Northwestern | INSTITUTE FOR POLICY RESEARCH

Working Paper Series

WP-20-47

Addressing the Challenges to Educational Research Posed by COVID-19

Larry Hedges
Board of Trustees Professor of Statistics and Education and Social Policy,
Co-Director of the STEPP Center, and IPR Fellow
Northwestern University

Elizabeth Tipton
Associate Professor of Statistics, Co-Director of the STEPP Center, and IPR Fellow
Northwestern University

Version: October 19, 2020

DRAFT
Please do not quote or distribute without permission.

ABSTRACT

The Covid-19 pandemic has disrupted many aspects of our society, including the conduct of ongoing education research, especially randomized field trials. This paper seeks to identify some of the problems that may arise because of this disruption, which may be different depending on the current stage of the trial. Hedges and Tipton identify some possible responses to the disruption with an emphasis on those that may permit investigators to capitalize on work already done and investments already made. They discuss tradeoffs of strategies such as ways to maintain statistical power of designs that could be compromised or dealing with designs that may have lower power than was initially planned. They also consider more radical changes in focus such as focusing on intervention or instrument development, methodological studies, or the codification of craft knowledge.

The writing of this paper was supported by the Institute of Education Sciences, U.S. Department of Education, through R305B170016 to Northwestern University. The opinions expressed are those of the authors and do not represent views of the Institute or the U.S. Department of Education. The authors also thank Tom Brock, David Francis, Mark Lipsey, Chris Rhoads, Jessaca Spybrook, and several others whose insightful comments on an earlier version of this paper led to dramatic improvements. The remaining failings are solely the fault of the authors.

I. Introduction

To say that the Covid-19 pandemic has caused disruptions in many – perhaps all – facets of life is an understatement. Beginning in the middle of March 2020, schools closed overnight, sending students and teachers home with little guidance. Most of these schools remained online for the remainder of the 2019/2020 academic year, and many – if not most – of these remain online or in a dramatically altered form as the 2020/2021 academic year begins. For those conducting research in schools – particularly randomized trials or longitudinal studies – this disruption has been especially acute, as implementation, training, and funding are all time-dependent. At an even deeper level, however, this disruption brings with it hard questions about the validity of inferences from data collections in which the context has changed radically.

This is certainly not the first of such disruptions to education research as a result of a natural disaster, though it is the broadest reaching. In 2005, Hurricanes Katrina and Rita closed schools in Louisiana, Mississippi, and Alabama, often for months. Since 2005, subsequent hurricanes (e.g., Harvey, Maria, Michael, Sandy) have closed schools in Texas, Puerto Rico, Florida, and New York, and multiple wildfires have closed schools in California and Oregon. While these natural disasters are more localized than the current Covid-19 crisis, for education researchers conducting studies in schools, the results are the same: studies, designed to answer questions and test the efficacy of interventions in one context, suddenly find that this context is no longer available (or, perhaps, even relevant).

What can a researcher – and the broader research community – do when these best-laid plans have to be abandoned? Here we speak specifically to research projects with external funding, the kind that typically involve an established and approved research plan, an obligation to the grantor to use funds appropriately, an investment on both sides, and a timeline for completing the work that has only limited flexibility. Given the scope of the Covid-19 pandemic, we assume that substantial additional funding is extremely unlikely. Furthermore, beyond the obligations to these funders, we assume that the researchers involved care deeply about improving schools, teaching, and student learning, and that this commitment is heightened given the effects of this disaster on schools, communities, and downstream inequality. Thus, it is imperative that the education research community find ways to ensure that the funding commitments already made will produce evidence that can enhance our practical or fundamental knowledge about education in ways that justify this investment.

The purpose of this paper is to provide education researchers dealing with the Covid-19 pandemic – or perhaps any of many other natural disasters – with a set of questions and options, as well as a framework, for thinking about and choosing the best course of action. In framing this paper, we have attempted to draw from the principles embodied in the Institute of Education Sciences (IES) Standards of Excellence in Education Research (SEER). These standards are meant to identify the domains and core questions of quality research in education (for more information, see here: https://ies.ed.gov/seer/index.asp) and we believe that investigators should be guided by the SEER standards as they evaluate their options. Finally, the analyses and suggestions provided here are not exhaustive but are meant instead as an introduction to those issues; researchers can pursue further the ones that are of most importance to them. We offer these ideas in the spirit of collaboration, to help researchers evaluate their options and make their own decisions in light of the difficult circumstances they face in their particular studies.

We begin this paper by discussing the general effects of the Covid-19 pandemic (i.e., the "pandemic") on education in general (Section II) and then turn to the effects on randomized field trials, in particular (Section III). In Section IV we address the important question of whether a study should proceed. Then we introduce four broad approaches researchers might take to address these problems (section V), followed by a handful of additional solutions that may be useful for continuing studies in specific situations (section VI), and a discussion of validity concerns (section VII). These approaches are by no means comprehensive, and so in Section VIII we focus instead on *how* researchers should consider evaluating different options they might have, considering potential threats to validity that result from Covid-19 and then putting this together into a general approach. Our hope is that this paper will provide researchers – and funders – with some new ideas and support them as they think flexibly about how to conduct research in the Covid-19 era.

II. The Chaotic Effects of the Pandemic

Perhaps the most obvious effects of the pandemic in educational research are on how schools are organized, how they deliver instruction, and how they deliver ancillary services that support instruction. Moreover, these changes have been rapid, rarely the result of long term planning, and have themselves been subject to rapid changes as conditions have, and continue to, evolve (e.g., rapid changes in schedules of the school years, changes from remote to in person and then back to remote instruction in schools inadequately prepared for remote instruction to begin with and no certainty in sight). Uncertainty and change have been the hallmarks of this disruption. The disruption has been almost as profound for the research enterprise itself. The institutions that support research are themselves subject to the same kinds of disruption (face to face meetings are impossible, classes and training have to be conducted virtually, procedures and policies are often vague and change rapidly).

The issues facing schools and researchers are not just about procedures and instruction but health of students, school personnel, and research personnel. These are literally life and death issues. While this document is written to focus narrowly on the problems of research in this era, we do not underestimate the magnitude of the other issues facing educators and researchers. Some might characterize this paper as akin to one about golfing during a hurricane (a phrase borrowed from another researcher to describe a much less disruptive situation of understanding administration in the face of "positional instability"). However crazy it may sound we believe there is a place for it.

III. Effects of the Pandemic Depends on Research Stage

A. Conceptual Considerations

In the sections below, we focus on common problems found when implementing a randomized trial. It is important to recognize that there are conceptual considerations that may cause the intervention impact to differ from that anticipated when the trial was planned. The SEER principles encourage researchers to "identify interventions' core components" and document treatment implementation and contrast." Consistent with these principles, it is essential to consider the theory of action of the intervention to evaluate how the changes caused

by the pandemic might influence the treatment impact realized under current conditions (as opposed to those anticipated earlier). Research on variation in treatment effects provides a useful perspective on the kinds of differences that could arise (see Weiss, Bloom, and Brock, 2014).

1. Changes in the Treatment Contrast

It is easy to imagine that changes caused by the pandemic could increase or decrease the treatment contrast. For example, the contrast between an online tutoring intervention and business-as-usual might be substantially reduced when all instruction moves online. Alternatively, the contrast between an online tutoring program and business-as-usual might be increased if business-as-usual becomes online instruction with less individual instruction than in a conventional classroom.

2. Changes to Client Characteristics

The pandemic may change the characteristics of all units found in the intervention or of the units most likely to remain in the study. Here the term 'unit' is intended to include not just students (and their families), but the teachers and school personnel that are part of the implementation of the intervention. It is often believed that unit characteristics moderate impacts of an intervention, so it follows that changes in unit characteristics as a result of the pandemic could change treatment impacts. For example, risk or need often moderates treatment impacts. If the disruption due to the pandemic increases risk (e.g., students or teachers have a greater need for an intervention), that might change the treatment impact. Alternatively, if the disruption due to the pandemic causes high risk units to drop out of the study, that would also change the risk profile of the units in the study. Of course, differential changes to unit characteristics might induce bias in estimated treatment impacts. Such changes would not necessarily *bias* the estimate of the treatment impact if they were uniform across treatment groups and did not cause differential attrition. They would, however, *change* the treatment impact that was estimated.

3. Changes to Context

It is well known that the context in which an intervention is used can moderate treatment impacts (see, e.g., Herbst, 2008). The pandemic has caused massive changes in the context of education. This is true in a myriad of ways in K-12 education, perhaps more profoundly so in special education and in higher education. While the mechanism of some of these changes may be through changes in client characteristics like risk, others might influence through distinct routes (like reduced opportunity costs for pursuing higher education).

B. Considerations Specific to Cohorts and Cases

Randomized field trials are complex endeavors and the temporal sequence of events is not identical in every trial. However, there is a sequence of steps that occur in every trial and it may be useful to consider this sequence to understand how the Covid-19 pandemic has disrupted education research[2]. In general, this sequence includes five stages – recruitment, random

[2] We assume that the intervention, product, or service to be tested (which we will hereafter refer to as the "treatment") has largely been developed before an efficacy or effectiveness trial is even funded.

assignment (and pre-testing[3]), implementation of the intervention (or business-as-usual), collection of data on implementation and mediators, and collection of data on final outcomes. Many studies divide this sequence across multiple years and multiple cohorts.[4] Figure 1 provides three different cases, each with potentially 1-3 cohorts of students.

Figure 1. Research flow and possible timing of the pandemic

Before COVID-19 | After COVID-19

	2018-19	2019/2020	2020/2021	2021/2022	2022-2023
Case 1			Planning	Cohort 1	Cohort 2*
		Planning	Cohort 1	Cohort 2*	Cohort 3*
Case 2	Planning	Cohort 1	Cohort 2*	Cohort 3*	
Case 3	Cohort 1	Cohort 2*	Cohort 3*		

*Note: 'Planning' indicates either that an award was just funded (typically in spring) or that there was a planning year included in the study design. The * indicates that additional cohorts are possible but not required.*

Figure 1 indicates that the effect of the pandemic on a randomized field trial depends both upon the number of cohorts of data collected before March 2020 and upon the stage of the research within the broader 2019/2020 and 2020/2021 cohorts, as the effect of the pandemic on schools is ongoing as we write this. Depending upon the case (Case 1, 2, or 3) the disruption could have differing effects on the resources left in the grant or contract, the cost of conducting further research to complete the research cycle, and the nature of the disruption to the research plan. In this section, we catalog the specific types of problems that likely are occurring in studies depending upon the case. Note, of course, that not all studies will fit nicely into one of these cases exactly.

