AERA Open
October-December 2019, Vol. 5, No. 4, pp. 1–15
DOI: 10.1177/2332858419891963
Article reuse guidelines: sagepub.com/journals-permissions
© The Author(s) 2019. http://journals.sagepub.com/home/ero

A Path to Greater Credibility: Large-Scale Collaborative Education Research

Matthew C. Makel
Duke University

Kendal N. Smith
University of North Texas

Matthew T. McBee
East Tennessee State University

Scott J. Peters
University of Wisconsin-Whitewater

Erin M. Miller
Bridgewater College

Concerns about the replication crisis and unreliable findings have spread through several fields, including education and psychological research. In some areas of education, researchers have begun to adopt reforms that have proven useful in other fields. These include preregistration, open materials and data, and registered reports. These reforms offer education research a path toward increased credibility and social impact. In this article, we discuss models of large-scale collaborative research practices and how they can be applied to education research. We discuss five types of large-scale collaboration: participating teams run different studies, multiteam collaboration projects, collaborative analysis, preregistered adversarial collaboration, and persistent collaboration. The combination of large-scale collaboration with open and transparent research practices offers education researchers opportunity to test theories, verify what is known about a topic, resolve disagreements, and explore new questions.

Keywords: *credibility revolution, open science, replicability crisis*

Opportunities abound for improving how academic research is conducted. For example, recent research in fields such as psychology (Open Science Collaboration [OSC], 2015) has revealed that many previously published findings cannot be replicated by independent research teams. The existence of this "replication crisis" (Pashler & Harris, 2012, p. 531) suggests that results of single studies should be viewed more skeptically and provisionally than had been previously appreciated. A second opportunity for improvement stems from the fact that education research does not always provide particularly informative findings. For example, Lortie-Forgues and Ingles (2019) analyzed 141 randomized controlled trials supported by the National Center for Educational Evaluation and Regional Assistance in the United States and the Education Endowment Foundation in the United Kingdom. These studies reported effect sizes from −0.16 to 0.74, with a median of 0.03. Under a Bayesian framework, 40% of the studies fell into the "uninformative" range (BF_{10} between 1/3 and 3). Although decisive null results are valuable, truly uninformative or indecisive results (in the Bayesian sense of the term) are tragic; the research project has been completed, but there is no convincing evidence for or against the phenomenon being investigated.

A third opportunity for improvement is connected to the fact that there are many ideas in education that are hugely popular among educators but lack a strong empirical foundation. Consider recent findings regarding the status of popular ideas such as *learning styles* (Pashler, McDaniel, Rohrer, & Bjork, 2009), *growth mindset* (Sisk, Burgoyne, Sun, Butler, & Macnamara, 2018), *grit* (Crede, Tynan, & Harms, 2017), or *multiple intelligences* (Waterhouse, 2006). All these are staples of teacher training and K–12 classroom practice, but

Creative Commons Non Commercial CC BY-NC: This article is distributed under the terms of the Creative Commons Attribution-NonCommercial 4.0 License (http://www.creativecommons.org/licenses/by-nc/4.0/) which permits non-commercial use, reproduction and distribution of the work without further permission provided the original work is attributed as specified on the SAGE and Open Access pages (https://us.sagepub.com/en-us/nam/open-access-at-sage).

their research basis reveals either lack of compelling supporting evidence or very small effects on academic achievement. The growing body of research results that cannot be replicated, yield uninformative results, and/or are often irrelevant to popular implementation, are clear signs that status quo research practices could benefit from improvement.

In this article, we discuss a set of large-scale collaboration strategies that could help education produce more replicable, informative, and relevant education research. In the following sections, we review some existing flaws of the modern academic research process and then introduce different types of collaborative research initiatives, explain why they are helpful generally, review what they could do to help improve education research, and provide resources for how to begin implementing these practices.

Issues in the Existing Academic Research Structure

In this section, we discuss several characteristics of common research practice that we believe contribute to the problems introduced above. These contributors are lack of replication, researcher flexibility, lack of statistical power, limitations of "big data," and misaligned incentives. Each on its own limits what education research as a field can accomplish. Together, they form serious limitations to the quality and credibility of the research being produced.

Lack of Replication

The research literature of every field contains false positives, but in education we know neither the rate nor the identity of these false claims because the field rarely engages in or publishes replication research. When the status of research findings is unclear, replication can help. Replication studies are a scientific field's immune system. The presence of a robust culture of replication where replication efforts are routine, valued, and disseminated can help a field establish greater reliability of results, or at least greater knowledge of constraints on generality (Simons, Shoda, & Lindsay, 2017). In such a system, false findings would be identified and purged from the intellectual bloodstream of the field. Without the immunity bestowed by widespread, frequent replication studies, fields can become infected by a form of antiknowledge—untrue notions that become accepted as factual. Theories become calibrated to fit these untrue facts, and eventually the whole discipline is in crisis, as has recently occurred in social psychology. Lack of replication is one of the many potential contributors to the creation and spread of antiknowledge.

In 2014, Makel and Plucker analyzed the complete publication history of the top 100 academic journals in the field of education (rated by impact factor) and found that only 0.13% of the articles (461 articles out of 164,589) were labeled as replications. When authors did engage in some type of replication, 69% reported successfully replicating the previous finding. This number shrunk to 54% when none of the replicating authors were also authors of the original study, meaning that nearly half failed to replicate the previous finding. Similarly, Chhin, Taylor, and Wei (2018) examined all causal impact grant applications funded by the Institute of Education Sciences (IES; 2018) between 2004 and 2016 ($N = 307$). Half were classified as conceptual replications that investigated a previously tested intervention but varied some aspect of the methods, context, outcomes, or the intervention itself. None were direct replications, and only 30% mentioned replication in the application. This is particularly concerning given that a single replication is likely insufficient for an unambiguous test of an effect (Hedges & Schauer, 2019).

