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

Avi Feller[1], Maia C. Connors[2], Christina Weiland[3], John Q. Easton[4], Stacy B. Ehrlich[5], John Francis[5], Sarah E. Kabourek[5] , Diana Leyva[6], Anna Shapiro[7], and Gloria Yeomans-Maldonado[8]

[1]University of California, Berkeley

[2]Start Early

[3]University of Michigan

[4]University of Chicago Consortium on School Research

[5] NORC at the University of Chicago

[6]University of Pittsburgh

[7]University of Virginia

[8]Children's Learning Institute at The University of Texas Health Science Center at Houston

Author Note: Following the first three authors, authors are listed in alphabetical order. We are grateful to Luke Miratrix and Elizabeth Stuart for helpful comments. 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. The opinions expressed in this document are those of the authors and do not necessarily represent 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.

Direct correspondence to Avi Feller, 2607 Hearst Ave, Berkeley, CA 94720, Email: afeller@berkeley.edu

Abstract

One part of COVID-19's staggering impact on education has been to suspend or fundamentally alter ongoing education research projects. This paper 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 the missing data, we advise against the standard recommendation of adjusting for missing outcomes (e.g., via imputation or weighting). Instead, changing the target quantity by restricting to fully-observed cohorts or by pivoting to focusing on an alternative outcome may be more appropriate.

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

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 paper 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]

We argue that analytic considerations in this context differ from standard pre-COVID-19 recommendations regarding missing data. In particular, the *What Works Clearinghouse* (WWC) handbook gives extensive guidance on whether and how to account for missing outcome information, such as by re-weighting the observed data or using imputation methods. These recommendations, however, rest on assumptions and other considerations that are unlikely to hold when entire cohorts are missing outcomes due to COVID-19.

There are two primary issues. The first is conceptual: researchers interested in adjusting for missing outcomes must be explicit about the assumed (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.

The second issue is statistical: cohorts with missing outcomes provide no information whatsoever about the impact of the intervention (see van Hoppel, 2007), and adjustment methods (e.g., imputation or re-weighting) are equivalent to generalizing impacts from the subset of fully observed participants to the original study sample. Thus, such methods essentially ask researchers 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.

Recognizing the many different considerations in this missing data situation, we provide a new, simple decision-making framework for empirical researchers. Our framework has two parts: (1) what is the quantity of interest? and (2) how should we estimate it?

[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 and Tipton, 2021). 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.

We first use this framework 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 pre-test, 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 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.

Applying the decision-making framework to these case studies for choosing an estimand, we argue that researchers should generally focus on *pre-pandemic* estimands, and caution against estimands that include outcomes missing due to pandemic-related disruptions. For estimation, we argue that complete case analysis — here, restricting the analysis to cohorts with outcomes *not* disrupted by the pandemic — or otherwise using fully observed outcomes are natural choices and caution against missing data adjustment methods.

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

A Statistical Framework for Reasoning about 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 *What Works Clearinghouse (WWC) Standards Handbook* (Miller et al.

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

[2019] give a thorough introduction) and Puma et al. (2009), among others. Logan (2020) provides an excellent overview of these ideas for COVID-19-related missingness.

The goal of this paper is to explore recommendations when entire cohorts have missing outcomes due to pandemic-related closures or disruptions. In the Appendix, we also explore related statistical issues in a stylized pre-pandemic study that mirrors our examples in the main text.

Idealized setup for COVID-19-related missingess

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. 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). For Cohort A, all enrollment and outcome data are collected as planned. For Cohort B, we observe the first follow up at Time 2.

While we can estimate longer-term effects for the "complete case subset" of Cohort A, the main challenge is that the pandemic prevents data collection for longer-term follow up for Cohort B at Time 3.

Figure 1: Schematic of missing data structure

	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 ✗

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.[3]

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

Decision-Making Framework: Overview

We now develop a simple decision-making framework for reasoning about missing data due to COVID-19. This framework has two main parts:

- **Estimands.**
    - *Possible estimands:* Does the target quantity involve missing outcomes in Spring 2020? If so, what are the assumptions about the (counterfactual) data collection?
    - *Measurement and construct validity:* Is the estimand a reasonable quantity?
    - *External validity:* How relevant is this estimand for future, post-pandemic studies?

- **Estimation.**
    - *Assumptions:* Are the assumptions for estimation satisfied, and can we assess them?
    - *Standard errors:* How precise is the estimate?

We now consider these two parts in more detail.

Decision-Making Framework: Estimands

The first question researchers must address is to specify the target estimand. There are four primary estimands of interest in this stylized example, which we divide into: *pre-pandemic estimands* **(**those based solely on outcomes collected prior to the pandemic) and *pandemic estimands* (those that involve missing outcomes due to pandemic-related disruption in Spring 2020 or later). For the pandemic estimands, we must further specify how (counterfactual) data collection would be possible.[4]

Pre-pandemic estimands:

1. *Longer-term effect for Cohort A only*.
2. *Shorter-term effect for Cohorts A and B*.

Pandemic estimands:

3a. *Longer-term effect for Cohorts A and B, world <u>without</u> 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.

3b. *Longer-term effect for Cohorts A and B, world <u>with</u> the pandemic:* Here we imagine that the pandemic continues to affect society, the education system, and study participants, but that data collection is possible despite all of this.

[4] See Cro et al. (2020) and Van Lecker et al. (2022). In the context of medical trials, Van Lecker et al. (2022) 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.

