Last Updated: 6.1.2021

education analytics

MISSING DATA GUIDE

Guidance for identifying key questions and next steps for limited or missing data

Introduction

The lack of standardized assessments in Spring 2020 and ongoing school closures due to COVID-19 has presented unique challenges around data availability and data quality for Education Analytics' (EA) state and school district partners. For many state education agencies (SEAs) and local education agencies (LEAs), the absence of Spring 2020 assessment data and limited School Year (SY) 2019-20 attendance data significantly impacted—and continues to impact—existing dashboards and analytics used by educators and decision-makers. To support our partners in navigating this reality, EA developed a systematic approach for investigating the impacts of missing data and making adjustments in order to account for data limitations.

How this guide can help you

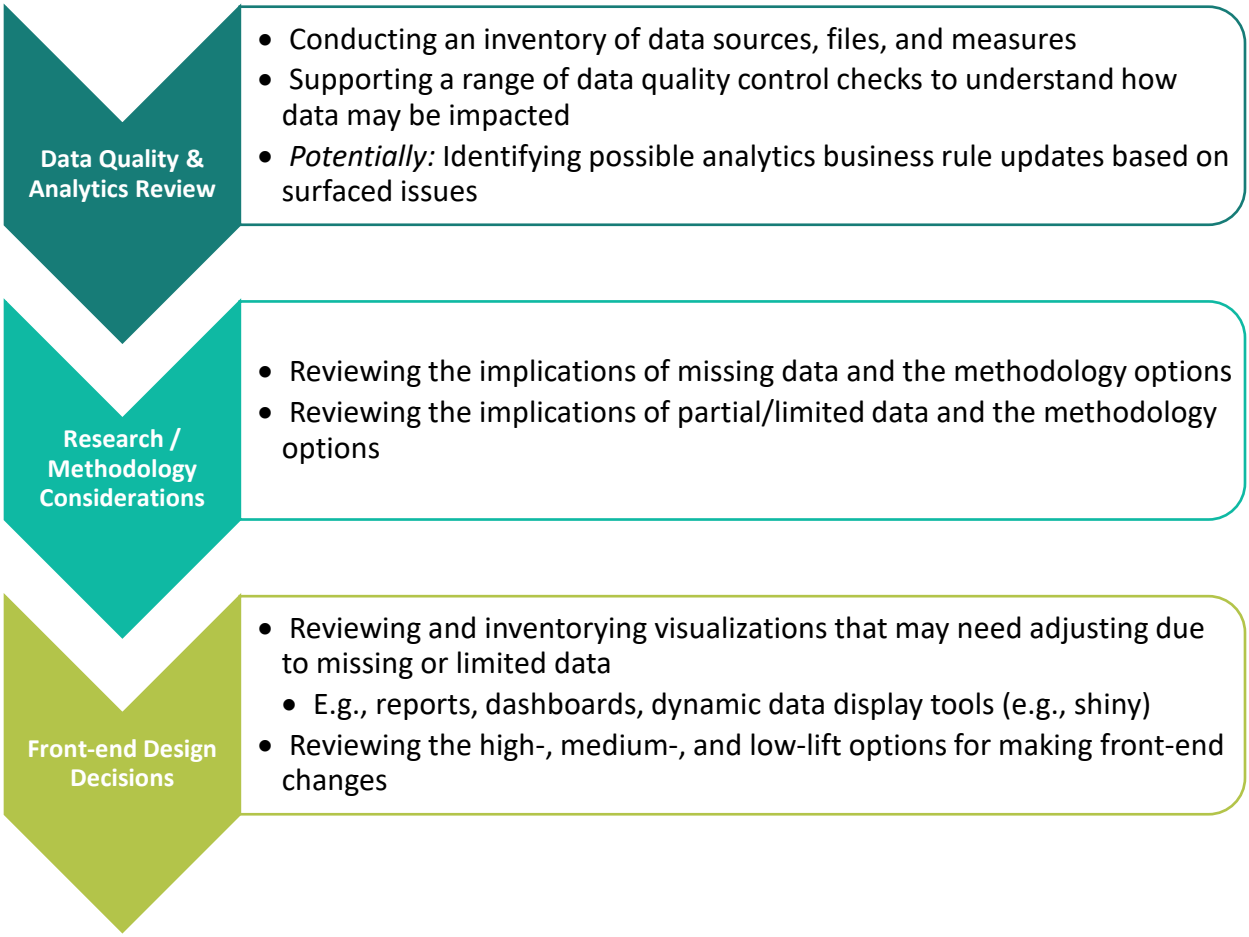
We created this external summary of EA's approach to offer a case study to LEAs and SEAs that may be having their own conversations about the technical implications of missing data. More broadly, our goal is to contribute to the education community's evolving understanding of best practices around addressing missing data to support thoughtful decision-making and to improve outcomes for students.