The lack of replication attempts is rooted in long-standing weaknesses in common research practices. One of the earliest signals to receive widespread attention was Ioannidis's (2005) paper titled, "Why Most Published Research Findings Are False." His analysis proved prescient; when disciplines began to experiment with systematic replication studies (e.g., Begley & Ellis, 2012; OSC, 2015), the results were grim. A decade later, a majority of physical and life scientists reported believing that there was a "significant crisis" with 90% believing that at least a "slight crisis" was occurring (Baker, 2016). This concern appears to be grounded in reality as over 60% of respondents reported trying and failing to replicate the results of other researchers.

Many broadly accepted research practices contributed to the replication crisis. For example, there is a long history of publication bias favoring statistically significant results (Polanin, Tanner-Smith, & Hennessy, 2016; Sterling, 1959), which is itself extremely troublesome. However, biases extend to which results are included within publications as well as which findings are cited in subsequent publications (de Vries et al., 2018). In education, Pigott, Valentine, Polanin, Williams, and Canada (2013) examined published dissertations and found that less than a quarter of the published versions contained all the outcomes reported in the dissertation. The results that were omitted did not appear at random; nonsignificant dissertation results were 1.30 times less likely to be reported in the published version than statistically significant results. This selective reporting of findings creates a false impression that researchers and interventions are more consistent and successful than they actually are.

Researcher Flexibility

Another existing limitation in academic research is the flexibility researchers have when making methodological decisions. Sometimes called researcher degrees of freedom (Simmons, Nelson, & Simonsohn, 2011) or a garden of forking paths (Gelman & Loken, 2013), this flexibility allows researchers to, wittingly or unwittingly, make analytic choices to obtain a particular result. Sometimes this flexibility includes behaviors that have been labeled as *questionable research practices*, such as selectively reporting outcomes, peeking at

results to decide whether to collect additional data, determining data exclusion based on its impact on results, rounding *p* values down so that they are below relevant thresholds, and reporting unexpected results as though they were predicted (John, Loewenstein, & Prelec, 2012). Surveys of active researchers in several fields show that large percentages admit to engaging in questionable research practices (e.g., Fiedler & Schwarz, 2015; Fraser, Parker, Nakagawa, Barnett, & Fidler, 2018; John et al., 2012).

This is not to suggest that there is only one "right" way to conduct research. In any scientific endeavor, there is almost always more than one reasonable way to measure key variables (Flake & Fried, 2019), prepare the data (Steegen, Tuerlinckx, Gelman, & Vanpaemel, 2016), and fit statistical models (Silberzahn et al., 2018). However, this flexibility produces an opportunity for researchers to see many different possible results during the course of an analysis. If researcher behavior is altered by observing those results, the reported results could be selected for their desirability, whether consciously or unconsciously. When this occurs, the study's evidence is overstated, error-control properties of statistical inference are damaged, and the risk of false-positive inference becomes much higher than was intended or reported. When this process becomes normative and publications fail to describe the true process that led to the result, a replication crisis seems almost inevitable.

Lack of Statistical Power

The problems of underpowered studies have been under discussion for quite some time (e.g., Cohen, 1988), but these critiques have not been persuasive enough to jolt educational science out of its existing equilibrium. Widespread replication failures in psychology, biomedical sciences, and other fields have reasserted the problems inherent to small, underpowered studies (Lindsay, 2015). Furthermore, these techniques interact with publication bias, resulting in substantially inflated effect size estimates (Gelman & Carlin, 2014). Preregistered replication studies, in which most questionable research practices cannot be employed, have revealed that true effect sizes tend to be roughly half as large as meta-analyses of published studies indicate (S. F. Anderson & Maxwell, 2017; OSC, 2015). This means that typical social science effect sizes are likely substantially smaller than typically reported. No matter how many bricks one might have, one cannot build a robust structure if few of the bricks can support any weight. This is why many believe that even meta-analyzing the existing literature is not a credible strategy for producing trustworthy evidence (e.g., van Elk et al., 2015). In fact, a recent preprint by Maassen, van Assen, Nuijten, Olsson-Collentine, and Wicherts (2019) was unable to reproduce 224 of 500 published meta-analytic findings based on reported information. Instead, reforms like increasing statistical power must be implemented that will increase the value and quality of individual studies.

Limitations of Big Data

Education differs from some of the other social sciences because large-scale state or school district administrative data sets are often available to researchers, as are nationally representative survey data, such as High School and Beyond or the Early Childhood Longitudinal Study. These products are valuable and have stimulated a great deal of work in the field. Although the sample sizes are typically more than sufficient for adequate statistical power, the retrospective nature of these data sources means that they generally cannot be used to evaluate the effect of novel interventions. Additionally, researchers are limited to observational designs. Estimating causal effects in observational designs requires confounding to be addressed through statistical adjustment (e.g., regression adjustment or propensity score techniques), by instrumental variables, or by a regression discontinuity design (Morgan & Winship, 2014). All these procedures, and particularly statistical adjustment, require the correct identification of variable roles in the causal system as well as precise and valid measurement of those variables. However, knowledge of the causal system is often lacking, and if present would often eliminate the need for the study in the first place.