An important but subtle point is that cohorts with missing longer-term outcomes (Cohort B in this example) *provide no information whatsoever* about the intervention's longer-term impacts (see supplementary materials and von Hippel, 2007).[5] As we discuss below, estimation approaches that target the pandemic estimands (e.g., imputation or nonresponse weighting) are therefore equivalent to generalizing longer-term impacts from Cohort A to Cohort B. Thus, our discussion of "pandemic estimands" is necessarily conceptual, since we can never estimate these quantities directly.

We now consider these estimands across several dimensions, turning to estimation and assumptions below.

Measurement and construct validity for data collection in a world with the pandemic

Unlike estimands that rely entirely on pre-pandemic quantities, focusing on longer-term effects for Cohorts A and B *in a world <u>with</u> the pandemic* (Estimand 3b) raises fundamental questions of measurement and construct validity. Among other concerns, researchers proposing this estimand should be prepared to grapple with these questions:

- *Measurement and interpretation.* 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. Had they taken the assessment, many students would likely have scored lower than their counterparts in previous cohorts due to lost learning time as well as added stress and trauma (for a detailed discussion of measurement questions, see, for example, Boyer, 2021). Moreover, emerging evidence indicates that many aspects of the COVID-19 pandemic and related challenges (e.g., access to high-speed internet) and economic repercussions were inequitably distributed—hitting some communities, and thus some students, harder than others (Weiland et al., 2021). Further, the theory of change of essentially all existing educational interventions and policies never included effects persisting through a historic global pandemic (Weiland & Morris, 2022). Accordingly, there is currently almost no guidance on how to interpret findings from longitudinal studies that pre-date the COVID-19 crisis. Combining cohorts, if data are available in Spring 2020, is thus in our view ill-advised.

- *Timing.* The timing of planned data collection is also an important consideration. For instance, consider data collection planned for mid-March 2020 versus data collection planned for mid-May 2020. We can more easily imagine that—had we been able to collect data in March 2020—children's outcomes (e.g., math skills, vocabulary, academic grades) would be more comparable to pre-COVID-19 levels a few days into the pandemic compared to a few months into it. In the Chicago Pre-K case study below, the outcome of interest is a third grade standardized assessment scheduled for May, at

[5] This is the case when all covariates are observed and we are only missing outcomes. As von Hippel writes: "Cases with imputed Y quite literally contain no information about the regression of Y on [treatment]" (2007, p. 88).

the end of the school year. Assessment scores thus reflect students' learning up to the assessment date — including the math that they learned in third grade. Had this assessment been given to third graders in early March 2020, prior to the pandemic, scores would reflect two months less learning time than intended. Alternatively, if the assessment had been given remotely in May 2020, scores would reflect two months of suboptimal learning conditions characterized by great uncertainty, upheaval, and trauma, all while schools were closed and/or shifting to other modes of instruction. Either scenario would almost certainly result in systematically lower assessment scores for the final cohort than for previous cohorts, all else being equal. Finally, while we largely focus on missingness in spring 2020, we can apply similar reasoning to missing data due to canceled or disrupted data collection later in the pandemic.

External validity for post-pandemic studies

The next question is how these estimands inform future decisions. Importantly, we consider this for the (possibly infeasible) quantities of interest, irrespective of our ability to estimate them. Thus, if we *were* able to obtain reasonable estimates during the most acute phases of the pandemic, how would that inform subsequent policy decisions? How likely are those quantities to generalize to, for instance, spring 2020 populations in different contexts or locations, *and to future populations*?

In considering these questions, we anticipate (as well as hope) that the conditions present in spring 2020 in the U.S. are anomalous. Even as the pandemic has stretched on far longer than most people expected, the context of fully remote learning environments, widespread sickness and death, no vaccines available, and traumatic levels of economic and social upheaval are unlikely to persist or simultaneously recur in the foreseeable future. For example, by spring 2021, more than half of schools in the U.S. had returned to in-person instruction (Burbio, 2021). Moreover, while it is difficult to forecast the future of education policymaking in a pandemic (or hopefully post-pandemic) world, we anticipate that policymakers and researchers will continue to care more about the persistence of effects than about impacts immediately post-treatment. We can then assess the different quantities of interest:

1. *Longer-term effect for Cohort A.* Since we are primarily interested in longer-term effects, the question for this estimand is whether Cohort A (the "complete case" subset) represents a well-defined group or population on its own. If so, it is reasonable to use this (more limited) estimand to inform future studies.[6]

2. *Shorter-term effect for Cohorts A and B.* This estimand avoids the issue of redefining the population but at the cost of losing information on longer-term effects.

[6] This same question arises in standard (pre-pandemic) missing data applications—and there is no clear consensus on the "correct" estimand there either. For instance, Puma et al. (2009) argue that we should prefer estimands that include both Cohorts A and B and that are closest to the study's original goal. However, Puma et al. also note that this is not a universally held view, especially if the corresponding schools or districts are a convenience sample.

3a. *Longer-term effect for Cohorts A and B, world without the pandemic.* This is a reasonable estimand when there is strong reason to target both Cohorts A and B as the study population of interest. Importantly, this estimand 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.

3b. *Longer-term effect for Cohorts A and B, world with the pandemic.* In principle, this estimand would be useful for understanding resilience during a disaster and other important mechanisms (Weiland & Morris, 2021). However, since we have no information on such behavior in practice due to data collection constraints, this estimand is less attractive.

Moving the goalposts

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., Gelbach & Robinson, 2018).