Case Study: CORE Dashboard

Callouts like these have been incorporated throughout this guide to highlight how we used this the Missing Data Guide in our work with CORE Districts as they considered the responsible calculation and reporting of school and district metrics on the CORE Dashboard.

Overall Approach

EA engaged in a multi-step approach to understand the impact that missing data might have on our projects and on partners' work. This approach focused on three different areas:



The sections below describe each of these areas in more detail and make recommendations around actionable next steps.

Data Considerations

As a first step, we recommend walking through the following list of data quality and data linkage questions to assess which metrics and measures are impacted by missing data. These data quality checks also use a data equity lens to consider the disproportionate impact that missing data may have on groups of historically marginalized and under-represented students. After a data quality review, some findings may lead to business rule changes to your metrics or metric code to account for missing/impacted data.

The following section may be particularly helpful to data and accountability teams that support data analytics and data quality control processes.

Data Quality Questions Checklist

Data and metric overview questions

- ☐ What are all the source files you expect to be missing or incomplete?
- ☐ Which files are necessary for each measure?
    - E.g., is the test file used for both test results & demographic information?
- ☐ What are all the measures you expect to be impacted by missing data?
- ☐ Do you provide any reporting (e.g., shiny, html) that may be impacted by missing data?
- ☐ Do your metrics make any explicit comparisons across years?

Linkage, business rule, and dependency questions

- ☐ Does your student-school linkage assume a full year of school?
- ☐ Does your code include any year-agnostic parameters that should be changed to year-dependent parameters?
    - E.g., is the school year length set to 180 days regardless of year, or is a measure index threshold applied to all years?
- ☐ Do you use any year-dependent parameters (e.g., number of school days is set to 180 for 2020) that need to be updated due to the impacted/limited school year?
    - Are dates explicitly used in any rules?

Case Study: CORE Dashboard

In California, a test file containing an end-of-year test date is typically paired with the Fall Census date to define the period of *continuous enrollment*. Because there was no test file or end-of-year test date in 2019-20, we set February 29th as the 'cutoff' date for end of school year or the last day of inclusion, for continuous enrollment (based on CDE guidance).

- ☐ Do any raw input variables have implicit time dependencies?
    - E.g., to be classified as a "Free and Reduced-Price Lunch" program participant, a student must be enrolled in the program for a certain number of days.
- ☐ Should any project metadata be stored/named differently for a partial year, as compared to a full year?
    - E.g., different directory names, file names, column names?

Data quality and data equity checks

- ☐ Do you anticipate that any data quality sources may be impacted and/or suffer this year, based on how your data sources are collected?
    - E.g., how is attendance data collected, and is that different this year?

Case Study: CORE Dashboard

In California, attendance rates were different in 2020-21 because there was no clear way to mark a student as absent.

- ☐ Are there any additional data quality control checks to perform that might not be run in a normal year?
    - ***Equity check example:*** Are certain data limitations specific to certain student groups?

Case Study: CORE Dashboard

In California, English Learner students were unable to take the ELPAC assessment in Spring 2020, which is used to determine English Learner Reclassification.

- ☐ How are metric distributions different this year, and what impact does this have on reporting?
    - ***Equity check:*** How are metric distributions different this year when disaggregated by student group? How are they different when you disaggregate by student group for semester 1 versus semester 2?
- ☐ How are n-sizes different this year?
    - What impact does this have on reporting (e.g., n-size suppression thresholds)?
    - ***Equity check:*** How are n-sizes/student counts different this year when disaggregated by student group?

Research/Methodology Considerations

Next, we recommend reviewing the impact of missing or limited data on metric methodology and statistical modeling decisions. For EA's scope of work with partners, our team identified two different areas of interest: (1) how to support growth metrics given the missing assessment data from the SY 2019-20 spring semester, and (2) how to support other metrics given the partial data availability in SY 2019-20. These research and methodology considerations are described below.

This section may be particularly helpful to research team members thinking though the implications of missing data on LEA/SEA metrics and how to adjust research methodology going forward. Note that specifics around research and method shifts will be, by necessity, specific to project methodology.

Methodology questions guidance

Research Question: How do we address the absence of missing standardized SY 2019-20 testing data with respect to our growth metrics?

For growth metrics, Student Growth Percentiles (SGPs), and Mean Growth Percentiles (MGPs), there are two potential options:

1. Predict missing assessment outcomes, which require multiple strong methodological assumptions and sets of simulations to verify them.
    a. Note that this approach requires two steps to measure growth: first, predicting spring 2020 test scores, and second, calculating growth from students' actual spring 2019 scores to their predicted spring 2020 scores. This additional complexity requires substantial additional time investments for methodological research, statistical coding, and producing high-quality results.
    b. An additional tradeoff is that, as states waived mandated standardized tests in SY 2019-20, there is no need to predict missing test outcomes in and of itself.
2. Treat SY 2019-20 as a "missing" year and produce metrics for 2020/21.
    a. This means measuring using growth over two years (spring 2019 to spring 2021) rather than over one year.