Moreover, existing data sets may not include all needed variables, or may only include noisy proxies for them. For example, socioeconomic status is a central concept in education research, but researchers have been content to accept free or reduced-price lunch status as an acceptable substitute for a direct assessment because it is easy to collect despite its well-established questionable validity as a measure of socioeconomic status or even poverty (e.g., Greenberg, 2018; Harwell, 2019; Harwell & LeBeau, 2010). In another example, Christakis, Zimmerman, DiGiuseppe, and McCarty (2004) studied the relationship between early television exposure and attention-deficit/hyperactivity disorder (ADHD) using the National Longitudinal Survey of Youth data set. In this study, ADHD (the outcome variable) was defined using an ad hoc, nonvalidated scale constructed from five items based on an arbitrary cutoff for classification. This exemplifies how the nature of big data in education often results in poor measurement and creates serious challenges for causal inference.

Poor measurement can have serious consequences. For example, when the focal variable (e.g., the X variable) is measured with error, its correlation with the response variable is attenuated (Crocker & Algina, 1986). Thus, poor measurement of a focal variable leads to a potentially severe underestimation of its relationship with the outcome. The situation becomes even more serious in observational designs, where confounding is omnipresent. In the presence of confounding, inference of the relationship between X and Y requires statistical adjustment for the confounding variables (Rohrer, 2018). However, statistical adjustment requires the confounders to be measured with high reliability and validity. When measurement of confounders is poor, a

substantial confounding effect persists even after adjustment, which can create a spurious correlation between the focal variable and the response variable. Westfall and Yarkoni (2016) extensively discussed this problem and showed in simulation that researchers would nearly always falsely conclude that *X* and *Y* are associated even after controlling for a confounder when that confounder is measured with moderate (rather than perfect) reliability. This implies that, for observational research to succeed, the mere quantity of data is not enough. The data set needs to consist of high-quality measures of not only the focal and outcome variables but also any covariates necessary to remove any confounding (Rohrer, 2018). These conditions are much more likely to hold when data collection is motivated by a deliberate process designed to answer a particular question than when the data were collected for some other reason and are repurposed by a researcher. With such limitations, it becomes clear that relying on existing "big data" will not solve all problems; instead, education researchers require large data sets that are calibrated in the service of specific questions. Pooling of data collection resources across teams is one method to achieve data sets that are not just big, but scientifically useful.

Misaligned Incentives

An additional factor contributing to the limitations of the current academic research model is the lack of alignment between the incentives motivating individual researchers and the goals of the broader field. For example, number of publications is a strong predictor for hiring and promotion (Alperin et al., 2019). Such incentives encourage researchers to focus on generating many publications. One efficient way to generate many publications is to publish with fewer participants in each paper. However, small sample sizes mean that most studies performed under this model are seriously underpowered to detect anything but strong effects (Chase & Chase, 1976; Sedlmeier & Gigerenzer, 1989; see also Schochet, 2008) and can produce only weak evidence. As a result, a field may generate many papers and claims, only a few of which address important questions with strong evidence. The philosophy seems to be that *though we are losing money on every sale, we can make it up in volume* (Miller, 1988).

For example, the What Works Clearinghouse (WWC) curates and summarizes findings from intervention studies that either use randomized treatment allocation or rigorous quasi-experimental designs and are free of excessive confounding threats or subject attrition (WWC, n.d.). Only a tiny proportion of published education work meets these standards. As one example, the WWC has summarized results from only six studies of the effect of Teach for America on mathematics achievement (see https://ies.ed.gov/ncee/wwc/Intervention/6 for details). However, a Google Scholar search for "Teach for America" and "mathematics achievement" returned 5,990 results at the time of this writing (August 2019). Even if only a tenth of the Google Scholar hits are empirical studies that should be counted, this means a mere 1% of the publications on this topic satisfied design standards required to merit WWC review. Similar (or worse) rates are observed in the other topical areas reported by the WWC.

The IES has long had a funding structure that supported diverse goals, including Exploration, Development and Innovation, Efficacy, Effectiveness, and Measurement (Brock & McLaughlin, 2018). Exploration and innovation are necessary and important precursors to subsequent research evaluating causal efficacy and effectiveness. But, in conjunction with the issues discussed above, it appears that researchers have been focusing far more heavily on exploration than on subsequent stages of research and implementation. For example, within these areas, between 2004 and 2016 IES funded over 300 Efficacy studies to generate initial causal evidence of intervention impact under ideal conditions (Chhin et al., 2018). However, IES has funded only 14 Effectiveness studies that replicated initial efforts under routine educational settings (Taylor & Doolittle, 2017). Because not all interventions "work," some winnowing is to be expected, but over 95% drop is quite steep. Such disparities in focus suggest that researchers may not be adequately incentivized to conduct the type of studies capable of generating strong evidential support.

Previously Proposed Solutions

The need to address these issues has motivated a variety of reform efforts that aim to increase transparency and rigor in standard education research practice. In 2018, three independent papers called for open science research practices in education (van der Zee & Reich, 2018), special education (Cook, Lloyd, Mellor, Nosek, & Therrien, 2018), and gifted education (McBee, Makel, Peters, & Matthews, 2018) to increase the veracity and internal validity of published research. All three papers suggested that using practices such as preregistering hypotheses, sharing data and research materials openly, and making research papers freely available as often as possible would improve the quality of research in education. A fourth paper (Gehlbach & Robinson, 2018) focused specifically on how preregistration could mitigate illusory findings in education research. These calls focus on important shifts in when various stages of work are done and how information is shared in hopes of improving alignment between practices with espoused values.

