Journal of Research on Educational Effectiveness

ISSN: 1934-5747 (Print) 1934-5739 (Online) Journal homepage: www.tandfonline.com/journals/uree20

Addressing Missing Data Due to COVID-19: Two Early Childhood Case Studies

Avi Feller, Maia C. Connors, Christina Weiland, John Q. Easton, Stacy B. Ehrlich, John Francis, Sarah E. Kabourek, Diana Leyva, Anna Shapiro & Gloria Yeomans-Maldonado

To cite this article: Avi Feller, Maia C. Connors, Christina Weiland, John Q. Easton, Stacy B. Ehrlich, John Francis, Sarah E. Kabourek, Diana Leyva, Anna Shapiro & Gloria Yeomans-Maldonado (2025) Addressing Missing Data Due to COVID-19: Two Early Childhood Case Studies, Journal of Research on Educational Effectiveness, 18:1, 226-245, DOI: 10.1080/19345747.2024.2321438

To link to this article: https://doi.org/10.1080/19345747.2024.2321438

Published online: 15 Aug 2024.

JOURNAL OF RESEARCH ON EDUCATIONAL EFFECTIVENESS
2025, VOL. 18, NO. 1, 226–245
https://doi.org/10.1080/19345747.2024.2321438

Routledge
Taylor & Francis Group

METHODOLOGICAL STUDIES

OPEN ACCESS

Check for updates

Addressing Missing Data Due to COVID-19: Two Early Childhood Case Studies

Avi Feller[a], Maia C. Connors[b], Christina Weiland[c], John Q. Easton[b], Stacy B. Ehrlich[d], John Francis[d,*], Sarah E. Kabourek[d], Diana Leyva[e], Anna Shapiro[f], and Gloria Yeomans-Maldonado[g]

[a]Goldman School of Public Policy and Department of Statistics, University of California, Berkeley, California, USA; [b]Institute for Policy Research, Northwestern University, Evanston, Illinois, USA; [c]Marsal Family School of Education and Ford School of Public Policy, University of Michigan, Ann Arbor, Michigan, USA; [d]NORC, University of Chicago, Chicago, Illinois, USA; [e]Department of Psychology, University of Pittsburgh, Pittsburgh, Pennsylvania, USA; [f]RAND Corporation, Washington, DC, USA; [g]Children's Learning Institute, The University of Texas Health Science Center at Houston, Houston, Texas, USA

ABSTRACT

One part of COVID-19's staggering impact on education has been to suspend or fundamentally alter ongoing education research projects. This article addresses how to analyze the simple but fundamental example of a multi-cohort study in which student assessment data for the final cohort are missing because schools were closed, learning was virtual, and/or assessments were canceled or inconsistently collected due to COVID-19. We argue that current best-practice recommendations for addressing missing data may fall short in such studies because the assumptions that underpin these recommendations are violated. We then provide a new, simple decision-making framework for empirical researchers facing this situation and provide two empirical examples of how to apply this framework drawn from early childhood studies, one a cluster randomized trial and the other a descriptive longitudinal study. Based on this framework and the assumptions required to address missing data, we advise against the standard recommendation of adjusting for missing outcomes (e.g., via imputation or weighting). Instead, we generally recommend changing the target quantity by restricting to fully observed cohorts or by pivoting to focusing on an alternative outcome.

ARTICLE HISTORY
Received 21 June 2022
Revised 31 March 2023
Accepted 17 February 2024

KEYWORDS
Missing data; COVID-19; early childhood education; multi-cohort study

One part of COVID-19's staggering impact on education has been to suspend or fundamentally alter ongoing education research projects; see Hedges and Tipton (2020) for a wide-ranging discussion. The goal of this article is to focus on one specific but widespread challenge: how to analyze data from a multi-cohort study in which student assessment data for the final cohort are missing because schools were closed, learning was virtual, and/or assessments were canceled or inconsistently collected due to COVID-19.[1]

CONTACT Avi Feller afeller@berkeley.edu Goldman School of Public Policy and Department of Statistics, University of California, Berkeley, 2607 Hearst Ave, Berkeley, CA 94720, USA.
*John Francis is currently affiliated with the Alan Turing Institute, London, UK.

Supplemental data for this article can be accessed online at https://doi.org/10.1080/19345747.2024.2321438.

[1]While the methodological literature on COVID-19-related complications is just beginning, there are many previous examples of evaluations affected by natural disasters, including hurricanes and wildfires (see Hedges & Tipton, 2020).

The standard recommendation for handling missing outcome information, such as in the *What Works Clearinghouse* (WWC) handbook, is to use missing data adjustment methods like re-weighting or imputation. The goal is typically to minimize bias in estimating the impact on the original study population. While this is a worthwhile objective in other contexts, we argue that considerations differ when whole cohorts are missing outcomes due to the pandemic. Instead, we recommend that researchers report estimates for quantities unaffected by the pandemic, such as by restricting the estimate to fully observed cohorts or by focusing on shorter-term outcomes. In this article, we walk through these choices in the context of a stylized example and two early childhood case studies.

As we show, the primary statistical issue with missing data adjustment methods in this pandemic scenario is that cohorts with missing outcomes provide no information about the impact of the intervention for those outcomes (see Von Hippel, 2007). Adjustment methods like imputation and re-weighting are then equivalent to generalizing impacts from the subset of fully observed cohorts to the original study sample. Thus, such methods essentially ask researchers interpreting the results to "generalize twice"—once from the observed cohorts to the cohorts with missing outcomes and once from the study to the post-pandemic context in which the study will be used. These methods also typically introduce noise relative to a simple complete case analysis alone.

Missing data adjustment methods can also introduce conceptual issues: researchers must make assumptions about a (counterfactual) state of the world in which data collection during the acute phase of the crisis becomes possible. In particular, researchers either assume a world without the pandemic entirely or assume a pandemic world in which data collection is nonetheless possible. Both choices lead to questions about the goal of targeting these estimands in the first place and, for the latter quantity, fundamental challenges with measurement and construct validity. By contrast, estimands based on pre-pandemic quantities alone, such as restricting to complete cases, avoid these concerns.

We assess these questions in some generality in the context of a stylized randomized trial. We then consider two case studies involving early childhood education that were affected by the pandemic. The designs—a cluster-level RCT and a descriptive cohort study—are two important special cases of this broader problem:

- The first case study is a multiple-cohort, cluster RCT evaluation of an assets-based, culturally responsive family intervention aiming to improve Latino kindergarten children's cognitive and academic outcomes in one of the largest school districts in the Southeast. Complete data for pretest, post-test, and follow-up are available for one cohort, but the pandemic precluded follow-up data collection for cohort two.
- The second case study is a decade-long descriptive study of public pre-K access and enrollment in Chicago that was interrupted in its final year by COVID-19. For the first five of six cohorts, researchers obtained standardized test scores from

Buttenheim (2010) and Moreno et al. (2011) give detailed case studies in the context of evaluations during and after earthquakes in Pakistan and Chile, respectively. See van Lancker et al. (2021) for a discussion of COVID-19-related disruptions in biomedical trials.