Decision-Making Framework: Estimation and Assumptions

Once we have defined the quantity of interest, we then consider possible estimation strategies:

1. *Longer-term effect for Cohort A only*: *Complete case analysis.* This is the estimated impact on longer-term outcomes using Cohort A only.[7]

2. *Shorter-term effect for Cohorts A and B: Alternative outcome*. This is the estimated impact on shorter-term outcomes using both Cohorts A and B.

3. *Longer-term effect for Cohorts A and B: Adjust for missing outcomes.* For both estimands 3a and 3b, we can adjust participants with observed outcomes (Cohort A) to have similar baseline covariates to the overall sample (Cohorts A and B together), such as by re-weighting or imputation.

As we note above, the estimated impact adjusting for missing outcomes is equivalent to generalizing (also known as "transporting") the estimated impact from Cohort A only to Cohort

[7] 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.

B. We discuss this in more depth in the Appendix; see Tipton and Olson (2018), Dahabreh et al. (2020), and Egami and Hartman (2020) for accessible overviews. Importantly, the estimated longer-term impact for Cohort B is the same for both Estimands 3a and 3b; that is, assumptions about a counterfactual world with data collection are critical for interpretation and validity but do not affect the estimate itself.

Missing At Random (MAR) Assumption

Methods that adjust for missing outcomes fundamentally rely on the assumption that outcomes are *Missing At Random (MAR)* given baseline covariates; that is, missingness only depends on observed (baseline) covariates and treatment assignment — and not on unobserved factors. While this assumption is not directly testable, researchers can assess it indirectly by examining differential attrition; see WWC guidelines (2021). We recommend this as part of standard missing data diagnostics.

Importantly, in cases in which entire groups are missing outcomes, adjustment methods also require additional assumptions on the outcome model (even if these assumptions are often implicit when using imputation). The most common assumption is known as (*mean) generalizability of treatment effects*, which states that: (1) the true subgroup impacts (based on baseline characteristics) are the same for Cohorts A and B, and (2) that all treatment effect moderators are measured.[8] See Dahabreh et al. (2019; 2020) and Egami & Hartman (2020) for closely related technical discussion.[9]

The assumption that subgroup effects are constant across cohorts is a strong assumption in pre-pandemic studies, but is especially strong with pandemic-related missingness:

3a. *Longer-term effect for Cohorts A and B, world <u>without</u> the pandemic.* For this estimand, the pandemic does not occur, so the assumption that there are no cohort differences is not as severe. Nonetheless, this again raises questions about the role of generalizing the impacts from Cohort A to Cohort B.

3b. *Longer-term effect for Cohorts A and B, world <u>with</u> the pandemic.* Even though this estimand is based on a world with the pandemic, the assumption of generalizable treatment effects requires that the pandemic in no way changed the intervention's impacts. This is a very strong restriction.

Regardless of the estimand, we can examine differences in *shorter-term* effects across cohorts, which again allows us to indirectly assess this assumption. We consider this explicitly in the

[8] There is extensive discussion of this point in the literature on generalizability and transportability. A closely related alternative discussion can instead be framed in terms of an "exclusion restriction." See Egami & Hartman (2021) for a thoughtful discussion.

[9] We can frame this technical assumption in two ways: (1) that the (reweighted) estimates from Cohort A are *internally valid* for the entire sample; or that (2) the (reweighted) estimates from Cohort A are *externally valid* for Cohort B. These are equivalent statement despite different language. See Appendix for additional discussion.

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.

Statistical power

There is a trade-off 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.[10] 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.

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. These case studies differ across several factors, including sample size, timing of the interruption due to COVID-19, and type of study (RCT vs a descriptive cohort study).

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 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) (Murphey et al., 2014; Evans et al., 2011). 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

[10] This is a heuristic ordering under some simplifying assumptions, such as homoskedasticity.

the transition to school is a "sweet spot" for Latino parents, as 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, 1 school) and a feasibility study was conducted in 2015-2016 (*N* = 68 families, 3 schools) (Authors, 2017), finding that children's language (vocabulary scores) increased from pre-test 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; Authors, 2021; 2022).[11] 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 impacts of FFT on several learning outcomes, pre-registering their hypotheses following best practices (Gelbach & 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: pre-test (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 pre-test and end-of-treatment due to COVID-19 pandemic.

Figure 2. Cohorts included in the FFT study.

	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

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

1. Effect at 5-month follow up for Cohort A

[11] 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.