Encouragingly, analogous efforts are beginning to permeate the field, including revision of the IES funding goals to include different types of replication studies; new IES and National Science Foundation (2018) companion guidelines on replication and reproducibility, and the recently created registry of efficacy and effectiveness studies in education

(D. Anderson, Spybrook, & Maynard, 2019). We seek to build on these efforts by extending their application to large-scale collaboration.

Collaborative Research

An approach to addressing many of the aforementioned challenges is to pool resources to conduct more rigorously designed studies through large-scale collaboration. We are not the first to highlight the potential for collaborative research (e.g., Uhlmann et al, 2019). Large-scale collaboration has been tremendously successful in the physical and life sciences. Two recent examples of large collaborations in physics are the discoveries of gravitational waves by the LIGO and Virgo collaborations (Abbott et al., 2017; a paper featuring 1,011 authors) and the Higgs boson announced in separate papers by the ATLAS and CMS collaboration groups with over 3,000 authors each (Aad et al., 2012; Chatrchyan et al., 2012). Such mega-collaborations have also been used outside of physics. The 2001 paper presenting the initial sequencing of the human genome was authored by the International Human Genome Sequencing Consortium and included about 2,900 individual authors. In large-scale collaborations, many individuals work on projects to produce rigorous, highly informative findings.

In recent years, psychologists have begun to experiment with large collaborative research activities. These efforts have garnered attention both within the academy and in the popular press (e.g., Aschwanden, 2015; Yong, 2018). We believe that education research is well suited for such collaborations because of the large number of researchers and schools spread across the country and world, the difficulty of accessing schools and students for research participation, and the relatively small amount of funding available. This creates a context in which pooling resources is especially valuable. In the following sections, we introduce five types of large-scale collaboration: participating teams run different studies, multiteam collaboration, collaborative analysis, preregistered adversarial collaboration, and persistent collaboration. Each type of collaboration helps address different types of problems, provides different benefits, and has varying limitations and barriers to entry (see Table 1).

Participating Teams Run Different Studies

In this model, the collaboration runs a set of separate projects with individual research studies assigned to individual research groups (teams). The goal of this form of collaborative effort is to make inferences regarding the *set* of studies or findings. Thus, the study selection process plays a central role in determining what conclusions can be drawn from the pattern of findings across the studies in this model. We use the term *team* to mean a group of collaborating researchers. This can be a "lab" in the traditional scientific sense or a group of researchers from different institutions working together on a specific project (like the authors of this article).

The OSC is an example of this type of collaborative effort. In 2015, the OSC published the results of the Reproducibility Project: Psychology, in which 270 researchers collaborated to replicate 100 psychology studies from a variety of subdisciplines using preregistered, high-powered protocols. They succeeded in replicating only 39% of the original studies, and the estimated effect sizes were typically only half the size of the original results. A more recent example of this type of collaboration was used to investigate the reproducibility of social science experiments published in the journals *Science* and *Nature* from 2010 to 2015, and the extent to which a betting market could classify the studies as replicable or nonreplicable (Camerer et al., 2018).

This model of collaboration has been typically used to answer metascience research questions and is well suited for replication efforts but may not be effective or efficient to answer substantive research questions that already have a lot of evidence. This is because the research produced by participating teams running different studies does not require a concentration of resources at the individual study level. Thus, this type of collaboration is not suited to definitively answer new questions; it assesses the replicability of previous findings. The value of the findings produced by this model directly relate to the credibility of the individual studies. When these are preregistered, openly share data and materials, and are executed independently from the original research team, each study can be relatively trustworthy, particularly when they are direct replications of past research. In direct replications, the original design, treatment/manipulation, and measurement practices are treated as a fixed feature of the protocol rather than subject to change or optimization.

This model of collaboration could be useful to assess any set of previously established findings (e.g., in introductory textbooks, used in schools, or highly touted as needing greater implementation). However, building on Ioannidis's (2014) recommendations for how to make more published research true, the most informative projects may choose to focus on previous studies that had small samples, flexible definitions and analyses, flexible thresholds for "success," weak measurements, nondiverse participants, or few independent direct replications. Such assessment will help establish the extent to which previous results can be trusted and when.

Multiteam Collaboration Projects

A second collaborative model is multiteam collaborations. Multiteam collaborations take the principle of independent replication and apply it to a single investigation conducted by several independent research teams that have

TABLE 1
Collaborative Research Initiatives, Their Benefits, and Example Resources