Kindergarten through third grade; for the final cohort, the third grade assessment, the primary outcome of interest, was canceled.

These study disruptions presented major challenges to both studies, including undermining statistical power and limiting information for future program scale-up and policy decisions. In both case studies, we argue that researchers should generally focus on estimands unaffected by the pandemic, and caution against missing data adjustment methods that target outcomes missing due to pandemic-related disruptions.

We conclude by discussing the challenges associated with the differential and inequitable impact that COVID-19 and the ensuing economic crisis have had on student learning. Our hope is that this case study can further discussion of best practices for education research at an extraordinary time.

Missing Data Due to COVID-19

There is an extensive literature on accounting for missing outcomes, also known as attrition, in experimental and non-experimental education studies.[2] Canonical references for education research include the *WWC Standards Handbook* (Miller et al., 2019 give a thorough introduction) and Puma et al. (2009), among others. Logan (2020) provides a useful overview of these ideas for COVID-19-related missingness.

The goal of this article is to explore recommendations when entire cohorts have missing outcomes due to pandemic-related closures or disruptions. We first outline an idealized setup for COVID-19-related missingness to explain the key ideas. We then discuss applying standard missing data adjustment methods in this idealized setting, before turning to alternative approaches.

Idealized Setup for COVID-19-Related Missingness

We begin with an idealized randomized trial with two (equal-sized) cohorts and three time periods, with randomization to treatment and control conditions within each cohort.[3] Figure 1 gives a schematic of this setup. Cohort A is enrolled at Time 0, with the first follow-up at Time 1 and second follow-up at Time 2; Cohort B is enrolled at Time 1, with the first follow-up at Time 2 and second follow-up at Time 3. At the time of randomization, the substantive goal is to estimate the impact of the intervention on the longer-term outcome, measured two periods after enrollment (at Time 2 for Cohort A and Time 3 for Cohort B). As in many multi-cohort studies, we assume that the original study design required both cohorts to be fully powered for the hypothesized effect size.

The key complication is that, while all enrollment and outcome data are collected as planned for Cohort A, the pandemic disrupts data collection at Time 3 for Cohort B. For this second cohort, we only observe the shorter-term follow-up at Time 2.

[2]Since we focus exclusively on missing outcomes, we will use the terms *missing outcomes* and *attrition* interchangeably.

[3]This is modeled off a similar idealized setup in Logan (2020).

	Time 0	Time 1	Time 2	Time 3 [pandemic]
Cohort A	[Enrolled]	Shorter-term Follow Up ✓	Longer-term Follow Up ✓	---
Cohort B	---	[Enrolled]	Shorter-term Follow Up ✓	Longer-term Follow Up ✗

Figure 1. Schematic of missing data structure.

In this idealized setting, the *overall* attrition rate, the rate of missing outcome data for the entire sample, is 50% (missing cohort B at Time 3). However, the *differential* attrition rate, the difference in rates of missing outcome data between treatment and control groups, is 0% (since treatment arms are equally affected). This would be considered a "low attrition" RCT under WWC standards and could meet WWC standards without reservations given appropriate adjustment.[4]

The study's original target estimand is the longer-term effect for both Cohorts A and B; that is, Cohort A at Time 2 and Cohort B at Time 3. We now turn to how and under what assumptions we could estimate this quantity using missing data adjustment methods. We then turn to alternative quantities of interest.

Missing Data Adjustment

Why Adjust for Missing Data?

Missing data can lead to two main problems; see Puma et al. (2009) for an extended discussion in the context of randomized trials in education. First, restricting the analysis to only those units with observed outcomes can introduce bias when there is *differential* missingness between the treated and control groups. Missing data adjustment methods, such as regression imputation and non-response weighting, reduce this bias by making the treatment groups comparable across baseline covariates.

Second, even when treatment groups are balanced, the cohorts with observed outcomes could differ in systematic ways from the originally planned study sample, potentially limiting the study's external validity. This is particularly relevant when the original sample of students and schools is substantively important, such as when sites are sampled at random. Regardless, as Puma et al. (2009) note, "the study's goal is presumably to obtain *internally valid estimates of the intervention's impact for that sample of schools* [emphasis in original]. If missing data problems lead to a sample of students with complete data in those schools that is not representative of all students in those schools, we believe this is a problem that should be addressed" (p. 16).[5]

[4]The WWC guidelines also provide an exception for "acts of nature" that affect both groups equally. This example would likely fall into that category.

[5]Little and Rubin (2019) also appeal to experimental design considerations: "The advantages of filling in the missing values in an experiment rather than trying to analyze the actual observed data include the following: (i) It is easier to

In general, the first problem is far more pressing than the second. Missing data adjustment methods, however, typically ignore this distinction and handle both problems at once, such as by re-weighting each treatment arm to have the same distribution of observed covariates as the original study population.

In the context of our stylized pandemic example, outcomes are missing for entire cohorts and so we do not need to address differential attrition. Instead, the main concern is that Cohort A alone is not representative of both Cohorts A and B together (i.e., there are important differences across cohorts), and therefore the estimate from Cohort A alone is no longer representative of the original study sample.

Under What Assumptions?

Methods that adjust for missing outcomes traditionally rely on the assumption that outcomes are *Missing At Random (MAR)*. This has two components. First, missingness only depends on observed (baseline) covariates and treatment assignment—and not on unobserved factors. In other words, under MAR, missingness is not a moderator for the treatment effect, after adjusting for baseline covariates. Second, there are no values of baseline covariates and treatment assignment where all outcomes are missing, known as *positivity*. In other words, missingness truly is "random" given observable characteristics.

With pandemic-related missingness, however, we argue that traditional MAR is insufficient because positivity no longer holds—outcomes are missing for an entire cohort. Instead, the necessary assumption is that we can *generalize* the impact from Cohort A to Cohort B, which is typically stronger than MAR.

To see this, first consider generalizing estimates across locations, such as from a study conducted in Chicago to Milwaukee (see, e.g., Dahabreh et al., 2020; Egami & Hartman, 2020; Tipton & Olsen, 2018). In practice, the estimation procedure for generalizing effects (e.g., via re-weighting) will be identical to standard missing data adjustment methods. The underlying assumptions will differ, however: it is challenging to view generalization from the perspective of missing data in part because it is hard to imagine that a student's location (Chicago or Milwaukee) is "random." And since the study was conducted entirely in Chicago, positivity does not hold. Instead, the necessary assumptions for generalizing are: (1) the true subgroup impacts (based on baseline characteristics) are the same for the original and target populations, and (2) that all treatment effect moderators are measured. Under these assumptions, we can then generalize the impact estimate from Chicago to Milwaukee by adjusting for differences in the mix of baseline covariates and school characteristics between the two locations. For pandemic-related missingness, we generalize across cohorts and over time, rather than across locations, but nonetheless require these same strong assumptions.

specify the data structure using the terminology of experimental design (for example as a balanced incomplete block), (ii) it is easier to compute necessary statistical summaries, and (iii) it is easier to interpret the results of analyses because standard displays and summaries can be used." Given the many other complications here, we view these as minor considerations.