C. Case 1: Disruption at Study Beginning

In some studies, 2019/2020 was a 'planning' year, in which the focus was on developing and implementing a sampling plan and securing commitments from at least a subset of schools (and perhaps teachers) to participate in the study (as well as other study development, e.g., measurement development). Similarly, in other studies, 2019/2020 was the year in which the grant application was under review, with final decisions and awards made in Spring 2020. In both cases – what we refer to as Case 1 – by the end of 2019/2020, the study was both funded and yet just beginning, with little of the budget spent or data collected. In general, for studies at this stage, we expect three types of problems to have arisen:

[3] Of course, not all studies include pre-testing, and in some such studies that do, this pre-testing occurs after random assignment, while in others before.

[4] While we have characterized studies as having explicit cohorts, it can also be true that treatment implementation occurs at somewhat different times for different treatment units, but not in discrete cohorts. The same issues arise for those studies as for those with explicit cohorts in that some units will have implementation partly or entirely disrupted but others will not.

1. Uncertain or Failed Recruitment of Sample

The recruitment of the sample involves a significant amount of effort in virtually every trial. If the project had not implemented the intervention by the time the pandemic began, much of the time, effort, and money expended on recruitment of a sample that cannot be used was wasted. The best-case scenario for the project going forward is that the intended sample will have to be approached again (essentially re-recruited) at a future time. The more likely scenario is that the sample will have to be substantially augmented or an entirely new sample will have to be recruited. However, it is not clear when schools will be operating in a normal fashion and when, or if, the study might be completed.

2. Inability to Collect Baseline Data

In some designs, baseline data for the following year is collected the previous year, often in the spring. For example, power analyses often assume the use of a student, teacher, or school-average prior-year test score in the outcome model. Given the timing of the pandemic shutdown, it is unlikely that these pre-tests were completed using researcher developed measures. Furthermore, since all 50 states cancelled spring 2020 state tests, even distal measures are not available.

3. Inability to Conduct Training

In some designs, the training for teachers (or other implementors) and research staff will have been disrupted by the pandemic. For example, professional development scheduled for spring or summer 2020 may not have occurred or may have been delivered is less than ideal circumstances that don't allow for a strong test of the underlying theory of change. Some of this training may be possible to be conducted remotely, but other kinds of training may be more difficult to do authentically without access to classrooms.

4. Change to business-as-usual and feasibility / utility of intervention

The theory of change for an intervention typically depends upon an assumed understanding of both current, business-as-usual practice in classrooms and schools and the resources, constraints, and context in which the intervention will be delivered. For some interventions, the pandemic may render the contrast between business-as-usual and the intervention to be negligible (e.g., if all students now learn math online and the intervention was online math), while others may simply not make sense in this new environment (e.g., a school-wide anti-bullying program may be difficult to implement online).

D. Case 2: Disruption Occurred in Middle or End of Study

In other studies, one or more cohorts of schools, teachers, and students were already recruited and engaged in the study when Covid-19 struck. In these studies, pretesting or other baseline measurements on participants (and perhaps even on classrooms) were likely already collected, the intervention was all (or mostly) delivered, and yet for at least some portion of sample, data was not yet collected on mediators or post-tests. For studies at this stage, we expect three types of problems:

5. Inability to Complete Treatment Implementation

In studies in which an intervention was intended to be either be full year (2019/2020) or full semester (spring 2020), the pandemic likely dramatically reduced the intervention dosage. For example, the intervention may have included three units of lessons, yet only two were implemented before March 2020. Even if the program was able to be implemented as students began working from home, doing so may have required changes to the format and implementation of the intervention (e.g., from in class to online).

6. Inability to Collect Mediators or Implementation Data

Most designs involve collecting data on implementation and possible mediating variables. For studies implementing the intervention in 2019/2020 or before, the pandemic may have disrupted collection of some or all of this data. Furthermore, implementation data that was collected before the pandemic may have changed in format from the kind able to be collected afterwards, introducing measurement comparability problems.

7. Inability to Collect Posttest or Follow-up Data

Even if intervention was entirely implemented before the pandemic, it may be impossible to collect the planned posttest or follow-up data. This has the potential to compromise the estimation of treatment impacts.

E. Case 3: Disruptions to Multiple Cohort Studies

Many studies involve recruitment and implementation of an intervention over multiple cohorts. This means that the effect of the pandemic on a study depends upon both the fraction of the total sample included in 2019/2020 and how far along into the study the pandemic occurred. For example, in the fourth row of Figure 1, the study includes both a cohort completed entirely before the pandemic (in 2018-2019) and a cohort after the disruption. This results in three additional considerations.

8. Study Budget Limitations

In most cases, the effect of the pandemic is essentially to reduce the overall budget of the grant moving forward, since money was already spent on recruitment, pre-testing, and training for a study cohort for which it is ultimately impossible to estimate a treatment effect. Depending upon how far along the study was overall, this may mean there is not enough funding left to entirely replace the 'lost' cohort, if that is even possible.

9. Comparability of Data Across Time

In some cases, the one or more cohorts will have been completed before or during 2019/2020, with a subsequent cohort taking place after 2019/2020. In these cases, it is unclear whether it is reasonable to pool treatment effects estimated across these two periods, since many features of business-as-usual, the intervention (e.g., dosage, delivery approach), and even post-tests may have changed as a result of the pandemic.

10. Generalizability Concerns Post-Pandemic

The simplest situation may be the one in which data collection (including post-testing or follow-up testing) was completed for the entire study before the pandemic. Studies in this happy situation face fewer practical problems, but still face the question of external validity: generalizability to the situation of schooling in the post pandemic era.

IV. Should the Study Proceed?

The most consequential decision facing an investigator and the funding agency is whether the study should proceed as planned (or with minimal modifications), proceed in modified form, or not proceed at all. The analysis of this question needs to focus on whether, given the current and future disruptions of the pandemic, it is ***scientifically appropriate*** to proceed with minimal modifications or even substantial modifications. If it is appropriate to proceed, then the decision to do so still depends on whether it is ***feasible*** to proceed. These two decisions are not entirely independent. Modifications may be necessary to make a study feasible, which will affect the decision about whether it is appropriate to proceed with the study, as modified to make it feasible.

A. Is It Scientifically Appropriate to Proceed?

For some studies at some stages of completion, it may be appropriate to proceed with relatively minimal modifications (for example a study which has already collected the data or a study of a context in which the disruption is likely to be minimal, such as a study entirely conducted in an existing online environment). For these studies, the questions of appropriateness will depend on whether the study can still yield a contribution to knowledge given the changed context that now exists and is likely to exist after the pandemic. For other (probably most) studies, modifications, and perhaps very substantial modifications will be required. In this case, the decision about whether it is appropriate to proceed will have to be made almost as if considering the study as a new study. While the investigator will need to satisfy themselves about appropriateness, when modifications are substantial, they will also have to articulate a persuasive argument to satisfy the program officer at the funding agency.

B. Is It Feasible to Proceed?

Even if it is scientifically appropriate to proceed with a modified study, it is still necessary to decide if it is feasible to proceed. The feasibility of proceeding will depend on the details of the modifications that need to be made. Section V of this paper anticipates some of the ways that studies could be modified to address problems created by the pandemic. A careful analysis of these or other modifications that may be used should suggest whether it is feasible to proceed.

C. Is There Enough Scientific Value to Proceed?

Finally, even if the question is scientifically appropriate and – with appropriate changes – the study is feasible, an important question is the degree to which there is adequate return-on-investment (ROI) to proceed. For example, in order for a study to feasibly continue, it may

require new personnel, additional time, changes to the study design, and so on, each of which comes with a cost. As a result of these new costs, the resulting study may have a smaller sample, more limited scope, or result in a less rigorous design. Put another way, the balance between 'scientific value' and 'operational costs' at the studies outset may have changed substantially as a result of the pandemic. Both PIs and funders will need to be involved in these ROI decisions, which are likely to be difficult and involve a great deal of uncertainty.

V. Four Broad Strategies for Addressing Problems

In this section we offer four broad strategies that might be used to address the problems resulting from the pandemic disruption. The choice of strategy likely differs for those in Case 1 versus Case 2 versus Case 3, as well as upon thousands of other decisions particular to a study in hand. We do not pretend to address all of them. For example, safety of students, school staff, and research staff are paramount concerns that will need to be addressed in ways that are responsive to the evolution of the pandemic. Many, perhaps all, of the strategies we offer will not be reasonable or feasible for any given study. We offer them as possibilities, to stimulate thinking. Whatever strategies that are chosen, it is important that, consistent with the SEER principles, that changes should be documented in the registration of the study.

A. Delay Startup of the Study

If the pandemic occurred near the beginning of the study (before recruitment was finalized or pretesting), then it may be feasible to delay the startup of the fieldwork for a year (or possibly longer). Delaying the startup may permit conducting all of the fieldwork at a time when the situation in schools has stabilized. If staff costs during the period in which fieldwork has been postponed can be minimized, this strategy can preserve much of the budget for future years. For projects that are just beginning, this strategy may be worth serious consideration.

Delaying the startup also has risks. It is not entirely clear when the pandemic will end or if education after the pandemic will have changed in ways that require significant changes in the research plan. If it takes more than one year for the situation to stabilize, then it is unclear that this strategy is feasible. It is also unclear if research sites previously committed will remain interested, if project staff will be available after fieldwork begins and if some or all of any training costs already expended will be lost.

B. Address Design Sensitivity Concerns

Problems in recruitment, retention, or having to repeat training, observation, or treatment implementation (under a fixed budget) all lead to the consequence of reduced sample size. For example, if a post-test is simply not possible for the 2019/2020 cohort, results from the study may only include cohorts before or after this year. This results in a smaller sample size – with a smaller number of schools (and possibly students) – than planned. For a given analysis, design sensitivity depends on (individual and cluster) sample sizes, significance level, effectiveness of covariates (if any), and effect size. If the design and analysis plans cannot be changed, the result of having a smaller sample size will be reduced design sensitivity.

1. Increase Design Sensitivity

Before accepting that the design must have reduced sensitivity, it may be wise to consider options that might increase the sensitivity (without changing the significance level). It is usually impossible to increase the effect of an intervention[5].

Design sensitivity may be increased somewhat by adding covariates, even if the analysis plan already includes covariates. These might be covariates added from administrative data. Cluster level covariates are much more important in determining design sensitivity than individual level covariates. Because clusters are usually schools, and school level administrative data are often more accessible than individual level data, the option of adding school level covariates may be quite feasible. However, researchers should keep in mind that adding covariates won't help much unless there is a large relationship between the post-test (outcome variable) and the covariate; when there is not a large relationship, this can result in less precision, since degrees-of-freedom are reduced for each additional covariate in the model.