Collaboration Type	Relevance	Benefits	Limitation and Barriers to Entry	Resources and Examples
Participating teams run different studies	Which existing results are replicable and generalizable?	Assess credibility of previous studies	Efforts focused on previously established findings	Open Science Collaboration: https://osf.io/vmrgu/
Multiteam collaborations	What about when a specific question or intervention needs a definitive answer?	Informs about generalizability, heterogeneity of effect	Massive resources to create Recognition for effort may not match existing incentive structure	www.manyclassess.org https://osf.io/wx7ck/ Overview of Many Labs 2: https://cos.io/our-services/research/many-labs-2-project-overview/
Collaborative analysis	What about when there are many plausible and reasonable analytic choices to answer a research question?	Makes analytic flexibility transparent and develops consensus assessment	Devotes large amount of analyst time to a single question	Many analysts, one data set: https://psyarxiv.com/qkwst/
Preregistered adversarial collaboration	What if there is disagreement within the field on how to interpret existing results?	Reduces post hoc "Whataboutism" and provides clarity regarding a program's strengths and weakness	Requires buy-in and participation of particular researchers who disagree. Researchers must be willing/able to change their view based on results	Example of Adversarial Collaboration Agreement: https://osf.io/deany Matzke, D., van Rijn, H., Wagenmakers, E. J., Slagter, H., van der Molen, M., & Nieuwenhuis, S. (2014, June 3). *The effect of horizontal eye movements on free recall performance. A purely confirmatory replication study*. Retrieved from osf.io/pxt3m
Persistent collaboration	How could education researchers organize to conduct large-scale collaborations?	Facilitate large scale data accumulation	Massive resources to create Recognition for effort may not match current incentive structure	PsyAccelerator: https://psysciacc.org StudySwap: https://osf.io/view/StudySwap/

all agreed to follow a common protocol. The general premise of such an approach is that multiple teams simultaneously conducting the same study will provide a more definitive answer to whether a finding exists and how variable it might be across contexts (i.e., whether it is generalizable). Multiteam collaborations must agree on a shared protocol (i.e., participants, method, analyses) prior to data collection. If posted publicly, this serves as a preregistration and limits the problems associated with researcher degrees of freedom discussed previously. After the protocol is implemented, the findings from each team are reported individually as well as in aggregate to show the variability in effect.

Multiteam collaborations do not have to be replications; they can test new questions or interventions. That said, testing previously established findings that are believed to have value can help focus what merits attention and resources. One multiteam collaboration tested 12 classic psychological findings (13 effects) to see if previously published findings could be replicated as well as assess the variability of the effect across samples (Klein et al., 2014). Each team followed the same specified protocol, collected data from at least 80 participants, recorded a video of their administration procedures, and documented any deviation from the established protocol. In the end, there were 36 samples with 6,344 total participants. Not every research team conducted each of the 12 studies, but every study included multiple teams, thereby evaluating intrateam variance in the effects.

Multiteam collaborations report the variability in effects across sites and researchers. For example, in the Klein et al. (2014) article, the authors found support for 10 of the 13 tested effects. One particular effect, Currency Priming (exposing people to money makes them more likely to endorse a system or policy), did not replicate and showed a consistent effect size clustered around zero (the original study effect was closer to $d = 1.0$). Compare this with Anchoring (estimating size or distance after first being presented with implausible values), which replicated, but showed a far wider range of effect (d ranged from <1.0 to

>3.0) than previously reported. Because multiple teams collaborated on a preregistered protocol, their results are likely more representative of expected findings others would get. Such expectation setting can be useful in education, where expected findings in specific settings or for specific populations are of interest to practitioners.

Multiteam collaborations may sound similar to traditional applications of meta-analysis, a tool to aggregate an existing body of work. Multiteam collaborations often make use of meta-analytic methods (e.g., weighted averaging of effect sizes, aggregating results of multiple studies), but because traditional meta-analyses include studies conducted across time, often by researchers applying different interventions with differing applications or protocols, and strongly affected by publication bias, the ability of meta-analysis to identify the true effect is limited (Fyfe, de Leeuw, Carvalho, Goldstone, & Motz, 2019). Traditional meta-analyses can rarely answer the question of how much of the variability in results or in replicability can be attributed to sample, researcher, statistical power, or bias (e.g., Lakens, Hilgard, & Staaks, 2016).

Research on the instructional strategy of Reciprocal Teaching is a good example of how challenging it can be to produce credible meta-analytic estimates of results. In Reciprocal Teaching, a cognitive strategy is modeled with the goal of students applying the strategy to novel content. Reciprocal Teaching strategies often take the form of summarizing, questioning, and predicting in analyzing text. Two meta-analyses found mean effects of Reciprocal Teaching to be $d = 0.32$ to 0.88 (Galloway, 2003; Rosenshine & Meister, 1994). However, sometimes the intervention was delivered in small groups as opposed to whole classes. In some studies, the outcome measure was a state standardized test, whereas in others it was a teacher-made test. These are all variations in protocol or intervention application whose unique effects cannot be tested via meta-analysis unless multiple studies implemented the same protocol. Multiteam collaborations control for extraneous factors to systematically estimate the true effect and the conditions under which that effect manifests.

By having multiple research sites and collaborators, a single multiteam "study" can reveal what factors influence the effect better than a meta-analysis of independently conducted existing studies. Fyfe et al. (2019) provide a relevant education example where the authors are using a multiteam approach to test the effect of timing of instructor feedback on subsequent class performance. Instead of using a single class with the same curriculum, the authors solicited a wide range of classes and disciplines. This allows the authors to examine the degree to which contextual factors such as assignment length, discipline, or class size influenced the observed effect. Such an approach is useful for educators and also more credible because it tested a theory across a wide range of contexts.

Collaborative Analysis

The benefits of collaboration can also be leveraged through a third collaborative model that focuses specifically within the data analysis stage of the research process. In this collaborative analysis approach, multiple teams independently analyze the same data set to answer the same research question. Similar to a specification curve (Simonsohn, Simmons, & Nelson, 2015) or multiverse analysis (Steegen et al., 2016) conducted by an individual or single research team, this model helps reduce the impact of particular analytic choices by making them transparent. This also reduces researcher degrees of freedom because no single analytic choice by an individual researcher will have as strong an impact on the final results. Collaborative analysis is relevant because recent metascience research has shown the extent to which data analytic choices influence subsequent research results. This influence can occur even in the absence of *p*-hacking or other questionable research practices (Silberzahn et al., 2018). Every decision made concerning missing data, outlier exclusions, assumption testing, variable aggregation and transformation, model selection, covariates, and so on comes with a host of alternative decisions that *could have* been made, any number of which might have been equally reasonable. One way to assess (and quantify) the magnitude of this variability is by having multiple analysts interrogate the same data set while testing the same hypothesis—all with an eye toward understanding how differing analytic choices influence the results.