The connection to generalizability also clarifies that Cohort B *provides no information* about the intervention's longer-term impacts, just as there is no information in Milwaukee about a study conducted in Chicago. Von Hippel (2007) formalizes this for the scenario when all covariates are observed and we are only missing outcomes: "cases with imputed Y quite literally contain no information about the regression of Y on [treatment]" (p. 88). In a multi-cohort scenario, any additional information necessarily comes from modeling assumptions, such as parametric restrictions or longitudinal structure. Without such structure, using imputed outcomes in the analysis, von Hippel notes, would "simply add noise" (2007, p. 85). Researchers must therefore assess whether obtaining a noisier estimate generalized to the original study sample is more valuable than a more precise estimate for Cohort A alone.

Conceptual Issues with Pandemic Missingness

In standard applications of attrition, we can usually imagine how data collection would be possible; e.g., a student who is absent on the day of the assessment instead attends school. This story is more complicated with pandemic-related missingness, which introduces a range of conceptual questions.

An important first step is to specify how (counterfactual) data collection would be possible.[6] This leads to two versions of the original study estimand:

- *A world without the pandemic.* Here we imagine that the pandemic never occurred, and thus there is no impact on the education system or on participants. Data collection proceeds accordingly.
- *A world with the pandemic:* Here we imagine that the pandemic continues to affect society, the education system, and study participants, but that data collection is possible nonetheless.

Both estimands are representative of the original study population, consistent with the goal of missing data adjustment. Moreover, this distinction is purely conceptual, and in practice, the resulting missing data adjustment procedure is the same for both estimands.

In principle, however, assuming a world *with* the pandemic is more useful for understanding resilience during a disaster and other important mechanisms (Weiland et al., 2021). In the Appendix, we discuss two main concerns with this estimand. First, measurement and interpretation are challenging: it is unclear what it would mean for students to take (counter to fact) a standardized assessment in the context of remote instruction and global uncertainty. Second, the timing of planned data collection is also an important consideration: it is easier to imagine counterfactual data collection occurring a few days earlier than a few months earlier.

[6]See Cro et al. (2020) and Van Lancker et al. (2023). In the context of medical trials, Van Lancker et al. (2023) also suggest an additional estimand: "the effect of the treatment in a post-pandemic patient population, where individuals can suffer from COVID-19 infections but in the absence of administrative and operational challenges caused by the pandemic." While not immediately relevant for this discussion, this formulation is promising for ongoing education trials.

Assuming a world *without* the pandemic avoids these conceptual issues by simply assuming them away; e.g., we can just assume that there are no underlying measurement issues in a counterfactual world without the pandemic.[7] As the underlying assumptions for generalizing from Cohort A to Cohort B are already quite strong, we take this as our working model for our extended discussion below.

Alternative Approaches

Thus far, we have considered estimands that involve missing outcomes due to pandemic-related disruption in Spring 2020 or later; we refer to these as *pandemic estimands*. As we discuss next, we instead recommend focusing on *pre-pandemic estimands* that are based solely on outcomes collected before the pandemic. In particular, we consider two alternative approaches that target pre-pandemic quantities:

1. Complete case analysis
   - *Estimand:* Longer-term effect for Cohort A only
   - *Estimation:* Difference in outcome means at longer-term follow-up for Cohort A only
2. Alternative outcome
   - *Estimand:* Shorter-term effect for both Cohorts A and B
   - *Estimation:* Difference in outcome means at shorter-term follow-up for both Cohorts and B

We refer to estimating quantity 1 as "complete case analysis" because the approach only retains the "complete cases" from Cohort A.[8] We refer to estimating quantity 2 as "alternative outcome" because we shift the outcome of interest from longer-term, as originally planned, to shorter-term outcomes. We can estimate both target quantities via standard estimators (e.g., difference in means or linear regression) for the relevant cohorts.

We note that focusing on pre-pandemic estimands necessarily "moves the goalposts" away from the original study target. We argue that such changes are largely inevitable for studies disrupted by the pandemic—even continuing with the original quantity of interest requires justification—and we, therefore, view this as a less salient concern. That said, for pre-registered studies, the study team should revise their plan before conducting additional analyses (see, e.g., Gehlbach & Robinson, 2018).

Decision-Making Framework

Our recommendation is that, when feasible, researchers should focus on pre-pandemic estimands, though the particular choice of estimand will necessarily depend on the specifics of the study. We now outline some considerations in deciding between approaches.

[7]See Logan (2020) for a helpful discussion of this approach. For example, the author writes that the estimate "represents what you would expect to occur should you run the study again on a new sample of participants drawn from the same population, and under the implementation conditions experienced by the students in the first [cohort]."

[8]In principle, complete case analysis could also target the longer-term impact for both Cohorts A and B under a much stronger Missing Completely At Random assumption. We avoid that approach here.

- *External validity and policy relevance.* A key initial question is: How relevant is the choice of estimand for future, post-pandemic studies? Choosing between the pre-pandemic estimands—a longer-term effect for Cohort A only or shorter-term effects for both Cohorts A and B—will depend on the context. The pandemic estimand is different because estimating that quantity essentially asks that researchers "generalize twice": once from Cohort A to Cohort B; and once from the combined study of Cohorts A and B to the post-pandemic research questions of interest.
- *Standard errors.* There is a tradeoff between the different approaches in terms of statistical power. In general, the estimated shorter-term effect for Cohorts A and B will be the most precise, followed by the longer-term effect for Cohort A alone, followed by the longer-term effect for Cohorts A and B.[9] The degree to which this additional precision weighs in researchers' decision-making may depend in large part on sample size: the smaller the study sample, the more important it may be to researchers to use as many cohorts as possible.
- *Missing data assumptions.* Finally, researchers should carefully assess the underlying assumptions before using missing data adjustment methods. While these assumptions are inherently untestable, we can examine differences in shorter-term effects across cohorts, which allows us to assess these assumptions indirectly. We consider this explicitly in the cluster RCT example below. Following Nguyen et al. (2017), we could also conduct a formal sensitivity analysis for violations of the generalizability assumption in this setting.

Two Empirical Case Studies

In the sections below, we assess each of these questions for our two case studies, demonstrating the utility of this framework for weighing the different approaches. The case studies highlight two substantively important use cases and complement each other. The evaluation of the Food For Thought program is an RCT with two cohorts and a more limited sample size overall. By contrast, the Chicago Pre-K study is a descriptive cohort study with a much larger sample over six cohorts. Thus, the tradeoffs in deciding on a missing data strategy differ between the two case studies and highlight the utility of our proposed framework in two very different applications.