If new cohorts of the randomized trial will begin after the pandemic, blocking to make the sample of clusters more homogeneous may also be a useful strategy to increase design sensitivity somewhat. The pretreatment variables that could be used to create blocks of clusters may also be available from school-level administrative data.

2. Accept Higher Type I Error

If other strategies to increase sensitivity cannot be used or cannot achieve the desired sensitivity, then an alternative is to increase the level of statistical significance used in hypothesis testing (or creation of confidence intervals) beyond the conventional 0.05. While this is completely statistically valid (significance levels chosen are arbitrary), there is an entrenched scientific preference for the significance level 0.05. It is reasonable to expect resistance to this option. However, there is an increasing effort on the part of statisticians, in conjunction with the move to greater emphasis on effect size, to dethrone the significance level of 0.05. In the current environment of the pandemic, we expect this effort to undermine the rigid adherence to the 0.05 standard will accelerate.

The statistical community and scientific societies such as the American Psychological Association and the American Educational Research Association have called for greater attention to effect size in interpreting the results of scientific studies as a supplement or replacement for interpreting results solely based on statistical significance (AERA, 2006; Wasserstein and Lazar, 2018; Wilkinson, 1999). These recommendations imply that estimates of treatment impacts and their confidence intervals are a more appropriate basis for framing interpretations of study results. Part of the reason for the emphasis on estimation is that statistical significance (or not) is a *qualitative* representation of results, reducing the finding of a study to a yes/no binary representation. Regardless of whether the treatment impact is statistically significant or not, the binary representation of treatment impacts discards a great deal of information about the range of effects that are consistent with the impact findings. While low design sensitivity typically leads

[5] It is not always impossible to increase the effect size of an intervention, for example by increasing the intensity or duration of treatment, but even if this is feasible, whether it would increase the effect size is usually a matter of speculation.

to wide confidence intervals, this approach emphasizes the quantification of what has been learned in a quantitative, not qualitative form.

Finally, we would argue that regardless of design sensitivity, researchers should make efforts to interpret the statistical effect size in practical terms, even if only as simple as where the intervention and comparison group means fall in percentile terms on the distribution of outcome scores (see, e.g., Valentine, Aloe, and Wilson, 2019). In this context it has the added advantage of de-emphasizing statistical significance.

3. Use a Bayesian Approach with Prior Information

A more radical approach in the same vein as emphasizing effect sizes and confidence intervals instead of statistical significance to use a Bayesian approach to reporting results. Bayesian methods incorporate prior distributions for analytic model parameters that are intended to provide a quantitative representation of what is known *a priori* about those parameters. Often these prior distributions are chosen in ways that attempt to reflect ignorance of precise values of the model parameters (so-called *uninformative* priors). The analysis combines the prior information (instantiated in the prior distributions) with information from the data (the so-called likelihood) via Bayes theorem to obtain an estimate of the treatment impact parameter and its uncertainty in the form of a *posterior distribution*. The posterior distribution can be summarized via its mean or a high-density posterior region much like a confidence interval. Although one can estimate the probability that the treatment impact was positive, significance testing is not typically part of standard Bayesian analyses.

Bayesian analyses are not magic. The amount of statistical information in the data determines both the design sensitivity for conventional (frequentist) statistical analyses and the uncertainty of the posterior distribution in Bayesian analyses (unless you have a very strong prior). If a research design has limited information, it will yield both an insensitive hypothesis test (or wide confidence intervals for treatment impacts) and a wide posterior high-density region (unless the prior distribution is very strong). However, because the Bayesian analysis focuses attention on the quantitative representation of what was learned from the study *in light of prior information*, it draws attention away from the binary representation of study results that accompanies significance testing.

One advantage of Bayesian analyses in the context of designs that are less sensitive than might be desired is that they can incorporate prior information that corresponds to greater knowledge than complete ignorance of the values of model parameters. For example, while we may not know the value of treatment impact precisely (that is why the study is justified), it is rare for impact parameter of most interventions tested in randomized field trials to correspond to a Cohen's *d* of bigger than 1.5 or smaller than -1.5. For many interventions, there may be persuasive arguments for narrower bounds, perhaps based on meta-analyses of similar interventions and outcomes. Bayesian analyses can use such information, quantified in the form of *informative* priors for model parameters, to reduce the uncertainty in the posterior distribution of treatment impacts. Informative priors would narrow the range of likely values of the treatment impact (the posterior high-density region), so that the conclusions about treatment impact are more sharply defined.

4. Accept the Low Power of the Design and Report Accordingly

Another alternative is simply to accept that the trial has low statistical power and report it in a conventional way (but certainly including an estimate of the treatment impact and its uncertainty). One of the effects of the pandemic is likely to be an epidemic of low power trials. In the spirit of never letting a good crisis go to waste, the epidemic of low power trials is a chance to better understand and appreciate their role in creating knowledge and debunk the myth of the single definitive study in science. Trials exist within an ecology of scientific research in which even low power trials can lead to knowledge that can improve the design of interventions or trials themselves. This is part of the interplay between empirical findings, methodological development, and theory building mentioned in Shavelson and Towne (2002).

Put another way, if none of the options suggested above seems feasible or desirable, another option is simply to accept the fact that the study will be a low power trial and interpret the results cautiously. When doing so, it is important to emphasize the design sensitivity when describing results. If the results do achieve statistical significance at conventional significance levels (e.g., 0.05), then interpretation is straightforward. If results are not statistically significant, it is important to emphasize that interpretations are made more complex by (unplanned) low design sensitivity. One device that may be helpful is to report the minimum detectable effect size at whatever significance level is used to carry out the significance test for treatment impact (Bloom, 1995). For example, one might say that the estimated treatment effect was $d = 0.15$ with a standard error of 0.1 (or a 95% confidence interval from -0.05 to 0.35) but add that the study did not have the sensitivity that was planned due to disruptions caused by the pandemic. Then add that if the full sample had been able to be achieved, the minimum detectable effect size would have been $d_{min} = 0.15$, but because of disruptions caused by the pandemic, the minimum detectable effect size for the design as realized is $d_{min} = 0.25$, so that strong conclusions about the treatment impact are not warranted. The value of such a trial is that it adds to the evidence base about the impact of the intervention studied.

It is important to note that this problem of 'low powered trials' is already common in the pilot studies often found in development work. A concern in that context – that we echo here – is that a variety of poor statistical practices arise when power is low. For example, there may be selective reporting bias (where only statistically significant effects are reported) and *p*-hacking (where many outcomes and models are considered until statistical significance is achieved). In order to combat these, efforts will be needed to ensure the outcomes and models specified at the beginning of the study remain the primary focus, that publication does not depend on statistical significance, and that treatment effect estimates are provided in a transparent and comprehensive way.

Finally, even when the best practices are used, low power trials do not provide as much information (in both the statistical and practical sense) as do higher power trials. To put it another way, the treatment impact estimates from low power studies are less precise than those from higher power trials. For this reason, they receive less numerical weight in meta-analyses and should receive less conceptual weight in any interpretation of findings from a group of studies.

C. Change the Measurement Design or Pool Meta-Analytically

If the intervention was successfully implemented in some form in 2019/2020 before the pandemic, the major problem posed may be the collection of the outcome or follow-up outcome measurements that were intended. Given the timing of the pandemic, we anticipate that this is a common problem, resulting in what many refer to as "pre-test only RCTs". The most obvious concern here is if it is even possible to collect a delayed post-test of some sort. But even once collected, there are two additional concerns. First, if the post-test differs across units in the study, there is a concern with how to link and equate these tests. Second, if this test differs across cohorts, there is a question, too, of how to analytically pool the results, since outcomes differ. We emphasize that that the problem is not merely to find "some outcome that can be measured," but, consistent with the SEER principles, to find a meaningful outcome that can be measured and (also consistent with the SEER principles) to be aware of and document that outcome and its limitations. In this section we begin by discussing these equating and pooling problems and then explore possible alternatives to the intended post-test, as well as their limitations.

Linking and Equating Tests

If outcome measures are available on all subjects, but they are measured by different tests of the same construct, it may be possible to represent all of the test scores on the same numerical scale by linking or equating the different tests. There is a huge amount of research on test equating and linking (see, e.g., Kolan and Brennan, 2004). Scholarship in this area recognizes several distinctions about the degree of comparability that an equating exercise may guarantee and the inferences it might be appropriate to draw from individual scores or population summaries such as means (see, e.g., Mislevy, 1992; Feuer, 1999). Statistical methods used to link or equate tests range from linear equating (using a linear regression equation to map test score on one scale to another scale), equipercentile equating (defining the score on one test to be the same as a score on the other test if both scores represent the same *percentiles* in a population distribution), and methods that rely on item response theory. The most sophisticated approaches to test equating can be technically complex, require a great deal of raw data (including item scores on tests to be equated), and are likely to be infeasible in the context of a single randomized trial. However, they may be possible by pooling data across many different studies; doing so may require researchers to collaborate with others in their domain.

Only the most sophisticated methods (involving item response theory) are suitable for equating assessments in which the interpretation of individual scores have significant consequences for those individuals or institutions (so-called high stakes assessments). However, there is evidence that simpler equating methods, such as linear equating, can be reasonably accurate for making inferences about comparing population groups (Phillips, et al., 2014). Treatment impact estimates in randomized trials are comparisons of group means. This suggests that linear equating may be adequate for creating a common metric from different tests (*as long as they measure the same construct*) to evaluate treatment impact. However linear equating works best for scores near the means of the groups used to create the equating, so some caution should be exercised in relying on linear equating with groups whose scores are far from the mean (e.g., gifted students or students with disabilities).

In conducting such equating, it is important to consider the reliability of the tests to be linked (it should be high for both tests and as similar as possible) and the correlation between the

tests (it should be very high, considered in light of the reliabilities, which limit the *possible* correlation). In some cases, this correlation may be available (e.g., between a proximal and distal measure available in one sample pre-pandemic), while in others it is not (if two completely different tests are used before and after the pandemic).

Finally, it is wise to obtain the advice of a competent psychometrician in this process. The validity of any linking of tests depends on the conceptual analysis of the tests themselves: Examination not only of the descriptions and purpose of the tests, but of the test specification, technical data, and the items themselves. No statistical magic can equate two measures in a meaningful way unless they measure the same underlying construct.