For example, Silberzahn et al. (2018) reported findings from a project in which 29 independent research teams analyzed the same data to determine whether soccer referees gave more red cards to dark-skinned players. Teams differed substantially in the choices they made, resulting in 20 statistically significant and nine null findings. Effect sizes ranged from moderately large to practically nil, although confidence intervals mostly overlapped with each other. Replacing "red cards" with educational outcomes like special education placement, suspensions, or expulsion would make such an analysis immediately relevant to education.

Collaborative analysis of a data set can also be combined with the Registered Report format, where a study's literature review, method, and analytic plan are reviewed prior to data collection (see https://cos.io/rr/). Hussey et al. (2018) announced the release of a massive data set that includes over 444,000 observations from 200,000 participants on 15 commonly used individual difference measures. The authors made 15% of the data set publicly available for exploratory analysis, along with the data code books and collection procedures. Research teams can use the exploratory data to help form hypotheses and data analysis plans that can then be submitted to a journal as a Registered Report. If granted in-principle acceptance, researchers will be given access to the rest of the data to conduct confirmatory analyses. As of

February 2019, 14 journals have agreed to accept Registered Reports based on these data.

In education, the collaborative analysis model could be applied to data routinely collected within schools for state, national, and international assessments. For example, a journal could put out a call for Registered Reports that would rely on the next release of state assessment, PISA (Programme for International Student Assessment), or NAEP (National Assessment of Educational Progress) data. Alternatively, researchers could agree on a set of hypotheses to test with these data sets before data are released. Additionally, collaborative analysis could extend to qualitative research; multiple analysts can facilitate triangulation of thematic or theoretical codes drawn from interview transcripts, open-ended survey responses, classroom observations, videos, social media posts, or student work products.

Because processing and modeling options abound for education data, using collaborative analysis can help ensure results are robust across those decisions and across researchers. Collaborative analysis clarifies what is not known by revealing the uncertainty of findings resulting from analytic flexibility. That said, to be successful, collaborative analysis requires multiple independent competent analysts. As we discuss below, to acquire this level of participation likely requires changes in incentive structure (i.e., Is authorship sufficient incentive?).

Preregistered Adversarial Collaborations

The previously discussed collaborative models are built on researchers agreeing on many fundamental premises, such as what worked and what is worth testing going forward. Clearly, this is not always the case. The following collaborative research model is built on preexisting *disagreement* within the field. A preregistered adversarial collaboration (PAC) is a joint research project conducted by individuals or research teams who disagree about an important theoretical or empirical question with the goal of resolving the disagreement. This kind of collaboration was suggested by Latham, Erez, and Locke in 1988 and was more recently advocated by Kahneman (2003) as an alternative to the protracted stalemate of the critique-reply-rejoinder method seen in academic journals. PACs are particularly useful in the field of education, where long-running disputes stretch across years, if not decades. For example, whether class size makes a major difference in student achievement has been investigated, and disagreed on, for decades (Li & Konstantopoulos, 2017; Slavin, 1990; Woods, 2015). Rather than the disagreeing parties serially waiting to point out the flaws of the other side's research, the two groups agree on what would (or could) be informative a priori, thus embracing the potential of resolving a disagreement that could otherwise drag on for years.

PAC collaborators should have a sincere desire to address differences, be objective, and be open to change (Latham et al., 1988). PACs begin with a discussion of the major theoretical differences and possible areas of disagreement to generate hypotheses (Latham et al., 1988). An initial meeting could take place during a national conference to brainstorm possible areas that could lead to collaborative research. All collaborators should agree that furthering knowledge should take precedence over any personal or professional motivations. It can feel like a professional risk when facing an antagonist, particularly if there are deep theoretical differences. For that reason, it is advisable to involve an individual who can arbitrate differences that arise (Kahneman, 2003; Mellers, Hertwig, & Kahneman, 2001). This arbitrator should be sufficiently knowledgeable about the theories and methods involved and be someone who collaborators trust to be impartial (Neir & Campbell, 2013).

The initial discussion is a success if it generates testable hypotheses about important issues (Mellers et al., 2001). It is possible that disagreements have been overstated. Therefore, collaborators should first tease apart rival explanations (Kerr, Ao, Hogg, & Zhang, 2018). For example, disagreements on the value of mindset interventions include differences regarding what constitutes an important effect size, different perspectives regarding whether mindsets are situational or universal, and disagreement about whether a focus on achievement rather than growth can be advantageous in some situations (DeWitt, 2017; Dweck, 2019; Elliot, 2019).