Overview of Culturally Responsive K Study

Background

Our RCT study example estimated the effects of the Food For Thought (henceforth, FFT) program, an assets-based, culturally responsive family intervention that leverages food routines to improve Latino kindergarten children's cognitive and academic outcomes. Family food routines are an ecocultural asset in Latino communities because

[9]This is a heuristic ordering under some simplifying assumptions, such as homoskedasticity. We briefly note recent claims (e.g., Logan, 2020) that missing data adjustment can essentially recover the statistical power of the (infeasible) original study. We argue that any gains in statistical power in the scenario we consider are driven by modeling assumptions and are therefore illusory.

through these practices, Latino parents transmit and preserve their culture and help children to develop their identity as Latinos and exercise the cultural value of *familismo* (strong sense of belonging and loyalty to family) (Evans et al., 2011; Murphey et al., 2014). FFT is a 4-week program taking place in schools; there is one 90-min family session per week where parents learn strategies to foster children's learning during food routines, watch videos of other Latino parents implementing such strategies, and have the opportunity to practice the strategies with their children and receive feedback from program facilitators. The theory of change was that parents who received the FFT program would increase their use of strategies facilitating children's learning during food routines (e.g., engaging in parent-child narratives during mealtime), which would in turn enhance children's learning outcomes (e.g., language). FFT focuses on kindergarten because the transition to school is a "sweet spot" for Latino parents when they are particularly likely to be involved in their children's education (Goldenberg et al., 2001). A pilot study of the FFT program was conducted in 2014–2015 ($N = 10$ families, one school) and a feasibility study was conducted in 2015–2016 ($N = 68$ families, three schools) (Leyva & Skorb, 2017), finding that children's language (vocabulary scores) increased from pretest to end-of-treatment post-test.

As the next phase in FFT development, a cluster RCT was launched in 2018–2019, involving two cohorts of kindergarten children and their families ($N = 261$ students in 13 schools; Cohort A's $N = 129$ in 2018–2019, Cohort B's $N = 132$ in 2019–2020; Leyva et al., 2022).[10] The cluster unit was schools; schools were randomly assigned to FFT or an active control condition. All schools were Title 1 and served at least 20% of Latino students in one of the largest school districts in the Southeast. The original study team estimated the impacts of FFT on several learning outcomes, pre-registering their hypotheses following best practices (Gehlbach & Robinson, 2018).

For this analysis, we focus on language outcomes, particularly vocabulary scores. As shown in Figure 2, assessments were conducted at three time points during the kindergarten year: pretest (September), end-of-treatment post-test (November), and 5-month follow-up (April). Children's language was assessed in schools during a pull-out session using a standardized test (Woodcock-Muñoz Picture vocabulary subtest; Woodcock, Muñoz-Sandoval, Ruef, & Alvarado, 2005) and a non-standardized test (expressive vocabulary items of the IDELA, International Development and Early Learning Assessment; Save the Children, 2017). While children in cohort A were assessed at the three time points, children in cohort B were only assessed at pretest and end-of-treatment due to COVID-19 pandemic.

Estimands. In this case study, the three possible estimands are:

1. Effect at 5-month follow-up for Cohort A;
2. Effect at end of treatment for Cohorts A and B;
3. Effect at 5-month follow-up for Cohorts A and B, in a counterfactual world *without* the pandemic.

[10]There was also a third study cohort planned for 2020–2021. Due to the crisis, this cohort was not recruited. For the purposes of this study, we discuss only the first two study cohorts as there is no data available at all for the third cohort.

	2018-2019			2019-2020		
	Pre-test	End-of-treatment	5-month follow-up	Pre-test	End-of-treatment	5-month follow-up*
Cohort A	Yes ✓	Yes ✓	Yes ✓	--	--	--
Cohort B	---	--	--	Yes ✓	Yes ✓	No ✗

*school year interrupted by COVID-19 pandemic

Figure 2. Cohorts included in the FFT study. *School year interrupted by COVID-19 pandemic.

Estimates

To estimate the impact of the FFT intervention on vocabulary, we conducted three sets of analyses: (1) using 5-month follow-up with Cohort A data only ("complete case"); (2) using 5-month follow-up with cohort A and with Cohort B data imputed;[11] and (3) using end of treatment for both cohorts ("alternative outcome"). In separate work, we have documented the results of the first two sets of analysis using an extended set of outcomes (Leyva et al., 2022). Across our three sets of results, we estimate the effect of being assigned to participate in the FFT program (i.e., Intent to Treat [ITT]) using linear regression of the outcome on the FFT treatment dummy and pretest assessment as well as a set of child, teacher, and school-level variables.[12]

Figure 3 displays the estimates from the three strategies described above. Overall, we found positive ITT estimates for the FFT intervention for vocabulary at the end of treatment and 5-month follow-up outcomes, though the treatment-control difference was statistically significant only for the IDELA vocabulary measure at the end of treatment. Of interest in the present study is the comparison of the impacts of the model with the 5-month follow-up data with Cohort A only and that with 5-month follow-up with Cohort A and with Cohort B data imputed. The magnitude for the cohort B imputed outcome for the IDELA was smaller than that of the model using only Cohort A (ES = 0.14 *vs.* 0.23). This pattern was consistent with the WM-Picture Vocabulary (our standardized outcome), with a slightly larger effect size for the Cohort A-only model (ES = 0.17) than the imputed model (ES

[11]We imputed the 5-month follow-up outcomes for Cohort B using multiple imputation with Stata 17. We imputed 50 data sets using multivariate normal regression. The imputation model included all variables that we specified in our statistical model (e.g., child covariates and pretest scores) as well as an additive treatment indicator. Our imputation model followed the What Works Clearinghouse Requirements relevant to imputing outcome data, specifically that (a) the imputation model must include an indicator variable for intervention status, (b) the imputation model must include all of the covariates used for statistical adjustment in the impact model, and (c) that the imputation must be based on at least five sets of imputations (What Works Clearinghouse, 2021).

[12]For child covariates, we included child's sex, test language of the pretest (e.g., English *vs.* Spanish), test language of the outcome assessment, and cohort. All regression models are adjusted for clustering using robust cluster-corrected standard errors at the school level.