2. Effect at end of treatment for Cohorts A and B

3a. Effect at 5-month follow up for Cohorts A and B, world without the pandemic

3b. Effect at 5-month follow up for Cohorts A and B, world with the pandemic

Estimates

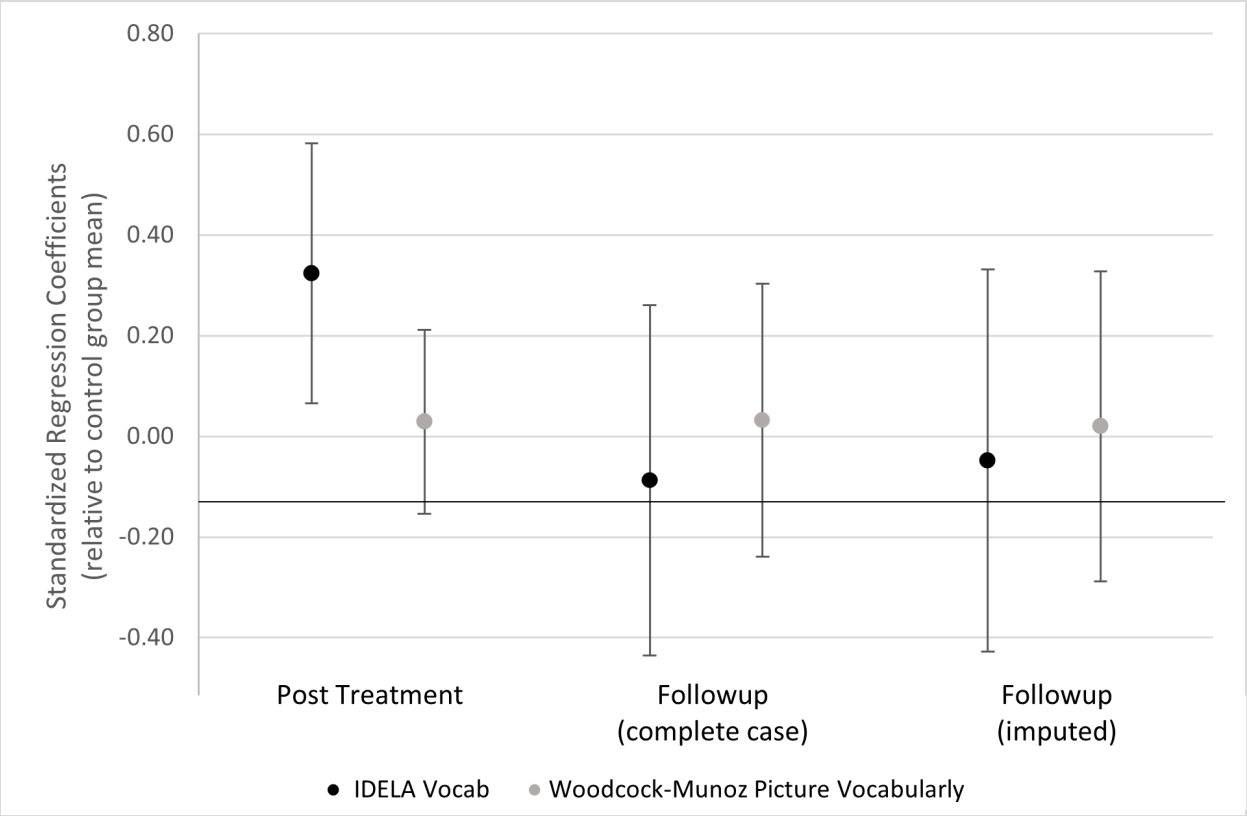
To estimate the impact of the FFT intervention on vocabulary, we conducted three sets of analyses: (1) using 5-month follow-up with Cohort A data only ("complete case"); (2) using 5-month follow-up with cohort A and with Cohort B data imputed;[12] and (3) using end of treatment for both cohorts ("alternative outcome"). In separate work, we have documented results of the first two sets of analysis using an extended set of outcomes (Authors, 2021). Across our three set 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.[13]

Figure 3 displays the estimates from the three strategies described above. Overall, we found positive ITT estimates for the FFT intervention for vocabulary for 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 for 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 = 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.

[12] 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).

[13] 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.

Applying the decision-making framework

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

- **Estimands: Measurement/construct validity.** The pre-pandemic estimands of interest for this study (Estimands 1 and 2) are well-defined. We assessed children using a standardized measure that is commonly used in RCTs with a language and literacy focus, as well as a non-standard assessment that has been used for evaluations in the international context.

  For the 5-month follow-up outcomes for Cohort B, we must also make assumptions about a counterfactual data collection world. We know that remote learning has widely varying effects on student learning, especially for Latino communities (Chen & Krieger, 2021; Sehra, Fundin, Lavery, & Baker, 2020). Thus, we do not know how Cohort B would have performed on the vocabulary assessments if we had tested the children remotely during the early stages of COVID closures in schools (April 2020).

- **Estimands: External validity and policy relevance.** The impact for the end-of-treatment outcome using data from Cohorts A and B provides information of 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 as well — at least in principle. Reasoning about a counterfactual data collection world, however, complicates this. A world with the pandemic (but where we can collect data) does not seem particularly useful for informing future policy decisions in this application, as we view Spring 2020 as anomalous. 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 pre-test) as well as data on all child-level assessments collected at pre-test (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:* For a world ***without*** the pandemic, 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. For a world ***with*** the pandemic, this is an unrealistic assumption. 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 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 prior to 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).

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,[14] and were thus eligible to enroll in school-based pre-K as a four-year-old during the 2010-11 through 2015-16 school years (see cohorts in Figure 4). Students in Cohorts 1-5 completed third grade prior to 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.[15]

[14] Plus those who enrolled in CPS for pre-K, but did not continue into CPS Kindergarten.
[15] MAP is produced by the Northwest Evaluation Association (NWEA), which markets standardized assessment used in all 50 states.

Figure 4. Cohorts Included in Analyses, Before and After Chicago's Pre-K Access, Application, and Enrollment Policy Changes.

	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.