Case Study: CORE Dashboard

For work in California, we decided to measure growth over a two-year period, from SY 2018-19 to SY 2020-21. Assessment scores in spring 2021 were the outcome variable, and scores in spring 2019 were the primary predictor variable. At the school level, we measure growth for a cohort of students over consecutive years (e.g., grades 4 and 5 at a given school over 2019-20 and 2020-21).

Research Question: How do we address partial SY 2019-20 data availability?

Educational outcomes that are recorded daily, such as enrollment, attendance, and suspension, are not observed after most school buildings closed in March 2020.

Given that only such partial-year data are available, there are two key research questions to consider:

1. To what extent are the partial-year data comparable to the full-year data?
2. To what extent do the metrics using partial-year data have different interpretations compared with those using full-year data?

Case Study: CORE Dashboard

For work in California, we compared full-year attendance data with partial-year data and calculated attendance and chronic absenteeism rates. Overall, the chronic absenteeism rates were highly correlated across years. However, using 2019-20 data as a proxy for full-year data changed the status of chronic absenteeism from chronically absent to not chronically absent ONLY for a small subset of schools. Ultimately, EA and CORE jointly decided to report the 2019-20 chronic absenteeism metric on the CORE dashboard because 1) weak correlations for *some* schools were an understandable outcome given the varied coverage 2019-20 data, and did not undermine the overall value of including the metric, and 2) the dashboard would not be used by CORE data collaborative members for high-stakes decision-making due to the state suspending its accountability requirements.

Front-end Options: User-Centered Design Guidelines and Considerations for Dashboarding

For metrics that are visualized, we recommend conducting a thorough review of any dashboards, reporting tools, or other visualizations that could be impacted by missing or limited data. The EA team took a User-Centered Design (UCD) approach to propose visual changes and recommendations. The following list details a three-step process your agency can use for considering visual and dashboard updates:

1. **Identify the type of data visualization under consideration:**
    - What type of visualization is currently being displayed on the relevant dashboard?
    - Is the missing dataset presented within an historical context of other data, or is it independent?
    - How are the missing data transformed between the front-end dashboard and the back-end database?
    - Is there a value available within the data to identify missing data or are the data simply omitted?
2. **Timeline considerations:**
    - **Internal**: What is the timeline from identification of missing/impacted data to delivery? (This could impact the categories offered to an external stakeholder depending on feasibility).
    - **External**: What is the timeline between the client's response to missing data concerns and the project's due date?
        - Create a timeline of the responses required from a client regarding missing data, depending on the types of project outcomes they need, in tandem with other teams involved in the project (e.g., web development, research).
3. **Solutions:**
    - Depending on the type of data visualization, compile a list of feasible solutions for each metric for internal review and decision-making (referencing the answers to the questions above).
    - Provide the team implementing the visualizations or reporting tools with specifications and designs (if necessary) for implementation.

Case Study: CORE Dashboard

There were not many visuals in the CORE dashboard that required changes due to missing or partial data. All predictive analytics were removed due to missing data, and "No Data" was noted for pages with metrics that had no data to display. We jointly decided to completely hide the front-end view of some pages, as they would not be informative given missing data and could cause confusion.

Case Study: CORE Dashboard

For metrics that had unique business rules ("BRule") for 2019-2020, a tooltip (denoted with an asterisk) was included that provided the BRule definition. Tooltip example: *"Chronic Absenteeism in the dashboard is based on enrollment days for students between the first day of school and the last date of reported data for the 2019-20 school year (may vary by school district)."*

For each project that involved front-end reporting or dashboarding, EA also created a list of front-end adjustments that considered the effort and staff capacity needed to implement them. When reviewing these options, we recommend considering whether the dashboard/reporting tool under consideration reports historical data and allows for comparisons over time; if so, front-end adjustments may include a visual demarcation and explanation for that year of data going forward.

Front-End Adjustment Options

Low Effort	Medium Effort	High Effort
• Append approved annotation or text to a visualization that explains the displayed data • Remove a visualization from the display for year of missing data	• Create "No Data" Scalable Vector Graphics (SVGs) that are rendered in place of a visualization. • If a flag or value is available from the database to indicate missing data, visually signal the data that are impacted (e.g., by adding opacity to impacted data) along with an accompanying annotation appended to the data visualization	• Redesign and rebuild a visualization to demarcate missing or impacted data (and to distinguish from non-impacted/historical data) • Add tooltips explaining data loss of specific points in a visualization that is impacted

Questions?

If you have any questions, please contact Kale Mabin at kmabin@edanalytics.org