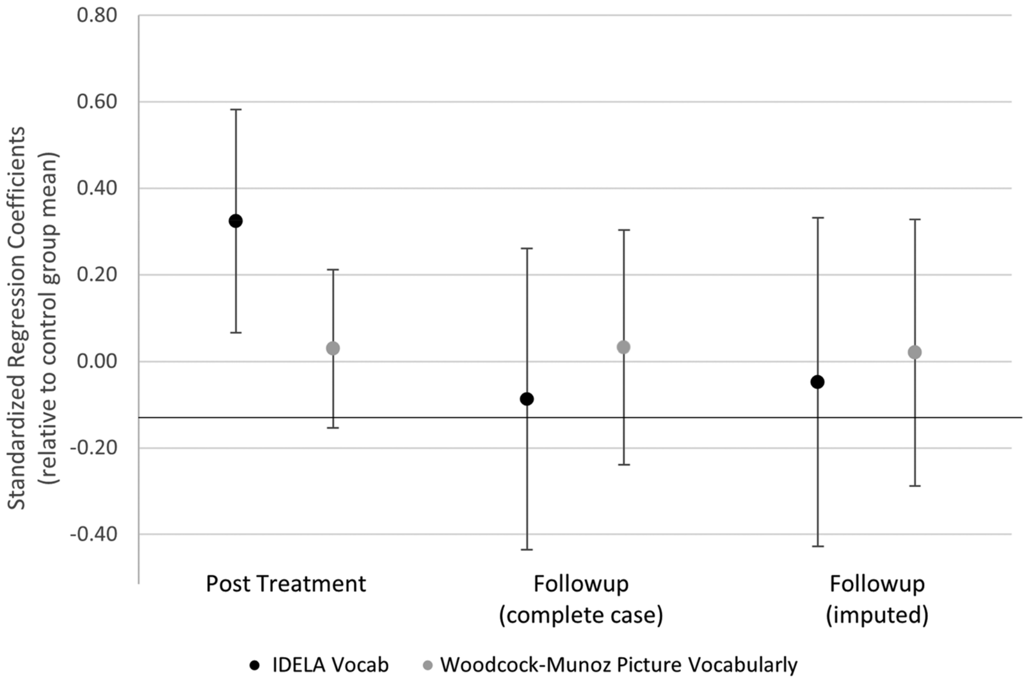

Figure 3. Standardized regression coefficients for RCT study. *Note:* Post-treatment $N = 219$ for IDELA Vocab, $N = 229$ for WM-PV. Follow-up complete case $N = 99$ for IDELA Vocab, $N = 102$ for WM–PV. Follow-up Imputed N = 261 for IDELA Vocab and WM-PV.

= 0.10). Standard errors were smaller for the Cohort A only follow-up estimates than the imputed estimates. As a reminder, each analytic strategy estimates a different quantity.

Applying the Decision-Making Framework

We now apply the decision-making framework to the FFT study.

- **External validity and policy relevance.** The impact for the end-of-treatment outcome using data from Cohorts A and B provides information on the immediate program efficacy for the FFT intervention but no information on the *persistence* of the impacts. The impact on the follow-up outcome data for Cohort A is therefore important for learning whether impacts persist beyond the program. Including this follow-up outcome for Cohort B as well is therefore attractive—at least in principle. Reasoning about a counterfactual data collection world, however, complicates this. A world without the pandemic is therefore a useful fiction, though its value beyond simply focusing on Cohort A is unclear.
- **Estimation: Standard errors.** The standard errors are smallest for the impact estimates immediately post-treatment, and largest for the estimates using imputed follow-up outcomes. These differences, however, are relatively modest overall.
- **Estimation: Missing data assumptions.**

- *Differential attrition:* To assess differential attrition by cohort, we examined children's demographic characteristics (i.e., sex and age at pretest) as well as data on all child-level assessments collected at pretest (e.g., Woodcock-Munoz, IDELA). Except for two assessments, we found no statistically significant differences between cohorts at baseline. For IDELA Math, children in Cohort B scored significantly lower than children in Cohort A ($b = -.07$, $p = .021$); for IDELA executive functioning, children in Cohort B scored higher than those in Cohort A ($b = .11$, $p = .01$).
- *Treatment effect generalizability:* The assumption that subgroup effects generalize from Cohort A to Cohort B seems reasonable; we are not aware of other unobserved, systematic differences between cohorts. The study team assessed this assumption indirectly by estimating overall treatment impacts at the first follow-up separately by cohort and across subgroups; they found no meaningful differences (Authors, 2021).

Summary

Across our decision making framework, focusing on the effect at 5-month follow up for Cohort A (estimand 1) best represents the quantity of interest in the original trial—the lasting impact of FFT, when children are attending in-person school and parents can safely gather in person for the workshops. It also has better construct validity and policy relevance. We then propose to estimate this quantity using the corresponding complete case analysis (i.e., Cohort A alone).

Overview of Chicago Pre-K Study

Background

Our descriptive study example explored whether and how Chicago's school-based pre-K system shifted enrollment patterns after the district implemented a set of policies focused on changing access to and enrollment in school-based pre-K. These policy changes were designed to increase enrollment among student groups identified as most likely to benefit from pre-K but who had historically low enrollment rates and lower school readiness. The goal was that these increases in pre-K enrollment would then lead to more favorable learning outcomes for students over time.

To assess this, we compared patterns of enrollment and geographic access (i.e., distance from home to a school with pre-K and the number of pre-K classrooms nearby) for different student groups before and after the policy changes. Initial results showed that following the policy changes, both access to and enrollment in full-day pre-K expanded substantially among Black students, lowest-income students, and students living in mostly-Black neighborhoods, even as overall school-based pre-K enrollment remained relatively constant (Ehrlich et al., 2020). Enrollment and geographic access patterns were assessed in years before the onset of the COVID-19 pandemic. However, the current case study asked whether these policy changes are also related to more favorable academic outcomes through third grade. For the final cohort of the study, this assessment period overlapped with the pandemic (Figure 4).

	2010-11	2011-12	2012-13	2013-14	2014-15	2015-16	2016-17	2017-18	2018-19	2019-20
Cohort 1	Pre-k	K	1st	2nd	3rd					
Cohort 2		Pre-k	K	1st	2nd	3rd				
Cohort 3			Pre-k	K	1st	2nd	3rd			
Cohort 4				Pre-k	K	1st	2nd	3rd		
Cohort 5					Pre-k	K	1st	2nd	3rd	
Cohort 6						Pre-k	K	1st	2nd	3rd

Cohorts that would have experienced pre-k enrollment prior to policy changes.
Cohorts that would have experienced pre-k enrollment after policy changes.
School year interrupted by COVID-19 pandemic.

Figure 4. Cohorts included in analyses, before and after Chicago's Pre-K access, application, and enrollment policy changes.

The sample for this study is our best approximation of the total number of students who might have considered enrolling in Chicago Public Schools (CPS) for pre-K as a four-year-old in the three years before and after the policy changes ($N = 141{,}938$). We defined six cohorts of students who attended CPS for Kindergarten[13] and were thus eligible to enroll in school-based pre-K as a four-year-old during the 2010–2011 through 2015–2016 school years (see cohorts in Figure 4). Students in Cohorts 1–5 completed third grade before the pandemic, but students in Cohort 6 were in third grade during the 2019–2020 school year, which was interrupted by COVID-19. In this case study, our primary outcome is students' math score from the Measure of Academic Progress (MAP), a standardized, computer-adaptive achievement test administered to all CPS students in grades 2 through 8.[14]

Estimands. In this case study, the four primary estimands are:

1. 3rd grade scores for Cohorts 4 and 5 (relative to Cohorts 1–3)
2. 2nd grade scores for Cohorts 4–6 (relative to Cohorts 1–3)
3. 3rd grade scores for Cohorts 4–6 (relative to Cohorts 1–3), in a counterfactual world *without* the pandemic

Estimates

As in the FFT study, we consider three strategies for estimating the association between math scores and the policy change: (1) using third grade math scores for the first five cohorts only ("complete case"); (2) imputing the missing third math scores for the final cohort; and (3) using second grade math scores for all students ("alternative outcome"). For all three strategies, our primary estimation approach is the same: we use simple linear regression to adjust for the cross-cohort comparisons. Specifically, we regress standardized math assessments for second/third grade on: standardized age in months, an indicator for an individualized education plan (IEP), an indicator for pre-K enrollment as a three year old, an indicator for male, an indicator for English

[13]Plus those who enrolled in CPS for pre-K, but did not continue into CPS Kindergarten.

