma^2_{j,t}\end{bmatrix}\right) \tag{12}$$

## 3.3 Model in Reduced Form

Before discussing estimation, it is useful to rewrite the model in reduced form, bracketing the fixed and random components of the model.

$$\begin{aligned} y_{i,j,t} &= [\tau_{00}(1) + \tau_{10}T + \tau_{20}P + \tau_{30}R] \\ &+\left[\mu_{0,j}(1) + \mu_{1,j}T + \mu_{2,j}P + \mu_{3,j}R + \gamma_{j,t}(1) + \epsilon_{i,j,t}\right] \end{aligned} \tag{13}$$

The reduced form expression leads to a much simplified, matrix expression:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\tau} + \mathbf{Z}\begin{bmatrix}\boldsymbol{\mu}\\\boldsymbol{\gamma}\end{bmatrix} + \boldsymbol{\epsilon} \tag{14}$$

with,

$$\begin{bmatrix}\boldsymbol{\mu}\\\boldsymbol{\gamma}\end{bmatrix} \sim N(\mathbf{0}, \mathbf{G}) \tag{15}$$

and,

$$\boldsymbol{\epsilon} \sim N(\mathbf{0}, \mathbf{R}) \tag{16}$$

When the within-school variance is homoscedastic, $\mathbf{R}$ can be written as an $N \times N$ identity matrix multiplied by a scalar: $\mathbf{R} = \mathbf{I}\sigma^2$.

For a heteroskedastic within-school variance, $\mathbf{R} = \mathbf{W}^{-\frac{1}{2}}\mathbf{I}\mathbf{W}^{-\frac{1}{2}}\sigma^2$, where $\mathbf{W}$ is a diagonal "weight" matrix:

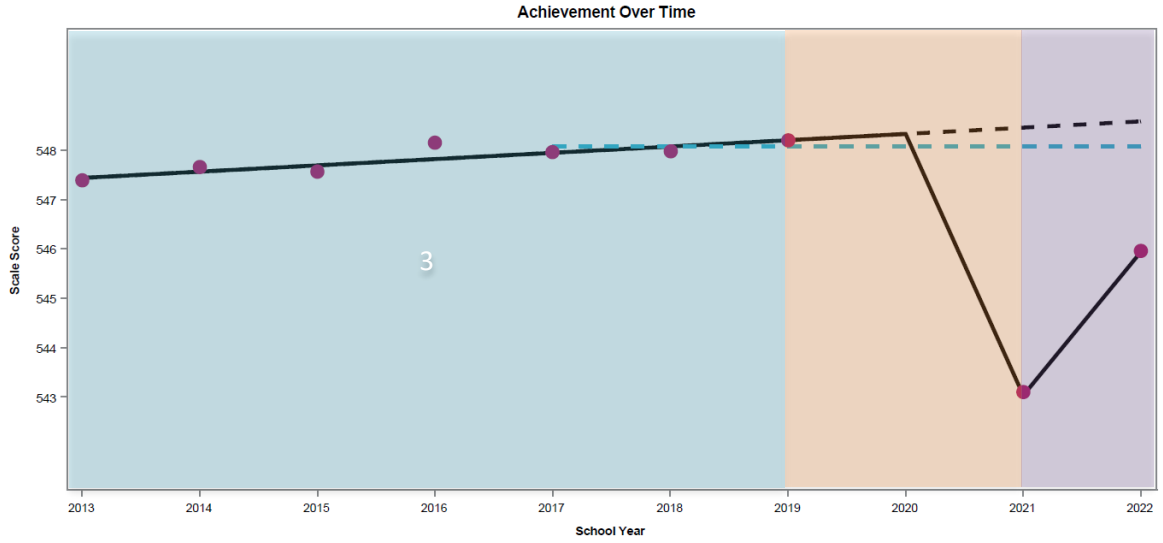
$$\mathbf{W} = \begin{bmatrix}\frac{w_{i,j,t}}{N^{-1}\sum w_{i,j,t}} & 0 & \dots & 0\\0 & \frac{w_{i,j,t}}{N^{-1}\sum w_{i,j,t}} & 0 & \vdots\\\vdots & 0 & \ddots & 0\\0 & \dots & 0 & \frac{w_{i,j,t}}{N^{-1}\sum w_{i,j,t}}\end{bmatrix} \tag{17}$$

where $w_{i,j,t} = \frac{m^{-1}\sum\hat{\sigma}^2_{i,j,t}}{\hat{\sigma}^2_{i,j,t}}$, and $\hat{\sigma}^2_{i,j,t}$ is the sample variance for the $j^{th}$ school. Estimation is carried out using the HPMIXED procedure in SAS.

## 3.4 Four Key Metrics (Post-Estimation)

The specification of the model described above does not directly produce all the state- and school-level estimates of interest. However, these estimates can be obtained as linear combinations of the fixed and random model parameters described above. Although the relevant linear combinations can be represented as matrix calculations, a visual representation of the estimates using the example below is more straight forward.