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

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)

3a. 3rd grade scores for Cohorts 4-6 (relative to Cohorts 1-3), world without the pandemic

3b. 3rd grade scores for Cohorts 4-6 (relative to Cohorts 1-3), world with the pandemic

Estimates

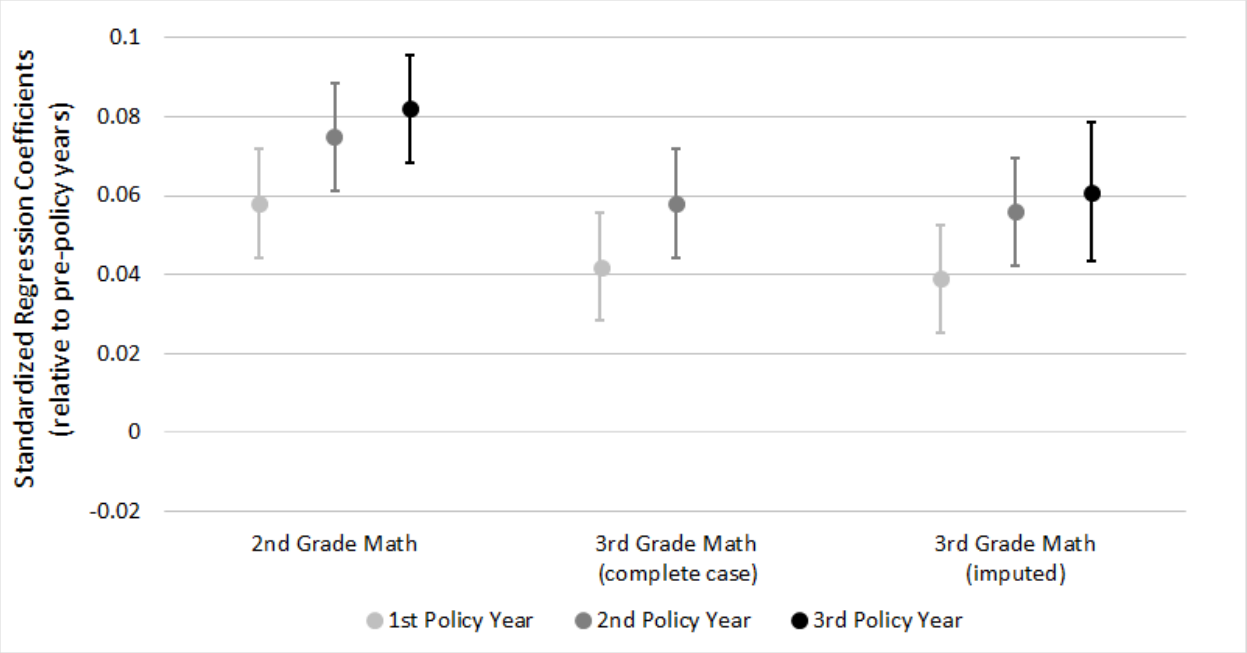
As in the FFT study, we consider three strategies for estimating the association between math scores and the policy change: (1) using third grade math scores for the first five cohorts only ("complete case"); (2) imputing the missing third math scores for the final cohort; and (3) using second grade 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 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.[16] 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

[16] 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.

outcome), which is 75% of the original sample.[17] 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.

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.

Applying the decision-making framework

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

- **Estimands: Measurement/construct validity.** The estimands that are the focus of both the alternative outcome and complete case analysis approaches are well-defined outcomes measured by assessments that we know to be valid and reliable in the context in which they were used in this study (i.e., routine use in pre-COVID, in-person public

[17] 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.

school settings).

For the third-grade outcomes for cohort 6, we must also make assumptions about a counterfactual data collection world. As with the previous case study, it is difficult to imagine a counterfactual world with the pandemic in which data collection still takes place in the context of Chicago preschools.

- **Estimands: 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 2013-14 when Cohort 4 was eligible to enroll, compared to 40% in 2015-16 when Cohort 6 was eligible. Since then, Chicago has continued to expand full-day pre-k as it works towards 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 with the pandemic (but where we can collect data) does not seem particularly useful for informing future policy decisions in this application, as we view Spring 2020 as anomalous. A world without the pandemic is therefore 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 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).

- **Estimation: 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.

- **Estimation: Missing data assumptions.**[18] 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, 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 trade-offs 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 at 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

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

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 (Gelbach & Robinson, 2018). This strategy accordingly requires revisiting and revising pre-registration plans *before* analysis.

Just as empirical education researchers have benefitted from other best practice guides (e.g., Bloom, 2012; Calonico et al., 2017; Duflo, Glennerster, & Kremer, 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 2020-21 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.

References

Bloom, H. S. (2012). Modern regression discontinuity analysis. *Journal of Research on Educational Effectiveness*, 5(1), 43-82.

Boyer, M. (2021). Focus, Fix, Fit: Understanding the Meaning of 2021 Test Scores: Finding a Path Forward with Existing Tools and Procedures. National Center for the Improvement of Educational Assessment. Available at: https://www.nciea.org/blog/state-testing/focus-fix-fit-understanding-meaning-2021-test-scores .

Burbio (2020-2021). K-12 school opening tracker. https://cai.burbio.com/school-opening-tracker/.

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.

Calonico, S., Cattaneo, M. D., Farrell, M. H., & Titiunik, R. (2017). rdrobust: Software for regression-discontinuity designs. *The Stata Journal*, 17(2), 372-404.

Claessens, A., & Engel, M. (2013). How important is where you start? Early mathematics knowledge and later school success. *Teachers College Record*, 115(6), 1-29.

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), 1-12.

Dahabreh, I. J., Robertson, S. E., Tchetgen, E. J., Stuart, E. A., & Hernán, M. A. (2019). Generalizing causal inferences from individuals in randomized trials to all trial‐eligible individuals. *Biometrics*, *75*(2), 685-694.

Dahabreh, I. J., Robertson, S. E., Steingrimsson, J. A., Stuart, E. A., & Hernan, M. A. (2020). Extending inferences from a randomized trial to a new target population. *Statistics in medicine*, *39*(14), 1999-2014.