[14]MAP is produced by the Northwest Evaluation Association (NWEA), which markets standardized assessment used in all 50 states.

language learner, a standardized poverty variable, a standardized social status variable, an indicator for neighborhood type, and a categorical variable for race/ethnicity (white, Black, Latinx, and all other races). Our quantities of interest are the coefficients on the cohort indicators (i.e., the number of years after the policy was implemented), measured relative to the pre-policy average.[15] We also restrict our analysis for all three strategies to the subset of students who have observed second grade assessment data (i.e., complete case using this outcome), which is 75% of the original sample.[16] While this introduces its own complications, it allows us to better isolate the key methodological questions around third grade outcomes.

Figure 5 shows the estimates obtained using these three strategies. Overall, we find small but positive associations of the policy changes with early elementary (second or third grade) math assessments; the magnitude of associations is slightly larger in each successive year of policy implementation. We reiterate that each analytic strategy estimates a slightly different quantity.

Applying the Decision-Making Framework

We now apply the decision-making framework to the Chicago Pre-K study.

- **External validity and policy relevance.** The Chicago Pre-K study was originally designed to investigate policy associations with third grade math assessment scores using data from all six cohorts. The number of school-based full-day pre-k classrooms—a key element of the policy that is the focus of this study—grew steadily across Cohorts 4–6 of our study: just 16% of schools offered full-day pre-k in 2013–2014 when Cohort 4 was eligible to enroll, compared to 40% in 2015–2016 when Cohort 6 was eligible. Since then, Chicago has continued to expand full-day pre-k as it works toward universal access for all four-year-olds. Therefore, the full-day pre-k available to Cohort 6 most closely resembles projected levels of full-day pre-k available to future, post-COVID cohorts. This means that including Cohort 6 in our analyses is an important way to make our estimand more relevant for post-pandemic decision making.
- Third grade assessment scores are typically used as outcomes of interest in early childhood research because third grade represents the first high-stake testing grade. But including third-grade outcomes for Cohort 6 again requires assumptions about a counterfactual data collection world. As in the previous application, a world *without* the pandemic is a useful fiction, though its value beyond simply focusing on pre-pandemic estimands is unclear. Instead, because Chicago administers the same standardized math assessment in second grade and third grade, and in fact uses second grade scores as a baseline for calculating third grade

[15]While this is inherently a descriptive study, we can also view this as a pre-post study in which we compare (adjusted) outcomes between pre- and post-treatment cohorts.

[16]Students who were dropped from the sample include: Pre-K students who did not enroll in CPS for Kindergarten, students who left CPS between Kindergarten and 2nd grade, and students whose second grade scores were simply missing from our dataset.

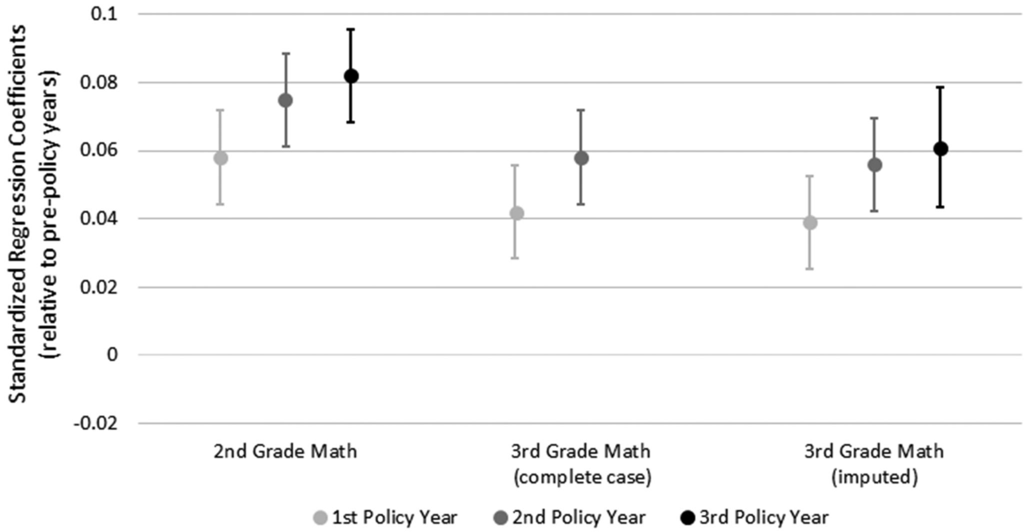

Figure 5. Standardized regression coefficients on year of policy implementation relative to pre-policy years (95% confidence intervals) across three analytic strategies. *Note:* 2nd Grade Math N = 141,938; 3rd Grade Math (complete case) N = 118,498; and 3rd Grade Math (imputed) N = 141,938.

student growth and teacher-level value added measures, we feel confident that second grade math scores also constitute a meaningful outcome of interest for this study. Moreover, other research has documented the importance of measuring math outcomes in early elementary grades given their predictive ability to later outcomes (Claessens & Engel, 2013).

- **Standard errors.** With a sample size of nearly 142,000, statistical power is not a central consideration in choosing among approaches for handling missing data due to COVID in the Chicago Pre-k Study: the standard errors for all estimates across all three approaches range from 0.007 to 0.009 standard deviations. That said, the complete case analysis has a smaller sample size (119,000 *vs.* 142,000), which slightly reduces power, albeit with little substantive difference.
- **Missing data assumptions.**[17] The assumption that subgroup effects are generalizable across cohorts seems difficult to reason about in this case study. In particular, even though baseline characteristics are largely similar across cohorts, it is not reasonable to assume there are no systematic differences between Cohort 6 and the previous Cohorts 1–5. In fact, the average second grade math score in Cohort 6 is more than one-tenth of a standard deviation higher than in Cohort 1, and average third-grade math scores in Cohorts 1–5 vary by more than .06 standard deviations. In this descriptive study, our pre-k policy change of interest is completely confounded with cohort: our study design compares the outcomes of Cohorts 1–3 (which were eligible for pre-k before the policy change) to the outcomes of Cohorts 4–6 (which were eligible for pre-k after the policy change). Moreover, policy implementation,

[17]Unlike for the FFT application, we cannot assess differential attrition in this application.

especially access to full-day pre-k, ramped up substantially from Cohorts 4 to 6. To the extent that the policy changes are associated with outcomes, we thus expect Cohort 6's third grade outcomes to be systematically larger than third grade outcomes in previous cohorts, all else being equal. Indeed, the average second grade math score in Cohort 6 is nearly .04 standard deviations higher than in Cohort 4, and nearly .02 standard deviations higher than in Cohort 5.