Pooling Data Across Different Outcomes

If there are only a few (2 or 3) different measures used, and if each measure is used by some clusters in both the intervention and comparison conditions, then another possibility is to conduct a separate analysis for each group of clusters (sharing the same outcome measure) and combine them via meta-analysis to create the summary treatment impact estimate for the trial. That is, each group of clusters is a small (underpowered) trial that yields an impact estimate and its standard error. The meta-analysis combines those impact estimates and permits testing the combined estimate with higher power than that of the individual small trials. The remarkable fact is that this process yields statistical tests for treatment impact that are only slightly less sensitive than would be obtained if all the outcome data could be analyzed together in a conventional analysis. The methods of meta-analysis are well established in statistics (see, e.g., Hedges and Olkin, 1985) and there are many comprehensive references available (see, e.g., Borenstein, et al., 2009: Cooper, Hedges, and Valentine, 2019).

Note that some scientists seem to think that meta-analysis is only appropriate if a large number of "studies" are to be combined. There is no statistical basis for this idea. Meta-analysis can be used with as few as 2 "studies," and in fact it has been adopted as the method for interpreting multiple studies in What Works Clearinghouse evidence reports, where there are frequently only 2 – 3 studies to be combined. Note also that the meta-analytic approach shares a logic that is similar to that which justifies the use of linear equating. If the outcome measures are linearly equitable, the effect size does not depend on the choice of outcome measures. This also implies that the considerations used in deciding whether linear equating of measures is appropriate also apply in deciding whether meta-analysis is appropriate. Careful examination of the test specification, technical data, and the items themselves is required to assure the validity of the meta-analytic conclusions as a representation of overall study results.

Possibilities for Post-Tests

Equating and pooling, of course, require that it is possible to measure some sort of post-test for all units in the 2019/2020 study. Here we discuss three possible approaches, noting that they could be used separately or in combination with one another.

1. Proxy Dependent Variable

The simplest strategy (and yet possibly the hardest to implement) may be to find a common proxy dependent variable, such as a test (or a subscale of a test) that is collected as part of

administrative data collection (if such data exist and is uncompromised by the consequences of the pandemic). While this strategy is appealing, there are some issues that must be evaluated. One is the degree of alignment of the proxy measure with the intervention and the intended outcome. There is a predictable mathematical relation between alignment and effect size under a model that says the proxy measure includes the dimension on which the intervention had an impact, plus other irrelevant components on which the intervention has no effect. Under this model, if δ_I is the treatment effect size on the intended outcome variable and ρ_{IP} is the correlation between the intended and proxy measures, then the effect size on the proxy dependent variable δ_P is related to δ_I via

$$\delta_P = \delta_I \frac{\rho_{IP}}{\sqrt{\rho_{II'}}} \quad (1)$$

(see Hedges, 1981). The symbol $\rho_{II'}$ in equation (1) refers to the reliability of the intended measure. The square root of the reliability serves as an upper bound for the possible value of ρ_{IP}. Notice that the less aligned the proxy variable is to the intended outcome (the smaller the value of ρ_{IP} for given values of the reliability of the two measures), the greater the attenuation of the effect size caused by using the proxy outcome. The attenuation of effect size has an impact on design sensitivity (it decreases the statistical power and widens the confidence interval associated with estimates of treatment impacts).

2. Use a Formative Assessment

Another option might be tests that are collected for formative or benchmarking purposes as the intervention is implemented. This might provide an outcome that is better aligned with the intervention as it was realized in the study than would the intended outcome (because presumably some of the intervention was not implemented). Of course, this may only be possible if such tests are given to both intervention and comparison groups.

Tests used to benchmark progress during an intervention period might have the advantage of being more aligned with the intervention than the intended outcome measure, but if this strategy is chosen, it is worth considering whether this proxy outcome is *over-aligned* with the intervention. Over-alignment would occur if the proxy measure focuses too specifically on the material used in the intervention, so that it becomes a virtual "manipulation check." For example, over-alignment would occur if the proxy measure for a vocabulary building intervention focused specifically on words explicitly taught or if a problem-solving intervention focused specifically on problems that have the same format and structure as those in the instruction.

There are other considerations. It may be that the formative test is not over-aligned, but it may be less reliable than the intended outcome (which presumably is highly reliable). Unreliability in the proxy measure ensures that the correlation between observed scores on the proxy measure and observed scores on the intended measure will not be too high (it will with certainty be less than 1). In order to make the impact of reliability on effect size explicit, we can rewrite equation (1) in terms of the true score correlation ($\rho_{T_I T_P}$), the reliability of the intended measure ($\rho_{II'}$) and the reliability of the proxy measure ($\rho_{PP'}$) as follows:

$$\delta_P = \delta_I * \rho_{T_I T_P} \sqrt{\frac{\rho_{PP'}}{\rho_{II'}}} \ . \quad (2)$$

It is clear from equation (2) that even if the intended and proxy outcome measures are perfectly aligned (that is, even if $\rho_{T_I T_P} = 1$), the effect size will be attenuated if the proxy measure is less reliable than the intended measure. Also note that the reliabilities place a mathematical constraint on the possible values of ρ_{IP} used in equation (1). In particular

$$\rho_{IP} \leq \sqrt{\rho'_{PP}\rho'_{II}}. \quad (3).$$

The impact of attenuation of effect size on statistical power is illustrated in Table 1. The table gives values of statistical power for a few examples of cluster randomized trial designs in terms of the alignment parameter (ρ_{IP}) and the proxy measure reliability ($\rho_{PP'}$) when $\rho_{II'} = 1$ (thus one can consider the values as relative reliability of the proxy outcome compared to the intended outcome). Recall that, because of the constraint (3) not all values of ρ_{IP} are possible given the value of $\rho_{PP'}$. For example, when $\rho_{PP'} = 0.8$, then $\rho_{IP} \leq 0.89$ and when $\rho_{PP'} = 0.6$, then $\rho_{IP} \leq 0.77$ (as reflected in Table 1). The first 6 rows of Table 1 assume perfect reliability and so reflect only at the impact of misalignment. The next six rows set ρ_{IP} to its maximum value given the reliabilities and so can be thought of as looking at the impact of unreliability of the proxy given perfect alignment. These entries all assume that the observed score correlation, ρ_{IP}, is the true score correlation times the square root of the reliability.

It is clear from the table that a small amount of misalignment has a moderate effect on statistical power, but more substantial misalignment can seriously erode power. In this example, when both proxy and intended outcomes are perfectly reliable, choosing a proxy outcome that is correlated 0.8 with the intended outcome reduces statistical power from 0.78 to 0.58, but choosing a proxy outcome that is only correlated 0.5 with the intended outcome reduces the power to less than 0.30. In this example, when the proxy outcome true scores are perfectly correlated with the intended outcome true scores, choosing a proxy outcome that is 80% as reliable as the intended outcome only reduces power from 0.79 to 0.68, but choosing a proxy outcome that is 50% as reliable as the intended outcome reduces power to less than 0.5. The joint effects of misalignment and a less reliable proxy (not shown in the table) are more profound than either that of misalignment or unreliability alone.

To address the probable effects of measurement error and misalignment, researchers might consider using the formulas given here to report the effects that might have been obtained using the intended measurements as a supplementary way of describing results. This would be analogous to reporting disattenuated correlations in other measurement contexts. We would suggest that any such results be explicitly labeled as reported for descriptive purposes only, that they dependent heavily on the assumptions involved (which should be stated explicitly), and they are not intended for hypothesis testing purposes (e.g., with no statistical significance attributed to them).

Table 1. The effects of measurement misalignment and measurement unreliability of statistical power

Number of Clusters	Cluster Size	Total N	Effect Size	$\rho_{PP'}$	$\rho_{II'}$	ρ_{IP}[a]	Power
50	100	5000	0.25	1.0	1.0	1.0	0.78
50	100	5000	0.25	1.0	1.0	0.8	0.58
50	100	5000	0.25	1.0	1.0	0.6	0.37
50	100	5000	0.25	1.0	1.0	0.5	0.27
50	100	5000	0.25	1.0	1.0	0.4	0.19
50	100	5000	0.25	1.0	1.0	0.3	0.13
50	100	5000	0.25	1.0	1.0	1.0	0.78
50	100	5000	0.25	0.9	1.0	0.95	0.73
50	100	5000	0.25	0.8	1.0	0.89	0.68
50	100	5000	0.25	0.7	1.0	0.84	0.63
50	100	5000	0.25	0.6	1.0	0.77	0.55
50	100	5000	0.25	0.5	1.0	0.71	0.49

Note: The computations in this table assume a cluster randomized design, that the intraclass correlation is $\rho = 0.2$, and that the significance level is $\alpha = 0.05$
a. Given $\rho_{PP'} < 1.0$, this maximum possible value of ρ_{IP} is less than 1.0

3. Collect a Post-Test from a Subset of Units

If no suitable proxy variables for the outcome are available, it might be feasible to collect the intended outcome or a close proxy for it from a smaller number of individuals in the trial. While larger sample sizes generally lead to greater design sensitivity, the relation between sample size and design sensitivity is more complex in multi-level situations such as cluster randomized trials. While sensitivity depends strongly on the number of clusters, it depends less so on the number of individuals (assuming the number of clusters is fixed). Moreover, there is a point of diminishing returns where adding more individuals within a given cluster has little impact on design sensitivity. This point of diminishing returns occurs at a sample size that most scientists find surprisingly small. Thus, a trial that collects 10 outcome observations per cluster may have almost as much statistical power as one that collects 25 observations per cluster. Collecting a small number of (ideally randomly chosen) outcome observations from each cluster by intensive means (using online or even specially arranged in-person testing) may make it possible to salvage a trial which would otherwise have no outcome measurements. The general principle is to invest heavily in collecting a small number of observations from each cluster rather than the larger number that may have been intended. However, because design sensitivity depends strongly on the number of clusters, it is crucially important that observations should be collected from as many clusters as possible.

The impact of within cluster sample size on statistical power is illustrated in Table 2, which gives the statistical power for some typical designs as a function of the within cluster sample size. The table shows that reducing within-cluster sample size from 40 to 20 has little

impact on statistical power. Even drastic reduction of cluster size from 100 to 10 reduces statistical power by less than 8% (from 0.78 to 0.72). The statistical power in most cluster randomized designs shows a similar relation between cluster size and power: Relatively small cluster sizes are needed to assure near maximal power for a design with a given number of clusters. Although we do not provide concrete examples in the table, direct computations show that the situation is quite similar for randomized block designs.