**Figure 1: Year-Over-Year Trends and Impacts from the Pandemic (EOG Math Grade 3) with Shading to Indicate Three Key Time Periods**




In the graph above, the information can be interpreted as follows:

- The light blue shading indicates the time period for the **Pre-Pandemic Trend Line**. The orange shading indicates the time period for the **Pandemic Impact,** and the purple shading indicates the time period for the **Recovery Thresholds**.
- **Average Scale Score (purple dots)** represents the state's average achievement and is reported in scale scores. It is based on the average of school-level achievement. As a reminder, the average shown in these figures are adjusted for version changes.
- **Achievement Trajectory (solid black line)** is the path achievement in the state took since 2013. Visible as abrupt changes in direction, the trajectory is broken into three distinct parts:
  - Pre-pandemic Trend Line (2013 to 2020)
  - Pandemic Response (2020 to 2021)
  - Recovery Response (2021 to 2022)
- **Extended Trend Line (dashed black line)** is the **Pre-Pandemic Trend Line**, extended into the **Pandemic** and **Recovery** periods. It represents an estimate of state achievement had the state's pre-pandemic trend line (shown by the black solid line) continued into the post-COVID period (2021 to 2022).
- **3-Year-Average Threshold (dashed teal line)** is the three-year average achievement threshold representing the state's average achievement from 2017, 2018, and 2019.

In this report, the information in Figure 1 is summarized in four key metrics constructed from the estimated model parameters.

- The slope of the **Pre-Pandemic Trend Line** or the average year to year change in achievement over the years prior to the pandemic.
- The **Pandemic Impact** or the distance from the achievement trajectory and the extended trend line threshold evaluated at the 2021 time point.
- **Distance to Full Recovery (extended trend line)** or the distance from the achievement trajectory and the extended trendline threshold evaluated at the 2022 time point.
- **Distance to Full Recovery (3-year-average)** or the distance from the achievement trajectory and the three-year average threshold evaluated at the 2022 time point.

Each metric has a state-wide version and a school-level version. The state-level version can be interpreted as the average over the school-level versions. For the state-level version, this report contains point estimates for each metric. The report does not contain estimated metrics for individual schools. Rather, the focus of the report is on characteristics of the school-level distributions for each metric.

## 3.5 Model Features

The model's features detailed above summarize changes in achievement over the pre- and post-pandemic period at both the state and school levels. Pandemic impact and recovery related estimates reported herein account for trends in achievement observed prior the pandemic (i.e., pre-pandemic trends). These state and school pre-pandemic trends provide the basis for a valid and transparent counterfactual, especially in cases where *measured* student achievement is changing rapidly over the years prior to the pandemic. Furthermore, even though the pandemic disrupted testing administration at the end of 2020 in most locales, the extrapolation of the pre-pandemic trend provides estimates of what 2020 achievement would likely have been had it been observed. This is important because the pandemic did not impact in-person schooling until late in the 2019-2020 school year, and the observed level of achievement in that year would have otherwise constituted a natural baseline for estimating the impact of the pandemic. When the pre-pandemic trend is allowed to continue indefinitely, it provides both the basis for estimating the magnitude of the setback as well as a threshold for a full recovery. However, its relevance as counterfactual diminishes over time.

Another important feature of the model relates to the treatment of schools as random effects. In principle, the model could have treated school-level model parameters as fixed effects by fitting the model to each school's data individually. However, this could produce wildly imprecise estimates of school-level pandemic and recovery effects, especially in cases where the number of students in a school is small. In contrast, by assuming a parametric form for school-level model parameters, school-level estimates are also informed by associations estimated at the state level. This leads to estimates with smaller standard errors than would otherwise be obtained if each school's data were considered in isolation.

It is important to note that the enhanced precision obtained from random effect models come at the cost of increased reliance on parametric assumptions. Model fit was assessed visually for each subject and grade at the state level to see if the pre-pandemic trend was being captured. When violations are detected that call into question the validity of the approach when applied to a particular assessment, this is noted in the report and remedial steps are taken when appropriate. Violations could be in the form of non-linear pre-pandemic trends or lack of data causing model fit issues.

Additionally with random effects, the school-specific estimates can be compared to the statewide averages or fixed effects upon which they are centered. They could also be compared to a value of zero, which, would indicate no pandemic setback or a flat pre-pandemic trend.

Finally, the model provides a comprehensive and concise summary of between and within school variation in student achievement in a manner that addresses a broad range of research questions.

For the state-level output, the model answers the following questions:

- For a specific subject/grade, what was the statewide trend in the school achievement in the years prior to the pandemic?
- Accounting for that trend, what was the pandemic impact on school-level achievement statewide in 2021?
- In 2022, how far are schools (on average) from a fully recovering the drop in achievement observed during the pandemic, and how much does that distance depend on the recovery threshold?

For school-level output, the model answers the following questions:

- How much variation is seen in the pre-pandemic trend of schools?
- Is there evidence that schools were differentially impacted by the pandemic?
- What proportion of schools have full recovered by 2022 and how much does it depend on the threshold used?

The Results section provides an overview of the modeling outputs and how to interpret them with these questions in mind.