D'Amour, A., Ding, P., Feller, A., Lei, L., & Sekhon, J. (In press). Overlap in observational studies with high-dimensional covariates. *Journal of Econometrics*.

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.* Chicago, IL: UChicago Consortium on School Research, NORC at the University of Chicago, and Start Early.

Gehlbach, H., & Robinson, C. D. (2018). Mitigating illusory results through preregistration in education. *Journal of Research on Educational Effectiveness*, 11(2), 296-315.

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., & Rubin, D. B. (2013). *Bayesian data analysis*. Boca Raton, FL: CRC press.

Graham, J. W. (2009). Missing data analysis: Making it work in the real world. *Annual review of psychology*, *60*, 549-576.

Hedges, L., & Tipton, E. (2020). *Addressing the Challenges to Educational Research Posed by Covid-19*. Chicago, IL: 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.

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.

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.

Logan, J. (2020). "Missing School-Based Data due to COVID-19: Some Guidelines," working paper, available at: 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., & Rolla, A. (2011). Aftershocks of Chile's earthquake for an ongoing, large-scale experimental evaluation. *Evaluation review*, *35*(2), 103-117.

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

National Academy of Education. (2021). Educational assessments in the COVID-19 era and beyond. Washington, DC.

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*, 225-247.

Tipton, E., & Olsen, R. B. (2018). A review of statistical methods for generalizing from evaluations of educational interventions. *Educational Researcher*, *47*(8), 516-524.

Van Lancker, K., Tarima, S., Bartlett, J., Bauer, M., Bharani-Dharan, B., Bretz, F., ... & Cro, S. (2022). 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 preprint arXiv:2202.03531*.

Von Hippel, P. T. (2007). Regression with missing Ys: an improved strategy for analyzing multiply imputed data. *Sociological Methodology*, 37(1), 83-117.

Weiland, C., Greenberg, E., Bassok, D., Markowitz, A., Guerrero Rosada, P. … & 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. Ann Arbor, MI and DC: 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 (2014). Assessing Attrition Bias. US Department of Education, Institute of Education Sciences. https://ies.ed.gov/ncee/wwc/Document/243

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

White, I. R., Royston, P., & Wood, A. M. (2011). Multiple imputation using chained equations: issues and guidance for practice. *Statistics in medicine*, 30(4), 377-399.

Supplementary Materials

How should we analyze school-randomized RCTs in which entire districts are missing post-test scores?

The technical literature on missing data is vast and can be contradictory for applied researchers. We highlight three important dimensions of this literature that are particularly relevant to decisions about how to handle data missing due to COVID-19:

- *What variables are missing?* Are observations (1) missing baseline covariates or pre-test scores; (2) missing outcomes or post-test scores; or (3) missing both?

- *Who is missing these variables?* In the context of school-level randomization (or other cluster-level intervention): (1) only some students in each school are missing these variables; or (2) entire schools (or coarser units) are missing these variables?

- *Why are these variables missing?* Does missingness depend on: (1) both unobserved and observed factors ("Missing Not At Random"); (2) observed factors only ("Missing At Random"); or (3) neither unobserved or observed factors ("Missing Completely At Random")? [see, for example, Little and Rubin, 2020]

Unhelpfully for practitioners, recommendations vary across these dimensions. Rather than focus on general advice, our goal here is to consider best practices for the specific missing data scenario of COVID-related missing outcomes.[19] Thus, we initially explore a close analogue: "how should we analyze a school-randomized RCT (with randomization blocked by district) in which entire districts are missing post-test scores?" In this stylized example: (1) only outcome variables are missing; (2) these variables are missing for whole schools or groups; and (3) missingness does *not* depend on unobserved factors (i.e., outcomes are MAR or MCAR).

Running Example Appendix 0

We begin with an idealized school-level randomized trial of a whole-school intervention in two school districts, Districts A and B, that were actively recruited to take part in this study. Figure 0 gives a schematic of this setup. These districts are the same size, but differ in terms of students enrolled: District A has 25% Title I schools and District B has 75% Title I schools. Schools in both districts are enrolled at baseline and pre-test scores are collected for all students. Schools are then randomized to treatment separately within each district. Post-test data collection proceeds on time for District A, but is canceled in District B due to an unfortunate software glitch that only affected districts starting with B. No other variables are missing.

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

Figure 0: Schematic of missing data structure

	Time 0	Time 1
District A [25% Title I]	Pre-Test ✓	Post-Test ✓
District B [75% Title I]	Pre-Test ✓	Post-Test ✗

In this idealized study, the *overall* attrition rate, the rate of missing outcome data for the entire sample, is 50% (missing post-test scores in District B). However, the *differential* attrition rate, the difference in rates of missing outcome data between treatment and control schools, 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.

How should the study team proceed with the analysis?

What's the estimand? The first step is to define the quantity of interest. In this initial example, there are two primary options:

- The impact on post-test for Districts A and B together, which was the original estimand of interest; or
- The impact on post-test for District A alone, which would "move the goalposts" from the original study.

Importantly, we are not particularly worried about fundamental measurement or construct validity issues here, and believe that both estimands are well-defined. For instance, we can easily imagine fixing the software glitch and collecting post-test data for District B at about the same time as originally planned, and we do not imagine that such a glitch would affect the underlying quantity. As we discuss below, these become major challenges with pandemic-related missingness.