Table 2. The effect of cluster size on power of a cluster randomized trial

Number of Clusters	Cluster Size	Total N	Effect Size	Power
50	100	5000	0.25	0.78
50	80	4000	0.25	0.77
50	50	2500	0.25	.077
50	40	2000	0.25	0.77
50	30	1500	0.25	0.76
50	20	1000	0.25	0.75
50	10	500	0.25	0.72
50	9	450	0.25	0.71
50	8	400	0.25	0.71
50	7	350	0.25	0.69
50	6	300	0.25	0.68
50	5	250	0.25	0.66

Note: The Randomized blocks designs (RBD) allocate one half of each cluster to each treatment group, the intraclass correlation is $\rho = 0.2$, the significance level is $\alpha = 0.05$

4. Collect a Delayed Post-Test from a Subset of Units

For some interventions, a delayed posttest may be a reasonable way to obtain outcome information if the testing process was disrupted by the pandemic. This strategy might be combined with the strategy of collecting a smaller number of observations in each cluster. In evaluating whether this strategy is reasonable, it is important to consider how quickly (if at all) it is believed that treatment effect sizes are likely to diminish over time. While this may be difficult to quantify, scientists may have some intuition about the rapidity of fadeout. A small amount of fadeout may not be problematic. However, if fadeout is substantial enough that, e.g., the effect size for a delayed posttest will likely be only 50% of the size that would have been observed at the immediate posttest (and for which the study was adequately powered) then the study will likely be very underpowered.

5. Collect Survey Data from a Subset of Units

Sometimes outcome data (and very often data on process, implementation, and mediation) is based on surveys, which may also have been disrupted by the pandemic. It may be that surveys or interviews with teachers and students about the implementation of an intervention, for example, were disrupted in Spring 2020. One option is to delay collection of these surveys until

the 2020/2021 academic year. Here we outline a few strategies. Researchers should keep in mind, however that existing surveys may need to be revised to address possible recall and history bias. For example, all teachers – regardless of intervention arm – may idealize their pre-pandemic teaching experience.

Beyond recall bias, delaying the timing of the survey may result in a reduced survey response rate. Because a secular decrease in response rate is a generic problem in surveys, the survey sampling community has carried out a considerable amount of research on the problem (see, e.g., Tourangeau and Plewis, 2013). The most general advice is to consider whether exhaustive sampling is necessary, or whether a smaller, but better (more representative) sample could serve the needs of the study. A smaller, more targeted sample (perhaps to a random or stratified sample) would allow greater resources to be devoted to gathering data on the desired respondents and reduce nonresponse. Repeated follow-up inquires to non-respondents (at least 3 or 4) are often the key to obtaining a higher response rate. This is easier if it is possible to know who has responded so that, even if the content of the survey needs to be anonymous, some identifier can be used to determine who has responded. There are a variety of ways this can be done while assuring a reasonable degree of privacy for respondents.

While the primary survey response may be requested via a particular mode (e.g., an internet survey tool), experience in the survey industry indicates that multi-mode surveys that allow respondents to choose the response mode that they prefer can increase response rates substantially. Thus, offering the respondent the choice of responding via the internet, a mail in form, a telephone interview, or even an in-person interview can increase response rates and may be feasible if the sample is a small one worthy of intensive follow-up.

If survey results are crucial to the research plan, then it may be appropriate to use incentives for response. The sample survey literature on the effectiveness of incentives is considerable and is worth considering.

Whatever procedure is followed to collect the survey data, if there is a considerable amount of non-response the characteristics of non-respondents should be carefully examined. For guidance see the National Center for Education Statistics Statistical Standards (NCES, 2012). A recent National Research Council report on the future of social science data collection examined the apparent trend towards increasing non-response rates in surveys and concluded that while increasing nonresponse rates *could* compromise survey data collection, there was good reason to think that it often did not (Tourangeau and Plewis, 2013).

D. Change the Focus of the Study

If none of the approaches to salvaging important impact information from the trial are feasible or desirable, the most radical approach to dealing with disruption caused by the pandemic is to change the focus of the study from an efficacy trial to a development study or a methodological study. Impact estimates are not the only important things that can be learned from trials. A great deal of other information can be gleaned from trials that does not depend on obtaining treatment impact estimates.

1. Further Develop the Intervention

If the study is near the beginning of the funding period and relatively little of the budget has been expended, one option might be to focus on the changes in the intervention that might make it more relevant to the current context of education. For example, an intervention that focused on training parents to support their child's education might be very attractive in the context of remote instruction if it could be repackaged to be delivered online. Similarly, computer-based interventions intended for school use might be relatively easily adapted to remote learning context (and seen as highly desirable to schools). In this situation, some evaluation of the efficacy of the intervention might still be possible.

If the trial has proceeded further, more of the budget has been expended, and the pandemic has disrupted a trial severely, it might be possible to reconceive the project to focus on further development, perhaps aided by fieldwork to gain better insight about how the intervention works and how it might be improved, especially for the context of education in the pandemic and post-pandemic world. It seems likely that the outcomes sought by interventions are likely to remain relevant. Therefore, it may be particularly useful for investigators to consider how the delivery platform (or platforms) for the intervention might be changed to accommodate the context created by the pandemic. For example, an intervention that was previously developed for face-to-face delivery may be more useful if transferred to an online delivery, thus requiring new development and refinement. Similarly, an intervention that was previously focused on individual, small-group, or whole class delivery may realize that, as a result of increased inequities, a different delivery method may be more practical moving forward. Changes to the delivery mode of the intervention, however, are substantial and often require different project personnel with different skills than those typically involved in evaluations. Thus, while this is obviously not an option for all, or even most trials, for some projects, it might be considered (with the caveat, of course, that the project's funder is on board).

2. Convert to a Methodological or Measurement Study

Similarly, instrumentation is often developed in trials for measuring proximal outcomes, to provide formative feedback, or to measure final outcomes. Procedures are developed for recruitment, training, and data collection. Technology is developed for administering and monitoring the implementation of an intervention. These instruments, procedures, and technology have value not only to the trials that develop them, but potentially to the broader education science community.

This kind of information is often called "craft knowledge", and it is highly valued in some scientific communities. A good example is the survey research community to which we have already alluded. The survey sampling industry is a big business that is essential to managing health, economic policy, education, government, and individual businesses. This industry includes many large institutions (such as Westat, RTI, and NORC) and many smaller ones. The main professional association of survey researchers is the American Association for Public Opinion Research (AAPOR), which has over 2,000 members. Sample surveys can be complex undertakings in which many problems of instrumentation, procedures, and technology can arise. Consequently, two of the three journals AAPOR publishes (*Public Opinion Quarterly* and *Survey Practice*) focus on craft knowledge of conducting surveys.

Randomized field trials are at least as complex and difficult to carry out as sample surveys. While there is a considerable literature on the technical and statistical aspects of randomized trials, there is very little literature on the *craft knowledge* of conducting randomized trials in education (one example is Lemons, et al., 2014). Instead much of the practical knowledge about conducting trials resides in the heads of experienced education scientists who have conducted trials, giving them what some might call unfair advantages in the competition for grants to carry out trials, but also hampering the progress of education science. This is not a criticism of education scientists who have labored and sometimes pioneered in the difficult business of conducting randomized field trials in education. Rather it is a statement about the state of our field—we are only just accumulating enough experience to begin to have much craft knowledge. In running a summer institute on the design, conduct, analysis, and interpretation of randomized field trials for established researchers for over a decade, we have noticed how much of the practical advice we give is based on the experience of the instructors and how little is based on codified literature.

Changing the focus of a study from impact evaluation to documenting aspects of carrying out the trial is a radical shift in focus, but one that could still result in tangible (and publishable) products that advance knowledge. There is a tradition in medicine of publishing material about the design and practical details of randomized clinical trials. Virtually every large medical trial publishes one or more papers about the design and procedures of a trial, independent of data about treatment impact.

Example: Recruitment study

In education, the products of such craft knowledge might address pervasive problems in conducting trials. For example, recruitment is a pervasive problem in trials, so case studies explicating recruitment strategies (along with data about the process) could be quite useful to other researchers. A trial that successfully completed recruitment but could not collect outcome data might be well positioned to provide such a case study if good records of the recruitment process still existed or could be reconstructed. For example, many researchers and research centers that have carried out a sequence of trials have developed informal (and sometimes formal) partnerships with schools. These partnerships serve both to engage schools in research and researchers in the problems of schools. Case studies of the formation, nurturing, and functioning of such partnerships could be very useful to researchers interested in developing such partnerships. In fact, the pandemic may provide a situation in which schools would be particularly congenial to the formation of partnerships with researchers who could help provide urgently needed help in dealing with challenges in providing effective instruction with minimal risk to the health of all involved.

Example: Measurement Study

Similarly, a trial that developed new instrumentation or technology could provide a great service to the field by sharing not only the *results* of that development but insights about the *process* of the development with other education scientists. Trials that experimented with novel methods of training observers (using novel kinds of stimulus materials, simulations, or online distributed training methods) might have a great deal to offer other education scientists.

As systematic exposition of craft knowledge of education trials develops, we would expect to see more systematic evaluation of alternative methods (as we now see in the survey research literature). Yet in the survey field, it took many years for the literature to evolve from the presentation of case studies to more systematic investigation of comparative methods. It seems plausible that there needs to be a literature describing alternative procedures before comparative studies are even warranted.

3. Convert to a Descriptive Study

The Covid-19 pandemic has impacted nearly every facet of education and, as many scholars have noted, will exacerbate inequities across race and class. Already we are seeing that schools have dealt with this crisis in very different ways – with some districts focusing only on in-person classes, others only online, and others using various hybrid models. Even within these choices, the delivery mode and expectations for students varies, with districts serving high proportions of students in poverty delivering fewer hours of school and with fewer resources per student. Understanding this new landscape of education and inequity is essential if we are to intervene in ways that can mitigate these effects.

Researchers with trials ongoing in schools are in a unique position to observe, document, and understand these changes. They may also be in a position to work with schools as advisors, providing additional supports as leaders make curricular decisions. Some researchers may then be able to take advantage of these connections to shift from intervention research to research focused on describing, exploring, and understanding the on-the-ground problems, contexts, and constraints schools will face moving forward. This descriptive work is certainly not for everyone or every study but work of this type will be essential for the field.