Summary

While there are tradeoffs for all choices, we argue that focusing on two pre-pandemic estimands is a reasonable default: (1) third grade scores for Cohorts 4 and 5; and (2) second grade scores for Cohorts 4–6. In the final Chicago Pre-k study, the research team chose to highlight Estimand 2 as the primary quantity of interest and Estimand 1 as a supplemental analysis. We can then estimate these via the corresponding sample quantities, rather than using any missing data adjustment methods.

Conclusion

The COVID-19 crisis presents many challenges to ongoing studies of educational policies and programs—challenges about which the field needs further discussion and guidance. Here, we tackled the common shared challenge of missing data on an entire cohort at a key follow-up time point. We reviewed best practice recommendations for addressing internal validity threats due to missing data (Miller et al., 2019). As we explained, these recommendations may fall short in studies disrupted by COVID-19 because the assumptions that underpin these recommendations were violated. We then provided a new, simple decision-making framework for empirical researchers facing this situation and then discussed two empirical examples of how to apply this framework drawn from early childhood studies—one a cluster randomized trial and the other a descriptive longitudinal study. We showed that what is often the most recommended strategy for addressing missing data problems pre-COVID-19, missing data adjustment methods, such as imputation and reweighting, is likely not advisable in situations with COVID-19-related missingness. Instead, a pivot to focusing either on a fully observed cohort (complete case analysis) or to focusing on an alternative outcome may be more appropriate in many situations. Note, however, that the alternative outcome strategy could undermine the strengths of pre-registration (Gehlbach & Robinson, 2018). This strategy accordingly requires revisiting and revising pre-registration plans *before* analysis.

Just as empirical education researchers have benefited from other best practice guides (e.g., Bloom, 2012; Calonico et al., 2017; Duflo et al., 2007; Imbens & Lemieux, 2008; Lipsey et al., 2015; Murnane & Willett, 2010), we hope our present work might do the same or at least spark further work on this topic. There is still much that can be learned from studies that were compromised by the COVID-19 crisis. As the U.S. and other countries seek to address learning setbacks, the need for rigorous empirical education research to inform evidence-based policymaking has only grown in importance and urgency. Moreover, the studies we present here—while still challenging to analyze—benefit from the relatively simple multi-cohort structure. There are many more complex missingness patterns that require more careful thought, such as studies where some participants are partially observed pre-COVID or studies with far fewer pre-COVID cohorts.

Though critically important, we have not addressed the highly variable experiences of students and their families throughout the pandemic, which likely impact assessment scores and all other measures of academic achievement including attendance, course grades, and disciplinary records. Some students in the United States returned to in-person schooling in Fall 2020, while others attempted "hybrid" models with some in-person learning combined with remote learning, and yet others remained remote well into the 2020–2021 school year. These variable patterns of district decision-making, as well as the degree to which students are able to learn within the paradigm made available to them, are no doubt associated with demographic characteristics (such as race) as well as social and economic characteristics (such as family and community wealth). For example, high-speed internet is not available in all communities, making remote learning difficult or impossible for some students. As such, the COVID-19 crisis interrupted schooling and, most importantly, *learning* differentially and likely inequitably; some students suffered little and others greatly. These disparities likely exist at multiple levels, such as by region, school district, neighborhood, student groups, and individual students. This means that during this time period, we cannot necessarily make usual assumptions about similarities across subgroups, or about the stability of relationships between student, school, neighborhood, or regional characteristics and learning outcomes.

Therefore any attempt to use or impute missing learning outcomes during the pandemic requires researchers to carefully account for all of these issues, which are similarly difficult to measure directly. While some smaller scale studies can address questions of "learning loss" in some cases and "resiliency" in others, it will be much harder to conduct larger scale studies to assess the impact of the COVID-19 and resulting economic crisis on student learning.

Acknowledgments

We are grateful to Erin Hartman, Luke Miratrix, and Elizabeth Stuart for helpful comments. The opinions expressed in this document are those of the authors and do not necessarily represent the views of the institutions involved. The authors thank Isabel Farrar and Amanda Stein for their collaboration on this project. This research would also not have been possible without the support of our dedicated colleagues at Chicago Public Schools and the Charlotte-Mecklenburg Public Schools. They also thank Angela Febles, Yarelin Rivera, Danielle Mayall, and Davidson College's undergraduate student research assistants who made this work possible.

Open Research Statements

Study and Analysis Plan Registration

There is no study and analysis plan registration associated with this manuscript.

Data, Code, and Materials Transparency

The data, code, and materials underlying the results reported in this manuscript are not publicly available.

Design and Analysis Reporting Guidelines

Not applicable.

Transparency Declaration

The lead author (the manuscript's guarantor) affirms that the manuscript is an honest, accurate, and transparent account of the study being reported; that no important aspects of the study have been omitted; and that any discrepancies from the study as planned (and, if relevant, registered) have been explained.

Replication Statement

This manuscript reports an original study.

Disclosure Statement

No potential conflict of interest was reported by the author(s).

Funding

The research reported here was supported by the Institute of Education Sciences, U.S. Department of Education, through grants R305A180510, R305B150012, and R305B170015, as well as the Brady Education Foundation and Davidson College.

References

Bloom, H. S. (2012). Modern regression discontinuity analysis. *Journal of Research on Educational Effectiveness*, *5*(1), 43–82. https://doi.org/10.1080/19345747.2011.578707

Buttenheim, A. (2010). Impact evaluation in the post-disaster setting: A case study of the 2005 Pakistan earthquake. *Journal of Development Effectiveness*, *2*(2), 197–227. https://doi.org/10.1080/19439342.2010.487942

Calonico, S., Cattaneo, M. D., Farrell, M. H., & Titiunik, R. (2017). rdrobust: Software for regression-discontinuity designs. *The Stata Journal: Promoting Communications on Statistics and Stata*, *17*(2), 372–404. https://doi.org/10.1177/1536867X1701700208

Claessens, A., & Engel, M. (2013). How important is where you start? Early mathematics knowledge and later school success. *Teachers College Record: The Voice of Scholarship in Education*, *115*(6), 1–29. https://doi.org/10.1177/016146811311500603

Cro, S., Morris, T. P., Kahan, B. C., Cornelius, V. R., & Carpenter, J. R. (2020). A four-step strategy for handling missing outcome data in randomised trials affected by a pandemic. *BMC Medical Research Methodology*, *20*(1), 208. https://doi.org/10.1186/s12874-020-01089-6

Dahabreh, I. J., Robertson, S. E., Steingrimsson, J. A., Stuart, E. A., & Hernán, M. A. (2020). Extending inferences from a randomized trial to a new target population. *Statistics in Medicine*, *39*(14), 1999–2014. https://doi.org/10.1002/sim.8426

Duflo, E., Glennerster, R., & Kremer, M. (2007). Using randomization in development economics research: A toolkit. *Handbook of Development Economics*, *4*, 3895–3962.