What's the estimator? We consider the two most common approaches here:

- **Complete case analysis** (also known as "listwise deletion"). Estimate the impact using District A only. This could target either (1) the impact for District A alone; or (2) the impact for both Districts A and B together.

- **Adjust for missing outcomes.** Adjust schools with observed outcomes (District A) to have similar baseline covariates to the overall sample (Districts A and B together), such

as by re-weighting or imputation. This targets the impact for both Districts A and B together.

Due to the simple structure in this stylized example, we can then compare the form of the resulting estimates directly.

Complete Case Analysis

In our initial running example, the complete case estimate is simply the estimated impact for District A. To aid comparison below, we can equivalently write this overall estimated effect as a weighted average of the subgroup estimates for Title I and non-Title I schools in District A:

$$\widehat{ATE}_{A} = \frac{1}{4}\widehat{ATE}_{A,\,Title\ I} + \frac{3}{4}\widehat{ATE}_{A,\,Non-Title\ I},$$

where the weights come from the share of Title I schools in District A.

An important but subtle point is that schools with missing post-test scores (District B in this example) *provide no information whatsoever* about the intervention's impact (see, e.g., von Hippel, 2007).[20] So, in a certain sense, the estimate in District A is the best we can do without additional structure. Or, as Allison emphatically concludes: "Among conventional methods for handling missing data, listwise deletion is the least problematic" (2002, p. 84). Partly as a result, the *What Works Clearinghouse* guidelines list complete case analysis alongside weighting and imputation as a viable analysis strategy in a wide range of missing data settings.

A common critique of complete case analysis is that units (schools) with missing outcomes might systematically differ from units with observed outcomes. Indeed, this is the main reason why complete case analysis has such an unfavorable reputation in the literature on missing data (see WWC 2020; Puma et al., 2009).[21] In our initial example, we know that District A is much more affluent than District B, and so we expect the true impacts to differ between districts.

As we note above, we can resolve this by either shifting the estimand or by assuming the problem away:

- **Move the goalposts: District A alone.** Our preferred approach is to "move the goalposts" and restrict our attention to the impact for District A alone. This is a

[20] This is the case when all covariates are observed and we are only missing outcomes. As von Hippel writes: "Cases with imputed Y quite literally contain no information about the regression of Y on [treatment]" (2007, p. 88).

[21] For example, the *What Works Clearinghouse* writes: "Many researchers have recommended against using complete case analysis to address missing data." Puma et al. (2009; p. 64) note that complete case analysis is "often criticized."

well-defined quantity of interest and is arguably the simplest option. This is not always appropriate, as the set of units with observed outcomes might not correspond to a well-defined group or population. But in the case where whole districts are missing outcomes, this seems like a reasonable option.

- **Assume the problem away: Districts A and B together.** The alternative is to continue to target the initial estimand of interest. We can do so via the *Missing Completely at Random (MCAR)* assumption, which states that outcome missingness is unrelated to both observed and unobserved factors. Under this assumption, the true impact for District A therefore equals the true impact for District B (and therefore equals the pooled impact for cohorts A and B together). This seems highly unlikely in this initial example, since Title I status is often an important moderator in practice. We largely advise against this approach.

Confusingly, the estimator is the same in both cases, even though the second approach requires much stronger assumptions. This highlights the importance of being clear about the target quantity.

Missing Data Adjustment

The main alternative to complete case analysis is to use statistical methods to adjust for missing outcomes (see WWC Standards Table II.6). To simplify exposition, we focus on *nonresponse weighting,* which re-weights the data using the estimated nonresponse probability, typically re-weighting the observed data to have the same distribution of baseline covariates as the full sample. The main alternative is *multiple imputation*, which uses an outcome model (e.g., linear regression) estimated on the observed data to predict what the missing outcomes would have been. While imputation might seem fundamentally different from nonresponse weighting, they are closely linked and both seek to adjust the observed sample to mirror the overall sample. We therefore do not focus on implementation details here and instead refer interested readers to Graham (2009), White et al. (2011), and the WWC Standards.[22]

In the context of our initial example, nonresponse weighting is straightforward because we observe a single, binary covariate, Title I status. Thus, the overall estimate is again a weighted average of the subgroup estimates, using weights that match the overall distribution of Title I schools:

$$\widehat{ATE}_{Adjusted} = \tfrac{1}{2}\widehat{ATE}_{A,\,Title\ I} + \tfrac{1}{2}\widehat{ATE}_{A,\,Non-Title\ I},$$

[22] While imputation might seem fundamentally different from nonresponse weighting, they are closely linked and both seek to adjust the observed sample to mirror the overall sample. Thus, the specific choice of adjusting method is typically less important than the “inputs” into the approach, such as the choice of which covariates to include in the imputation model.

where 50% Title I Schools (25% in District A + 75% in District B). Equivalently, we can think of nonresponse weighting as first *predicting* the impact for District B, and then taking the unweighted average of the impacts for Districts A and B:

$$\widehat{ATE}^{*}_{B} = \frac{3}{4}\widehat{ATE}_{A,\,Title\ I} + \frac{1}{4}\widehat{ATE}_{A,\,Non-Title\ I}$$

$$\widehat{ATE}_{Adjusted} = \frac{1}{2}\widehat{ATE}_{A} + \frac{1}{2}\widehat{ATE}^{*}_{B}.$$

As we note above, schools with missing post-test scores (District B in this example) provide no information whatsoever about the intervention's impact. Thus, nonresponse weighting is equivalent to generalizing the impact estimate from District A to District B, and then taking the average of the District A estimate and the (extrapolated) District B estimate.