Although we have a history of working in quantitative research traditions, we would argue that some of the most important work that might be done to understand the effects of the pandemic is essentially qualitative. The case studies suggested above are but a few examples. There are, however, broader questions about how schools are responding to the pandemic. While researchers who usually work on randomized trials may not be the most experienced in qualitative research methods, those involved in ongoing, but disrupted, trials already have created relationships and gained access to schools. Moreover, partially completed trials may have already collected quantitative data (about baseline achievement, composition of schools and instruction, etc.) that can help provide a descriptive framework for understanding the particular schools involved in trials and studying how those schools respond to the pandemic. A set of contrasting cases (e.g., of similar schools that responded differently or quite different schools that responded similarly) could prove remarkably enlightening.

4. Convert to a Study of a Different Intervention

An even more radical shift in focus would be to exploit the fact that the disruption caused by the pandemic has raised a huge number of new and pressing research questions. It has also created a vast natural experiment from which much might be learned. For example, systematic instruction in some schools essentially ceased before the end of the school year. What effects will that have on learning? How long will it take to make up the losses in learning incurred due to the pandemic? What effects will the pandemic have on dropping out of school or going to college?

Will there be effects on the type of college to which students matriculate? How are each of these kinds of effects different for different societal groups? How do these differences contribute to inequality? What impacts has the pandemic had on school personnel? Will it hasten the departure of older or minority personnel (who are more at risk from Covid-19) from teaching and administration? The pandemic has caused new teaching arrangements (virtual classes and presumably in Fall 2020, classes and whole schools reorganized to involve social distancing). What will be the effects of these new teaching arrangements? As state and local revenues plummet due to the economic effects of the pandemic, schools are likely to face serious budget shortfalls. How will this impact schools, academic attainment, and academic achievement?

Few of these kinds of questions are likely to ever be addressed experimentally. Yet the pandemic has created a vast natural experiment which may permit us to investigate some of them using quasi-experimental methods. It creates an opportunity to learn much by using difference in difference, (comparative) interrupted time series, regression discontinuity, and propensity score methods. For example, if some schools provided a more organized and intensive (or just a different) program of instruction when schools were closed than others, comparing before and after levels of achievement may yield interesting insights. In some cases, these changes may be staggered – for example, with students with IEPs receiving priority in returning to in-person classes before other student. Furthermore, there may even be opportunities to examine questions that could never be studied without an extreme disruption like a pandemic. For example, if there are some schools that essentially ceased organized instruction, while others did not, we might obtain insight about how much, in an absolute sense, schooling (as opposed to out of school experience and just getting older) increases achievement.

Studies that are likely to be most disrupted by the pandemic (those that already began, but did not finish, data collection) may be the ones that are most poised to exploit the natural experiment created by the pandemic. Having already collected pre-pandemic data in this situation is an opportunity that cannot be duplicated later. Moreover, having already established relationships with schools and their personnel gives these trials an advantage in negotiating future work with those schools. However, exploiting natural experiments requires nimbleness, great creativity, and a certain amount of good luck. It might involve abandoning one research focus for others. Not every investigator will find this appealing. Others may be more attuned to serendipity and see this as an exciting opportunity to address interesting research they had never anticipated tackling. Researchers poised to take this leap, however, should discuss these ideas with their grant funding agency and program officer, as such a change is far beyond the initial scope of work.

VI. Additional Solutions

In the previous section, we focused on the four broad paths forward that we believe will be most widely useful to researchers. In this section, we address three other possibilities that, while perhaps not as broadly useful, we expect may provide options for at least a few researchers with specific types of problems.

A. Use Proxy Pretests

If the pandemic disrupted collection of pre-tests (baseline or before baseline), but recruitment and the intervention can proceed in 2020/2021 or 2021/2022, then the use of proxies for the planned pretests might be a reasonable strategy. Pretests are primarily used for two purposes:

1) Increasing statistical power and precision in estimates of treatment impact, and
2) Checking that random assignment achieved minimal baseline differences between treatment groups.

In cluster randomized designs, a cluster level covariate almost always contributes the most to increasing design sensitivity. While pretests of the same construct measured as the outcome are optimal for increasing statistical power and precision in estimates of treatment impacts, proxies that are quite different can do almost as well in cluster randomized (hierarchical) designs (Bloom, et al., 2007). Note, too, that pretests on reading may be almost as effective as a pretest in math for measuring impact on math (or math pretest for measuring impact on reading). Moreover, pretest data measured on the same clusters in previous years (that is, not involving the students whose outcomes are being measured) are also almost as effective in increasing design sensitivity as covariates measured on the individuals whose outcomes are being measured (e.g., annual state-test scores). Not only that, the effectiveness of covariates can be increased modestly by using multiple covariates. These considerations suggest that administrative data or data originally collected for other purposes might serve reasonably well to help increase design sensitivity.

Proxies for pretests may be less convincing for assessing baseline differences. If baseline measures of roughly the same construct as the outcome, but obtained from administrative data, are available, these could be used to assess baseline equivalence. Using more measures (if several are available) could serve to make the arguments about baseline equivalence more convincing. Obviously, proxies for pretest that are not gathered from the same participants as those on which the outcome is measured cannot provide convincing information on baseline equivalence among treatment groups.

Occasionally pretests are used to determine the composition of treatment groups (for example, selecting struggling students when the intervention is specifically targeted at such students). It may be worth considering such pretests as covariates if they are available.

The major risk in using proxy covariates is that they may seem counter-intuitive to some scientists. For example, the What Works Clearinghouse distinguishes between balance assessed on baseline covariates of the same construct and covariates that are not measures of the same construct. Another risk is that, although administrative data is seductive because it exists, actually obtaining it may be expensive and time consuming. For example, there may be major administrative and practical hurdles to obtaining the data (e.g., it may exist only in a form that requires hand transcription).

B. Change to Within-School Randomization

If it is not feasible to sufficiently increase design sensitivity, and if a low power study is not considered desirable, then substantially changing the research design moving forward might be an option. For some interventions, it might be feasible to consider assignment of both intervention and comparison conditions within the same cluster (school), so that the design

becomes a (generalized) randomized blocks design. The actual assignment might involve the assignment of intact classrooms to interventions within schools. Sometimes it may be possible to randomly assign students either across an entire grade level (e.g., in a program not administered as part of a classroom) or individually within classrooms (e.g., when the intervention is embedded within technology used by every student). Such designs are typically more sensitive than cluster randomized designs with the same sample size.

Many researchers have rejected randomized blocks designs because they fear diffusion of an intervention between the intervention and comparison conditions (usually classrooms) will reduce the treatment impact. Treatment diffusion does have the potential to fatally compromise the treatment contrast (for example if there is active subversion, such as in a tutoring intervention where comparison group children are allowed to attend tutorial meetings with the treatment children). However, interventions that involve complex changes in behavior or instruction are unlikely to be spontaneously adopted in comparison group classrooms. Interventions that involve special technology or extensive training seem particularly unlikely to experience diffusion to comparison group classrooms. Extensive investigation of the comparative sensitivity of cluster randomized and randomized blocks designs show that the amount of diffusion of an intervention has to be very large to overcome the sensitivity advantages of randomized blocks designs (Rhoads, 2011). To put it another way, even if diffusion of an intervention reduces the treatment impact by 25%, the randomized block design will almost always be more powerful than a cluster randomized design with the same number of clusters and individuals.

The statistical power of randomized blocks designs depends on the significance level (which we fix in this discussion at the conventional 0.05), the effect size, and the intraclass correlation, just as it does for cluster randomized designs. But in randomized blocks designs, power also depends on another parameter representing the heterogeneity of treatment effects across blocks (clusters). The greater this heterogeneity, the lower the power, when everything else is equal. For an explanation of how to compute power for both types of designs see Hedges and Rhoads (2009).

The discussion of randomized blocks designs above assumes that the blocks (clusters or schools) are treated as random effects. This analytic choice supports generalization of treatment impacts to a population of clusters that are like those in the sample. The sensitivity of randomized block designs can be increased even further by making the analytic choice to treat the clusters as fixed. This changes the analysis and increases the sensitivity of the design, but at a cost. By considering the clusters (blocks) as fixed, generalizations are limited to the clusters in the sample or to clusters "sufficiently" like them. Thus, the results of such an analysis are more appropriate to demonstrating "proof of concept," rather than demonstrating a widely generalizable treatment impact. However, analyses using fixed block effects are statistically valid for the intended inferences and are widely accepted in science. Choosing fixed block effects can greatly increase the sensitivity of the design, permitting stronger inferences to be drawn from studies with smaller than conventional sample sizes (relative to when clusters are considered random).

Table 3. Comparison of the statistical power of hierarchical (cluster randomized) and randomized blocks designs

Design	Total Number of Clusters	Cluster Size	Total N	Effect Size	Power
Hierarchical (CRT)	50	100	5000	0.25	0.78
RBD Blocks Random	25	200	5000	0.25	0.84
RBD Blocks Fixed	25	200	5000	0.25	> 0.99
Hierarchical (CRT)	40	100	4000	0.25	0.68
RBD Blocks Random	20	200	4000	0.25	0.75
RBD Blocks Fixed	20	200	4000	0.25	>0.99
Hierarchical (CRT)	40	20	800	0.25	0.67
RBD Blocks Random	20	40	800	0.25	0.74
RBD Blocks Fixed	20	40	800	0.25	>0.99
Hierarchical (CRT)	40	20	800	0.25	0.65
RBD Blocks Random	20	40	800	0.25	0.72
RBD Blocks Fixed	20	40	800	0.25	0.98
Hierarchical (CRT)	30	15	450	0.25	0.51
RBD Blocks Random	15	30	450	0.25	0.56
RBD Blocks Fixed	15	30	450	0.25	0.84

Note: The Randomized blocks designs (RBD) allocate one half of each cluster to each treatment group, the intraclass correlation is $\rho = 0.2$, the significance level is $\alpha = 0.05$ and in the randomized blocks design with random blocks, the heterogeneity parameter described in Hedges and Rhoads (2009) is $\omega = 0.4$.

To illustrate the relative sensitivity of the three designs, we offer some comparisons of the statistical power of cluster randomized, randomized block designs with random block effects and randomized block designs with fixed block effects in Table 3. Examining the table, we see that when the treatment effect size is $\delta = 0.25$ and the intraclass correlation is a not unreasonable $\rho = 0.2$, a cluster randomized design randomizing $m = 50$ schools and 100 students per school (and reasonable values of level 1 and level 2 correlations between covariates and outcome of $R_1^2 = 0.5$ and $R_2^2 = 0.8$)[6] has statistical power of only 0.78. On the other hand, a randomized blocks design with the same total number of students (split evenly across 25 schools) treating blocks as random (with modest treatment effect heterogeneity of $\omega = 0.4$) has somewhat larger power of 0.84, and the same randomized blocks design with fixed block effects has power exceeding 0.99. It is instructive to compare the cluster randomized design described in the first row of Table 3 with the randomized block design described in the last row of the table with fixed block effects.