Next, collaborators create a comprehensive study protocol that includes design, data collection, analysis, and reporting of findings. Exhaustive planning is important to ensure true collaboration and minimize the researcher flexibility discussed above. Because the protocol is preregistered, collaborators are motivated to provide constructive criticism that can produce higher quality methods and analysis when it can still be useful—prior to data collection. This planning also helps prevent researchers from explaining away unexpected or undesired results as a lack of fidelity (Kaimal & Jordan, 2016; Missett & Foster, 2015). For example, it has been difficult to determine the effectiveness of integration of instructional technology in the classroom because of variance in commitment to learning new teaching methods (Delgado, Wardlow, McKnight, & O'Malley, 2015; Nicol, Owens, Le Coze, MacIntyre, & Eastwood, 2018). For a PAC to work, collaborators must define a priori the sufficient level of fidelity and how it will be measured.

Collaborators should also anticipate possible interpretations of the outcomes and explicitly identify what kind of results would be consistent/inconsistent with their expectations (Mellers et al., 2001). Namely, what results would change or falsify beliefs? If no result would falsify beliefs (e.g., a fundamental disagreement about the role of public education in society) it is not a scientific disagreement and empirical assessment would not be fruitful. PACs help prevent researchers from overselling results because researchers can avoid becoming entrenched in a single perspective

(Jussim, Crawford, Anglin, Stevens, & Duarte, 2016), but expectations should be reasonable. The process begins with a substantial disagreement, and it is possible that there will not be a complete resolution. Discussion of these disagreements are often part of the coauthored paper (Mellers et al., 2001). Although the preregistered protocol should eliminate post hoc interpretation of effects, thinking about post hoc explanations is unavoidable. Researchers can agree to jointly plan follow-up research based on the results of the initial collaboration. The process can help narrow differences and increase mutual respect. In fact, PAC was recently used to address the differing explanations posited by the original researchers and the replicators for a failed replication of studies suggesting that the experience of emotion is affected by one's own facial expressions (i.e., smiling will cause you to feel happy; Coles et al., 2019). Bringing diverse perspectives to the table at the research design stage ensures that concerns from multiple sides are incorporated into the research process.

Numerous educational issues could benefit from PACs. For example, the implementation of restorative justice is gaining popularity in school systems worldwide, but research is far behind practice (Song & Swearer, 2016). Restorative justice as applied to schools is a philosophy of discipline that focuses on rebuilding relationships in the school community by bringing together offenders and those affected by an infraction to decide together how to repair the harm caused to others (Suvall, 2009; Zehr, 2015). Although some scholars and policymakers support the potential of restorative justice (Pavelka, 2013; Wearmouth, Mckinney, & Glynn, 2007; Zehr, 2015), reports from school districts suggest that restorative justice can negatively affect school climate, safety, and staff morale (Augustine et al., 2018; Eden, 2019; Gray et al., 2017). Areas of disagreement regarding restorative justice include the relative importance of prescriptive policies, requirement of adherence to philosophical values, the level of necessary implementation in a school, and the role that restorative justice plays in addressing issues of racial equality (Anfara, Evans, & Lester, 2013; Morrison & Vaandering, 2012; Song & Swearer, 2016).

Rather than debate for decades while millions of students age through the system, researchers supporting the potential of restorative justice can collaborate with those who are skeptical to subject their perspectives to an empirical test through the production of joint research under an agreed on protocol. The researchers would define ahead of time what type of results adequately address their differences in empirical perspective. Differences in values (e.g., How much must GPA increase to make an intervention worthwhile?) may still exist across researchers. However, if PACs can shift the conversation from the empirical question "Is there evidence of an effect?" to the policy and value question of "Is it worth it?" then we believe science will have done its job.

One concern may be that few researchers would be willing to risk this type of collaboration. Many scholars establish themselves based on their work studying a particular theory or intervention, and this leads to presentations, books, and grant funding (Lilienfeld, 2017). If the effect of their research is found to be less powerful than they have presented, prestige can be lost. We believe greater numbers of stakeholder groups, who have different roles or viewpoints, engaging with each other would generate numerous benefits that outweigh the risks of having one's work scrutinized. A PAC can clarify under what conditions an educational intervention is effective (Jussim et al., 2016) resulting in improved implementation and stewardship of resources. We believe that funding agencies should strongly consider earmarking resources to PACs as a means of incentivizing them.

Persistent Collaboration

As mentioned above, the creation of large-scale collaborations may be difficult, particularly for researchers who are not well established within the field and/or those at institutions with fewer research resources. However, there are new initiatives within psychology that may serve as a fifth type of collaborative model that could be useful to leverage in education research: persistent collaboration infrastructure. PsyAccelerator (https://psysciacc.org; Moshontz et al., 2018) is a network of psychology research labs (over 500 as of August 2019) from 60 countries from every populated continent. According to its website, its mission is "to accelerate the accumulation of reliable and generalizable evidence in psychological science, reducing the distance between truth about human behavior and mental processes and our current understanding." Its founder views PsyAccelerator as psychology's equivalent of CERN, home of the Large Hadron Collider mentioned previously (Chartier, 2017). In essence, it is an infrastructure to create multiteam collaborations to facilitate persistent large-scale collaboration. With an infrastructure to support multiteam collaborations, the organizational effort required for each multiteam collaboration shrinks.

Similarly, StudySwap (https://osf.io/view/StudySwap/) is a platform on which researchers can offer and request assistance in data collection (either whole studies or additional participants; Chartier, Riegleman, & McCarthy, 2018). Using StudySwap, researchers can post "Haves" or "Needs" as part of an online exchange. Both StudySwap and PsyAccelerator help crowdsource and coordinate resources so that an individual researcher is not responsible for all aspects of the research cycle (or collecting all data).