## 3.6 Conversion of Metrics from Scale Score Units to Effect Size Units

In order to facilitate comparisons across assessments that employ different scaling units, the metrics are converted from scale score units to effect size units by dividing each metric estimate by the standard deviation of the scale-score in 2019. In all cases, this converts the metric into a standardized distance or "effect size." Effect sizes can also be classified as small, medium, or large to assist with interpretation. Various researchers have offered thoughts on what defines a small, medium, and large effect size.

- Cohen describes 0.20 as small, 0.50 as medium, and 0.80 as large (Cohen, Jacob. *Statistical Power Analysis for the Behavioral Sciences.* 2nd ed. Mahwah, NJ: Lawrence Erlbaum, 1988).
- Hattie describes an effect size of 0.40 as the average seen across all interventions, and 0.40 as the "hinge point" (Hattie, John, *Visible Learning: A Synthesis of Over 800 Meta-Analyses Relating to Achievement.* London: Routledge, 2008).
- Kraft suggested < 0.05 as small, 0.05 to 0.20 as medium, and > 0.20 as large based on the distributions of effect sizes and changes in achievement (Kraft MA. "Interpreting Effect Sizes of Education Interventions." *Educational Researcher*. 2020; 49 (4):241-253).

Most researchers agree that what constitutes a small, medium, or large effect size should be considered in the context of previous research conducted in a relevant field or substantive area of research. To place the findings in the context of the educational intervention literature, this report uses Kraft's definitions. This enables researchers to characterize a distance to a full recovery metric in terms of effect of a typical intervention in education (e.g., a tutoring program). For example, recent studies have indicated that double-dose Math instruction (where students have an extra period of Math) provided over the course of the entire year yields, on average, a positive effect size of about 0.20.[2]

[2] In an interview with "The 74," Harvard researcher Tom Kane referenced studies by Stephen W. Raudenbush and Takako Nomi's in Chicago schools and Eric Taylor's similar research in Miami-Dade schools. https://www.the74million.org/article/harvard-economist-offers-gloomy-forecast-on-reversing-pandemic-learning-loss/

Some of the results are also color-coded according to Kraft's definitions as shown below to assist with interpretation.

| Color | Effect Size | Definition |
|---|---|---|
| | Large Negative | Student effect size is less than -0.20 |
| | Medium Negative | Student effect size is -0.20 or greater and less than -0.05 |
| | Small Negative | Student effect size is -0.05 or greater but less than 0.0 |
| | Small Positive | Student effect size is between 0.0 or greater but less than +0.05 |
| | Medium Positive | Student effect size is +0.05 or greater but less than +0.20 |
| | Large Positive | Student effect size is +0.20 or greater |

# 4 Results

The model's three key metrics (Pre-Pandemic Trend, Pandemic Impact, and Distance to a Full Recovery) track the educational achievement of North Carolina students. Collectively, these metrics offer a comprehensive context to understand their collective achievement trajectory before, during, and after the pandemic.

These trajectories are summarized in two different formats: summary tables across assessments and visualizations for a specific assessment.

## 4.1 Summary Tables Across Assessments

- Table that shows the pre-pandemic trend, pandemic impact, and distance to each recovery threshold as an effect size for the state by assessment.
- Bar chart that shows the pre-pandemic trend, pandemic impact, and distance to each recovery threshold as an effect size for the state by assessment. This is the same information as described in the previous bullet point but in a different format.
- Table that shows the percentage of schools in the state that obtained the following benchmarks by assessment:
  - The school's achievement level improved from 2021 and 2022.
  - The school's 2022 achievement level met or exceeded the recovery threshold based on the three-year average achievement.
  - The school's 2022 achievement level met or exceeded the recovery threshold based on the extended trend.

## 4.2 Visualizations for a Specific Assessment

- Graph that shows the year-over-year trends and impacts from the pandemic for a specific assessment. This information is explained in Section 3.3 and it is also summarized across assessments in the table showing pre-pandemic trends, pandemic impact, and distance to each recovery threshold and the bar chart described in Section 4.1.
- Table that shows the sample size and percentage of students meeting or exceeding proficiency for a specific assessment
- Boxplot that shows the school-level distributions for the pre-pandemic trend, pandemic impact, and recovery thresholds for a specific assessment.
- Table that shows the percentage of schools meeting recovery thresholds in 2022. This information is summarized across assessments in 1.c.ii and 1.c.iii described above.

## 4.3 Reflections

When reviewing the results, consider the following questions for reflection:

- During the pre-pandemic period, was achievement increasing, decreasing, or holding steady?
- Assuming the Pre-Pandemic Trend had continued, what was the average magnitude of the pandemic's impact on state achievement across different assessments?
- Which assessments had the largest negative impact from the pandemic?
- In comparison to the Pre-Pandemic Trend extended to 2022, what is the continued gap in achievement?

- In comparison to the achievement threshold based on the most recent three-years prior to the pandemic, what is the continued gap in achievement?
- For each assessment, how does the magnitude of the pandemic compare to the remaining impact? Which assessments have closed the gap in achievement, and which are furthest from a full recovery?