[6] The values $\rho = 0.2$, $R_1^2 = 0.5$, and $R_2^2 = 0.8$ are close to the national averages computed from representative samples by Hedges and Hedberg (2007).

The randomized block design would achieve higher statistical power, and does so with only 15 schools, each randomizing 30 students to treatments and a total sample size of one tenth that of the cluster randomized design described in the first row of the table (450 versus 5,000 students). The important point we wish to illustrate is that by changing the design from a cluster randomized design to a randomized block design, it may be possible to preserve design sensitivity even if the realized sample size is smaller than that initially planned.

C. Change to a Quasi-Experiment

It is important to recognize that random assignment to intervention and comparison groups, in and of itself, does not guarantee internal validity. Disruption caused by the pandemic has the potential to compromise randomized trials. For example, disruption is likely to cause higher than expected attrition. It is also quite plausible that the attrition will not be uniform across these groups. For example, discontinuing participation in an experimental study might be one of the first things schools consider as they adapt to the changing conditions of education in the pandemic. Even if attrition appears to be uniform across treatment groups, high enough attrition raises questions about internal validity. This is why many education scientists (and the What Works Clearinghouse) regard high attrition randomized trials as having essentially the same status as quasi-experiments. Attrition is not the only way a randomized trial can be compromised (non-compliance with treatment assignment is another obvious way). Thus, it is crucial that the disruption caused by the pandemic is not ignored in the analysis of the trial. The literature on "broken" randomized experiments provides some guidance (see, e.g., Barnard, Du, Hill, and Rubin, 1998; or Holmes, 2014).

Nonetheless, a much more radical design change would be to select a design that does not involve randomization. This might happen, for example, if the pandemic made the intervention so attractive (e.g., remote learning platform) that schools would rather seek to implement some version of it themselves than risk not being able to do so if randomized to the comparison group. We would advise that this is an extreme change, that such a choice should be considered very carefully, and that it should not be made unless there is no feasible alternative.

Randomized trials are the strongest designs for making causal inferences. While quasi-experiments *can* provide minimally biased estimates of causal effects under certain circumstances, it is difficult to know if those circumstances exist in the real world. Moreover, strong quasi-experiments often require larger sample sizes (for instance, because extensive matching may require a large reservoir of comparison units to ensure enough control units that have an adequate match). Thus, if difficulty in recruiting an adequate sample size is motivating the change to a quasi-experimental design, switching to a design that requires an even larger sample size than would the randomized trial may not be the best way to resolve the difficulty. In addition, adequate analyses and interpretations of quasi-experiments tend to be different and more complex than the analyses of randomized trials. It is also important to remember that even computation of statistical power for the testing of treatment impact in quasi-experimental designs is more complicated than in experimental designs. Thus, not only does the quasi-experimental design have less internal validity, it can be difficult to know if a quasi-experimental alternative to the randomized design is adequately sensitive to detect the likely effects of the intervention. It is quite likely that a project changing its design from an experimental to a quasi-experimental

design would need additional methodological personnel with somewhat different qualifications than would be necessary for a randomized trial.

VII. General Considerations and Validity Concerns

We have argued that salvaging the considerable investment of the investigator's time and research agency's resources that have already been invested in ongoing trials is of paramount importance. However, the scientific value of any evidence that is produced is ultimately limited by the validity of that evidence. Virtually all of the suggestions for modifications to trials in response to the disruption caused by the pandemic raise at least some validity concerns which should be considered as investigators decide how to respond to the challenges created by the pandemic. In doing so it is wise to recall that validity (and any subtype of validity) is not a binary variable. Studies are not valid or invalid but are more or less valid on a conceptual scale that may be difficult to quantify. This, of course, was also true of the study that was intended in the first place—even the best study is not without validity concerns.

Most scientists understand that no study is without validity concerns and that progress in science depends not on conducting "perfect" studies but on clearly understanding the limitations of studies while working diligently to minimize them. Sophisticated critics can identify validity concerns and their goal in evaluating studies is not so much to determine whether the study has flaws, but whether the investigator knows that, why they chose one research strategy over another (also flawed) research strategy, and whether, holistically, something can be learned from the study. We urge all investigators to take a self-critical approach and articulate the potential validity concerns in their work and their logic for doing what they chose to do.

In this section, to help investigators evaluate their choices, we exploit the well-known validity framework suggested by Shadish, Cook, and Campbell (2002). This framework includes four major types of validity concerns: Statistical conclusion validity (validity of the statistical analysis), internal validity (freedom from bias), external validity (generalizability), and construct validity of cause. This section is not intended to free the reader from careful evaluation of validity concerns—they will be somewhat different in every trial. Instead, we hope to suggest some ideas that could be the beginnings of such an evaluation.

A. Statistical Conclusion Validity

Statistical conclusion validity is concerned with whether the statistical analysis will yield the correct conclusion about the relation between an intervention and the measured outcome. Anything that compromises design sensitivity compromises statistical conclusion validity. A major problem created by the disruption caused by the pandemic is reduced sample sizes due to failed recruitments or attrition from trials, which leads to reduced design sensitivity.

Changes in the measurement model may change the measurement properties of baseline or outcome measurements. Design sensitivity (e.g., statistical power) depends in part on the treatment effect size, and effect size depends on the reliability and validity (e.g., alignment to the outcome construct the intervention is intended to change) of the outcome measurement.

A more complex concern related to the measurement model is whether changing the data collection mode may create differential functioning of the measures. The concept of differential functioning is often invoked at the test item level to describe how some items are more difficult

for one societal group than another. The same issue can arise with respect to entire measures. For example, does switching to an online measure disadvantage poor children who may have weaker internet connections or less congenial settings in which to take the test, which might affect their performance? While a decrease in performance of one subgroup might not affect the internal validity of the study, it might increase variation which would decrease the effect size and reduce design sensitivity. Note that this is not just a phenomenon that could increase within-cluster variation. If, for example, poor children are more concentrated in some clusters than others, this could increase between-cluster variation which is the major contributor to imprecision in estimation of treatment effects and could have an important effect on statistical power.

Many of the other decisions that might be made to deal with the disruption due to the pandemic can also have consequences for statistical conclusion validity. For example, the decision to evaluate an intervention that is only partially implemented may be reasonable, but it is likely to reduce the magnitude of the treatment effect size, which will reduce design sensitivity.

B. Internal Validity

Internal validity is concerned with whether the observed relation between an intervention and an outcome is *causal*. Concern about internal validity is what privileges randomized trials among all other designs in estimating causal effects. Thus, any change from a randomized design to a quasi-experimental design – whether purposeful or as a result of attrition – raises concerns about internal validity.

It is important to remember that while perfectly implemented randomized trials have high internal validity, there are many ways that the implementation of randomized trials can be compromised. Attrition can compromise the internal validity of randomized trials, and attrition is quite likely as a consequence of the pandemic, so it is particularly important to document and address attrition and particularly differential attrition. Addressing attrition may be a contentious business and may involve the use of several approaches in concert. The development of robust multiple imputation models is one way to address attrition. This approach is likely to be more convincing if several models are considered and evaluated and a principled substantive (as well as statistical) rationale is given for the preferred model. Another approach is the development of "best" and "worst" case bounds for likely treatment effects (e.g., by imputing the plausible scores that lead to the largest and smallest treatment impacts). Yet another approach is to examine the treatment impact on proximal outcomes (e.g., an intermediate test or classroom test) for both those who are missing final outcome scores (leavers) and those who do not have those final outcome scores (leavers). With substantial or differential attrition, no one approach may be convincing, but the convergence of several might be.

The pandemic is a powerful external shock and its effects are as yet unknown. If the pandemic occurred between randomization and collection of outcome data, then interactions between treatment and the pandemic are serious threats to the internal validity of randomized trials. While it is difficult to evaluate how serious this threat might be, it is incumbent upon researchers to try to understand how the pandemic might have had a differential effect on the intervention and comparison conditions. This may involve posing hypotheses and gathering whatever data can be assembled to try to evaluate whether these differential effects occurred.

C. External Validity

External validity is concerned with the extent to which the causal effects found in a study would occur in other situations than those in which the trial was conducted. There has recently been a considerable amount of research on the generalizability of randomized trials (Stuart, Cole, Bradshaw, and Leaf, 2011; Tipton, 2013; for an overview, see Tipton and Olsen, 2018) and the SEER principles encourage researchers to "facilitate generalization of study findings." The disruption caused by the pandemic may compromise external validity of trials in a variety of ways. One very serious question is whether the context of post-pandemic schooling will be at all like that of pre-pandemic schooling and how much that will affect the generalizability of pre-pandemic findings. While that kind of question is essentially unanswerable, there is at least some reason to believe that basic science about teaching and learning will remain valid and interventions that are based on that science will remain efficacious.

A more practical question arises when the implementation of the intervention is changed by the pandemic. For example, if an intervention designed for in-class use was administered partially or entirely online, are the results of the trial generalizable to in-class use of that intervention? While it may not be possible to gather definitive evidence about generalizability in this case, some indications may be obtainable. For example, is there evidence that the intervention was adapted to the online setting? What were those adaptations and do the developers of the intervention believe that they change its active (or enabling) ingredients? Even interventions intended to be delivered online could have different effects when administered in a classroom setting with teacher supervision than when they are used at home without teacher supervision.

Another example concerns measurement. If measures were administered online, we discussed that these measures may have different psychometric properties than they would have had if the measures been used in class. This may compromise statistical conclusion validity, but also generalizability because the trial would have had a somewhat different outcome had it been conducted in class.

A particularly difficult issue arises for trials that are conducted in multiple cohorts and treatment implementation and/or data collection for one or more of those cohorts was completed before the pandemic. This raises a fundamental question of whether it is meaningful to combine the data from pre- and post-pandemic data collections. There is always a question of whether it is meaningful to combine different cohorts, but usually there are minimal differences in context between subsequent cohorts. The pandemic makes this question more difficult by imposing more dramatic differences between the contexts experienced by different cohorts. The usual devices of checking for cohort effects and treatment by cohort interactions are certainly warranted, although these are like to involve relative insensitive (low power) tests.

D. Construct Validity of Cause

Construct validity of cause concerns whether the correct attribution of the causal agent has been made. This is a very broad and conceptual kind of validity claim which is often difficult to evaluate. However, the large shock to the education system caused by the pandemic makes misattribution of cause a particularly serious possibility. Whether the examples given below

properly fall under the category of internal or construct validity might be debated, but they certainly do constitute misinterpretation of the causal ingredient.