Power. In this simple example, the standard error for the adjusted estimate will tend to be larger than the standard error for cohort A alone. To see this, note that, in cohort A, there are three times more non-Title I schools than Title I schools, and so we expect $se\left(\widehat{ATE}_{A,\,Title\ I}\right) > se\left(\widehat{ATE}_{A,\,Non-Title\ I}\right)$. The adjusted estimator puts greater weight on the less-precise subgroup estimate, thus increasing the overall standard error. It is generally the case that estimates that adjust for missing outcomes have larger standard errors than the corresponding complete case estimates (see, e.g., Little, 1992; von Hippel, 2007).

Assumptions. Generalizing the estimated effect from District A to District B assumes that outcomes are *Missing At Random (MAR);* that is, missingness only depends on observed (baseline) covariates and treatment assignment --- and not on unobserved factors. In the context of our initial example, the MAR assumption implies that the *subgroup* treatment effects are the same for Districts A and B. Equivalently, this approach assumes that we observe all possible treatment effect moderators.[23] That is, the average impact for Title I schools is the same across districts and the average impact for non-Title I schools is the same across districts. See, e.g., Dahabreh et al. (2019; 2020) and Egami & Hartman (2020) for closely related technical discussion.

An important practical question is why researchers should bother with complex adjustment methods when District B has no information whatsoever about the intervention's impact (and also adds noise to the estimate). The most prominent answer is that the initial study population of Districts A and B together is a more important estimand than District A alone. This is especially true when the schools or districts are themselves randomly sampled from a well-defined population. It is worth quoting Puma et al. (2009) at length here:

[23] There is extensive discussion of this point in the literature on generalizability and transportability. This version is closest to the assumption of "mean generalizability of treatment effects," though this is sometimes framed in terms of an "exclusion restriction." See Egami & Hartman (2021) for a thoughtful discussion.

> In many RCTs in education, the schools in the sample constitute a sample of convenience: they are not selected randomly and thus cannot formally be considered representative of any larger population [...]. Therefore, some analysts may argue that if the study is not designed to produce externally valid estimates, we should be less concerned about missing data problems that make the analysis sample "less representative." However, some RCTs in education do in fact select schools or sites randomly. Furthermore, in those RCTs that select a nonrandom sample of convenience—schools willing to participate in the study—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.

While there are other justifications for missing data adjustment here, we view these as less compelling in the context of this very specific missing data pattern.[24]

Adjusting for Missingness: WWC Guidelines

Once researchers have assessed the underlying assumptions and chosen an appropriate statistical adjustment method, the next step is to assess whether the resulting estimates are still reliable. The goal of the WWC recommendations is to limit possible bias due to missing data adjustment to be below 0.05 standard deviations in typical settings, with specific limits guided by plausible parameters from existing education RCTs (WWC, 2014).

There are two key quantities for assessing the possible bias due to attrition:[25] (1) the *overall* attrition rate, or the rate of missing outcome data for the entire sample; and (2) the *differential* attrition rate, or the difference in rates of missing outcome data between treatment and control groups. Following the WWC guidance, the dangers from differential attrition are typically much larger than from overall attrition. For example, an individually-randomized trial that is missing

[24] For instance, Logan (2020) writes: "Even when data are MCAR there are some disadvantages of dropping incomplete cases. Primarily, dropping cases results in a decrease in statistical power to detect effects, and this can be avoided through modern missing data methods." While this will hold in cases with more complex missingness patterns, in the special case with only missing outcomes, the only way to increase statistical power is by incorporating additional data (e.g., intermediate outcomes) or by introducing stronger parametric modeling assumptions. Little and Rubin (2020) 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 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."

[25] The WWC guidelines also require researchers to assess baseline equivalence between intervention and comparison groups in the analytic sample used. This issue is less concerning in our context because we focus on scenarios in which we have baseline data and later lose entire cohorts in follow up. Nonetheless, examining baseline differences will be a useful check on the assumptions we discuss below.

half of all outcomes (overall attrition = 50%) but with differential attrition less than 1 percentage point is still considered “low attrition.” If the combination of overall and differential attrition exceed the WWC thresholds, which vary based on the overall attrition rate, the study is considered “high attrition” (see WWC Sec II). Low attrition studies that adjust for missing outcomes can meet WWC standards “without reservations”; high attrition studies can instead meet standards only “with reservations.” Importantly, WWC standards have an exception for *acts of nature*, though it is not yet clear if that standard applies here.[26]

If these guidelines have been met, researchers can proceed with a range of analytic methods, including both nonresponse weighting and complete case analysis.[27]

Summary: We return to our initial question, “How should we analyze a school-randomized RCT in which entire districts are missing post-test scores?” While there is no universal answer, we generally recommend restricting the analysis to the fully observed district alone — and being clear that the target of interest has changed.

[26] From the WWC guidelines: “Losing sample members after random assignment because of acts of nature, such as hurricanes or earthquakes, is not considered attrition when the loss is likely to affect intervention and comparison group members in the same manner. However, when sample loss due to an act of nature was concentrated in one group, the loss will be considered attrition.”

[27] “Many researchers have recommended against using complete case analysis to address missing data (for example, Little et al., 2012; Peugh & Enders, 2004). Nevertheless, the WWC considers complete case analysis to be an acceptable approach for addressing missing data because possible bias due to measured factors can be assessed through the attrition standard and WWC’s baseline equivalence requirement.” (WWC, 2020; p. 36)