Egami, N., & Hartman, E. (2020). Elements of external validity: Framework, design, and analysis. *Design, and analysis*. Working paper.

Ehrlich, S. B., Connors, M. C., Stein, A. G., Francis, J., Easton, J. Q., Kabourek, S. E., & Farrar, I. C. (2020). *Closer to home: More equitable pre-K access and enrollment in Chicago*. UChicago Consortium on School Research, NORC at the University of Chicago, and Start Early.

Evans, A., Chow, S., Jennings, R., Dave, J., Scoblick, K., Sterba, K. R., & Loyo, J. (2011). Traditional foods and practices of Spanish-speaking Latina mothers influence the home food environment: Implications for future interventions. *Journal of the American Dietetic Association*, *111*(7), 1031–1038. https://doi.org/10.1016/j.jada.2011.04.007

Gehlbach, H., & Robinson, C. D. (2018). Mitigating illusory results through preregistration in education. *Journal of Research on Educational Effectiveness*, *11*(2), 296–315. https://doi.org/10.1080/19345747.2017.1387950

Goldenberg, C., Gallimore, R., Reese, L., & Garnier, H. (2001). Cause or effect? A longitudinal study of immigrant Latino parents' aspirations and expectations, and their children's School performance. *American Educational Research Journal*, *38*(3), 547–582. https://doi.org/10.3102/00028312038003547

Hedges, L., & Tipton, E. (2020). *Addressing the challenges to educational research posed by COVID-19*. Northwestern University Institute for Policy Research.

Imbens, G. W., & Lemieux, T. (2008). Regression discontinuity designs: A guide to practice. *Journal of Econometrics*, *142*(2), 615–635. https://doi.org/10.1016/j.jeconom.2007.05.001

Leyva, D., & Skorb, L. (2017). Food for thought: Family food routines and literacy in Latino kindergarteners. *Journal of Applied Developmental Psychology*, *52*, 80–90. https://doi.org/10.1016/j.appdev.2017.07.001

Leyva, D., Weiland, C., Shapiro, A., Yeomans-Maldonado, G., & Febles, A. (2022). A strengths-based, culturally responsive family intervention improves Latino kindergarteners' vocabulary and approaches to learning. *Child Development*, *93*(2), 451–467. https://doi.org/10.1111/cdev.13698

Lipsey, M. W., Weiland, C., Yoshikawa, H., Wilson, S. J., & Hofer, K. G. (2015). The prekindergarten age-cutoff regression-discontinuity design: Methodological issues and implications for application. *Educational Evaluation and Policy Analysis*, *37*(3), 296–313. https://doi.org/10.3102/0162373714547266

Little, R. J., & Rubin, D. B. (2019). *Statistical analysis with missing data* (Vol. 793). John Wiley & Sons.

Logan, J. (2020). Missing school-based data due to COVID-19: Some guidelines. Working paper. https://edarxiv.org/24bsu/

Miller, D., Spybrook, J., & Cassidy, S. (2019). *Missing data in group design studies: Revisions in WWC Standards Version 4.0*. US Department of Education, Institute of Education Sciences. https://ies.ed.gov/ncee/wwc/Docs/Multimedia/WWC-Missing-Data-508.pdf

Moreno, L., Treviño, E., Yoshikawa, H., Mendive, S., Reyes, J., Godoy, F., Del Río, F., Snow, C., Leyva, D., Barata, C., Arbour, M., & Rolla, A. (2011). Aftershocks of Chile's earthquake for an ongoing, large-scale experimental evaluation. *Evaluation Review*, *35*(2), 103–117. https://doi.org/10.1177/0193841X11400685

Murnane, R. J., & Willett, J. B. (2010). *Methods matter: Improving causal inference in educational and social science research*. Oxford University Press.

Murphey, D., Guzman, L., & Torres, A. (2014). America's Hispanic children: Gaining ground, looking forward. Publication #2014-38. Child Trends. Retrieved from http://www.childtrends.org

Nguyen, T. Q., Ebnesajjad, C., Cole, S. R., & Stuart, E. A. (2017). Sensitivity analysis for an unobserved moderator in RCT-to-target-population generalization of treatment effects. *The Annals of Applied Statistics*, *11*(1), 225–247. https://doi.org/10.1214/16-AOAS1001

Puma, M. J., Olsen, R., Bell, H., & Price, C. (2009). *What to do when data are missing in group randomized controlled trials (NCEE 2009-0049)*. National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, U.S. Department of Education.

Tipton, E., & Olsen, R. B. (2018). A review of statistical methods for generalizing from evaluations of educational interventions. *Educational Researcher*, *47*(8), 516–524. https://doi.org/10.3102/0013189X18781522

van Lancker, K., Tarima, S., Bartlett, J., Bauer, M., Bharani-Dharan, B., Bretz, F., Flournoy, N., Michiels, H., Parra, C. O., Rosenberger, J. L., & Cro, S. (2021). Estimands and their estimators for clinical trials impacted by the COVID-19 pandemic: A report from the NISS Ingram Olkin Forum Series on Unplanned Clinical Trial Disruptions. arxiv [stat.ME] 2202.03531.

Van Lancker, K., Tarima, S., Bartlett, J., Bauer, M., Bharani-Dharan, B., Bretz, F., Flournoy, N., Michiels, H., Olarte Parra, C., Rosenberger, J. L., & Cro, S. (2023). Estimands and their estimators for clinical trials impacted by the COVID-19 pandemic: A report from the NISS Ingram Olkin Forum Series on unplanned clinical trial disruptions. *Statistics in Biopharmaceutical Research*, *15*(1), 94–111. https://doi.org/10.1080/19466315.2022.2094459

Von Hippel, P. T. (2007). Regression with missing Ys: an improved strategy for analyzing multiply imputed data. *Sociological Methodology*, *37*(1), 83–117. https://doi.org/10.1111/j.1467-9531.2007.00180.x

Weiland, C., Greenberg, E., Bassok, D., Markowitz, A., Guerrero Rosada, P., Luetmer, G., Abenavoli, R., Gomez, C., Johnson, A., Jones-Harden, B., Maier, M., McCormick, M., Morris, P., Nores, M., Phillips, D., & Snow, C. (2021). *Historic crisis, historic opportunity: Using evidence to mitigate the effects of the COVID-19 crisis on young children and early care and education programs*. University of Michigan Education Policy Initiative and Urban Institute Policy Brief. https://edpolicy.umich.edu/files/EPI-UI-Covid%20Synthesis%20Brief%20June%202021.pdf

What Works Clearinghouse (2021). *What Works Clearinghouse standards handbook version 4.1*. US Department of Education, Institute of Education Sciences. National Center for Education Evaluation and Regional Assistance. https://ies.ed.gov/ncee/wwc/Docs/ref-erenceresources/WWC-Standards-Handbook-v4-1-508.pdf