For example, if the intervention is particularly well packaged and organized (e.g., for online administration) while the comparison condition is usual classroom instruction, and all instruction was unexpectedly online because of the pandemic, then it is plausible that the intervention would outperform the hastily prepared comparison instruction simply because it was better *online instruction*, not better *instruction*.

Many similar scenarios could be imagined. For example, teachers might be so demoralized by the adaptations required due to the pandemic that their preparation or delivery of lessons was not what it would have been in class, but the intervention unit was already prepared for them and thus more effective simply because it was better prepared. Thus, it is plausible that the intervention outperformed the comparison group because the lesson preparation in the comparison condition was disrupted preparation, not because the intervention provided better instruction than the comparison condition would have under "normal" circumstances.

It is possible that teachers in the intervention group became more enthusiastic about the intervention (e.g., it gave them a sense that they could do something special for their students in a difficult time) enabling them to put more effort into instruction in the intervention group while teachers in the comparison group were demoralized by trying to cope with the difficulties brought about by the pandemic. Alternatively, it is possible that intervention teachers regarded the intervention as "just too much" to try to cope with given the pandemic and reduced their effort in implementing it while the comparison group teachers pursued something closer to their usual effort.

Yet another possibility is that because the overall amount of instruction was reduced as a consequence of the pandemic, some material might have been omitted from instruction to the comparison group that was dutifully included in the instruction to the intervention group because there was an external impetus to teach that material.

These scenarios are not necessarily relevant to any trial and they are certainly not the only ones that can be imagined. Yet scenarios like these might be conceivable in some trials. It is incumbent on investigators to consider how conditions caused by the pandemic might lead to misattribution of causal effects and to gather evidence that their proposed attribution is likely to be correct.

VIII. Evaluating Possible Strategies and Practical Considerations

Evaluation of possible strategies for dealing with disruptions caused by the pandemic must consider the stage of the trial, the nature of the intervention and the context in which the trial is being implemented. Important practical issues are the amount of the budget already expended and the feasibility of reducing the level of expenditures until the situation in schools returns to something like pre-pandemic conditions. The paramount considerations should be assuring that the investment already made in the trial (not only by the supporting agency, but also the investigators' human capital) results in knowledge which has value commensurate with the investment made (once health and safety of students, school personnel, and researchers is assured). Any such evaluation will be complex and a matter of judgment (of both the investigator and their program officer).

Substantial changes in research strategy or timelines have implications (even painful ones) for budget allocations and for staffing. It is always difficult to risk losing talented research staff because of delays or because they may not be needed in the new project configuration, but it may be inevitable. It is made more difficult when new staff (with different qualifications) may need to be hired. However, this will have to be managed with as much grace and sensitivity as possible.

If substantial changes are made in the collection of data from human subjects, these must be cleared with the Institutional Review Board (IRB) with jurisdiction over the research (typically a university or research center IRB). Clearance may also be required by school district IRBs for changes if their approval was originally required for the research.

Substantial changes in the research plan or schedule must also be approved by program officers at the funding agency. A request for approval of such changes will need to be accompanied by a cogent rationale for those changes, but it is safe to assume that every program officer will be interested in salvaging as much scientific value from their portfolio of grants as possible. In the current environment, it is difficult to imagine that such requests would be unexpected. In fact, it is probably wise to assume that program officers will receive many such requests and that rapid replies should not be anticipated. Ample time for program officers to evaluate requests for changes should be allocated.

IX. Conclusions

Investigators who have recently obtained funding for randomized trials or are currently in the field with their studies face the quandary of what to do about the inevitable disruption caused by the pandemic. There are no easy answers and there is no reason to assume that there is an optimal answer for any particular trial. Yet scientists face similar dilemmas when they plan research studies in "ordinary" circumstances. Evaluation of possible research strategies involves creatively choosing from among many possibilities, each involving unknowns with no assurance that unforeseen problems will not arise. Despite all this uncertainty, some choices do seem (in the consensus opinion of scientists involved in the peer review process) to be better than others. We believe that this is largely because some choices are supported by arguments that are better articulated, more soundly warranted, and more grounded in conventional scientific practice. While the choice of a research strategy in the face of the pandemic may involve greater uncertainties and contexts that are less well known, we believe that the same considerations of scientific judgment will ultimately determine which research choices are most wise given what was known when those choices were made.

References

American Educational Research Association. (2006). Standards for reporting on empirical social science research in AERA publications. *Educational Researcher, 35*, 33 –40. https://doi.org/10.3102/0013189X035006033

Barnard, J., Du, J., Hill, J. L., and Rubin, D. B. (1998). A broader template for analyzing broken randomized experiments. *Sociological Methods and Research, 27*, 285–317.

Bartlett, C. W., Klamer, B. G., Buyske, S. A., Petrill, S. A., & Ray, W. C. (2019). Forming big datasets through ltent class concatenation of imperfectly matched database features. *Genes, 10*, 727; doi:10.3390/genes10090727

Bloom, H. S. (1995). Minimum detectable effects: A simple way to report the statistical power of experimental designs. *Evaluation Review, 19*, 547-556.

Bloom, H. S., Richburg-Hayes, L., & Black, A. R. (2007). Using covariates to improve precision for studies that randomize schools to evaluate educational interventions. *Educational Evaluation and Policy Analysis, 29(1)*, 30–59.

Borenstein, M., Hedges, L. V., Higgins, J. P. T., & Rothstein, H. R. (2009). *Introduction to Meta-analysis*. London: John Wiley

Cooper, H. M., Hedges, L. V., & Valentine, J. (2019). *The handbook of research synthesis and meta-analysis (3rd edition)*. New York: Russell Sage Foundation.

Feuer, M. J., Holland, P. W., Green, B. F., Bertenthal, M. W., & Hemphill, C. (1999). Uncommon measures: Equivalence and Linkage Among Educational Tests. Washington, D. C.: National Academies Press. https://www.nap.edu/catalog/6332/uncommon-measures-equivalence-and-linkage-among-educational-tests

Hedges, L. V. (1981). Distribution theory for Glass's estimator of effect size and related estimators. *Journal of Educational Statistics, 6*, 107-128.

Hedges, L. V. & Hedberg, E. C. (2007). Intraclass correlations for planning group-randomized experiments in education. *Educational Evaluation and Policy Analysis*, *29*, 60-87.

Hedges, L. V. & Olkin, I. (1985). *Statistical methods for meta-analysis*. New York: Academic Press.

Hedges, L. V. and Rhoads, C. (2009). *Statistical Power Analysis in Education Research (NCSER 2010-3006)*. Washington, DC: National Center for Special Education Research, Institute of Education Sciences, U.S. Department of Education.

Herbst, C. M. (2008). Do social policy reforms have different impacts on employment and welfare use as economic conditions change? *Journal of Policy Analysis and Management, 27*, 867-894

Hill, C. J., Bloom, H. S., Black, A. R., & Lipsey, M. W. (2008). Empirical benchmarks for interpreting effect sizes in research. *Child Development Perspectives, 2*(3), 172–177. https://doi.org/10.1111/j.1750-8606.2008.00061.x

Holmes, W. M. (2014). *Using propensity scores in quasi-experimental designs*. Thousand Oaks, CA: Sage Publications. https://dx.doi.org/10.4135/9781452270098

Kolan, M. J. & Brennan, R. L. (2004). *Test equating, scaling, and linking*. New York: Springer.

Lemons, C.J., Fuchs, D., Gilbert, J.K., and Fuchs, L.S. (2014). Evidence-Based Practices in a Changing World Reconsidering the Counterfactual in Education Research. *Educational Researcher, 43* (5), 242–252.

Mislevy, R. J. (1992). *Linking educational assessments: Concepts, issues, methods, and prospects.* Princeton: Educational Testing Service Policy Information Service.

NCES. (2012). *2012 revision of the NCES statistical Standards*. Washington, DC: National Center for Education Statistics, Institute of Education Sciences. https://nces.ed.gov/statprog/2002/stdtoc.asp

Phillips, G., Jia, Y., Xu, X., Wise, L. L., Wiley, C., Diaz, T. E., & Rahman, T. (2014). *2011 NAEP-TIMSS linking study: Technical report on the linking methodologies and their evaluations. (NCES2014-461)*. Washington, DC: National Center for Education Statistics, Institute of Education Sciences.

Rhoads, C. (2011). The implications of contamination for experimental design in education research. *Journal of Educational and Behavioral Statistics, 36*, 76-104.

Shadish, W. R., Cook, T. D. & Campbell, D. T. (2002). *Experimental and quasi-experimental designs for generalized causal inference*. Boston: Houghton Mifflin.

Shavelson, R. & Towne, L. (2002). *Scientific Research in Education.* Washington, DC: National Academies Press.

US Institute of Education Sciences (2020). *Standards for Excellence in Education Research.* Last retrieved March 20, 2020. https://ies.ed.gov/seer/

Stuart, E. A., Cole, S. R., Bradshaw, C. P., & Leaf, P. J. (2011). The use of propensity scores to assess the generalizability of results from randomized trials. *Journal of the Royal Statistical Society, Series A*, Part 2, 369-386.

Tipton, E. (2013). Improving generalizations from experiments using propensity score subclassification: Assumptions, properties, and contexts. *Journal of Educational and Behavioral Statistics, 38*: 239-266.

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

Tourangeau, R. & Plewes, T. J. (Eds.) (2013). *Nonresponse in social science surveys*. Panel on a Research Agenda for the Future of Social Science Data Collection, Committee on National Statistics. Division of Behavioral and Social Sciences and Education. Washington, DC: The National Academies Press.

Valentine, Aloe, A. M., & Wilson, S. J. (2019). Interpreting effect sizes. Pages 453-470 in H. Cooper, L. V. Hedges, and J. C. Valentine (Eds.) *The handbook of research synthesis and meta-analysis (3rd edition).* New York: The Russell Sage Foundation.

Wasserstein, R. L. & Lazar, N. A. (2018). The ASA statement on *p*-values: Context, process, and purpose. *The American Statistician, 72*, 129-133.

Weiss, M. J., Bloom, H. S., & Brock, T. (2014). A conceptual framework for studying the sources of variation in program effects. *Journal of Policy Analysis and Management, 33(3)*, 778-808. https://www.jstor.org/stable/24033389

Wilkinson, L. (1999). Statistical methods in psychology journals. *American Psychologist, 54,* 594-604.