For an EduAccelerator, imagine 50 research teams across the country forming a research consortium. The goal of the consortium is not to conduct a specific study or answer a specific question but to create an infrastructure in which those things can happen more easily and at a larger scale. A

StudySwap for education research could facilitate data collection from hard to reach populations that may not be concentrated within schools or districts (e.g., valedictorians, students who've been suspended, students with low-incidence disabilities such as visual impairments). They could help provide results that help local policymakers understand not just "Does this intervention work?" but how that intervention works in specific educational contexts (e.g., high-poverty urban schools or rural schools).

Broadly, persistent collaboration consortia could help more efficiently allocate existing resources, reduce barriers to entry, and increase inclusivity, while also increasing transparency, rigor, and reliability of the results produced (Uhlmann et al., 2019). Establishing consortia provides many advantages, including larger sample sizes, more diversity among potential participants, increased depth and breadth of available expertise, and greater inclusivity within the research process. Persistent collaborative infrastructure can help assure that the sample size of any study can be much larger than what individual researchers may be able to acquire. This can help with many facets of research design, including statistical power and precision of estimates. Another advantage is several forms of diversity. Diversity of potential participants (e.g., geography, age, demographic, rurality) can help assess generalizability or whether there are moderating factors of any result. For example, an individual researcher may only collect data in one state or one school district, thus limiting our understanding of whether we should expect results to generalize to other contexts. A consortium can facilitate assessment of generalizability and the existence of potential moderators on relevant results (see Fyfe et al., 2019, example above).

Diversity of expertise allows the research conducted to cover a broader range of topics and use a broader range of methods because the group can draw from a pool of expertise when selecting topics and designing studies. Not every consortia member need be an expert on every aspect of any study being conducted. This diversity can create massive scale of efficiency in the types and amount of work that can be done, enabling individual researchers to contribute their unique perspective to any given consortia project. Moreover, any researcher, regardless of career status or university affiliation, is able to join and make contributions to the project, even if they are not providing data. This type of crowdsourcing makes the research process more inclusive of researchers with limited resources to conduct studies on their own.

Together, these facets of persistent collaboration infrastructure have the potential to improve the quality, precision, and generalizability estimates of research being produced while simultaneously accelerating the rate at which this improved knowledge is developed. All that said, creating the infrastructure would take substantial investment of resources and shared will across many stakeholders. Regardless, the payoff over time would be at both the micro (individual consortia member) and the macro (across the consortia) levels. Each contributing school learns something about itself while also contributing to a larger project that indicates whether the results from any individual consortia member are the exception or the norm. Knowing how their specific context fits within the larger spectrum of results could have immense value for schools. It could help districts select and prioritize specific interventions as well as set realistic expectations for effects of those interventions in their specific environment. In noncrowdsourced research, constraints on generality of results (Simons et al., 2017) can be guesswork. In crowdsourced research, diverse consortia can directly investigate constraints on generality as part of the typical process.

The Future of Collaborative Education Research

In this article, we reviewed how research has not always lived up to its scientific aspirations and proposed that education research would benefit from greater adoption of collaborative research methods. It is our belief that if the education research community were to adopt these collaborative models more frequently, it will more credibly answer research questions and earn the trust of practitioners and policymakers. Positive strides have already been made in education research. For example, the "ManyClasses" project (www.manyclasses.org) is an application of the multiteam collaboration concept across K–12 classrooms. Their stated goal is to "examine the same research question in dozens of contexts, spanning a range of courses, institutions, formats, and student populations." Similarly, after finding a "developer effect" where studies included in the WWC that had been commissioned by the intervention developer averaged 1.5 times the effect size of independently conducted evaluations, Wolf, Morrison, Slavin, and Risman (2019) suggested that requiring preregistration should become necessary for inclusion in the WWC.

We are not proposing that every project must become large-scale collaboration (although what a great learning opportunity for student training![1]). The role and importance of small-scale studies developing and testing new ideas remains relevant. That said, we believe that the more important the issue, the more resources we should devote to understanding what it is, what causes it, and what can expand, reduce, or remove it, as necessary. Obviously, "importance" is not universally agreed on. What may be important to some may be less important to others. Regardless, like the current research model, individual researchers, schools, and funders would all still be able to rely on their own value systems to determine which collaborative research projects they choose to participate in. Factors such as personal interest or relevance, expected magnitude of effect, number of students affected, and many others would all continue to play important roles when selecting research project priority. The primary change would be in the scale collaborative research projects undertake, leading to an expected synergistic return in the value derived from the subsequent results.

Which model of large-scale collaboration is preferred depends on the goals of the research team as well as the state of the research community. For example, if the goal were to assess which existing findings can be replicated by independent research teams, then the first model we discussed, of a coordinated effort of participating teams running different studies, would be beneficial. When different research teams have already reported differing results, multiteam collaboration may help advance the conversation while deepening understanding beyond "if" an effect exists to include "when/where" an effect exists. However, if entrenched disagreement exists within a research community, a multiteam collaboration may not be sufficient to resolve the debate and a PAC may be preferable. For example, there have not only been numerous studies, but numerous meta-analyses on the effects of ability grouping on students with essentially polar opposite conclusions made by various authors over several decades (e.g., Slavin, 1987; Steenbergen-Hu, Makel, & Olszewski-Kubilius, 2016). In this case, a PAC could help bring the various sides together to advance toward agreement.

If the education research community (or a subset of it) finds value in conducting more large-scale collaborative research, then creating an infrastructure to support persistent collaboration across studies may yield lucrative returns. This infrastructure could serve as the foundation for multiteam collaborations as well as collaborative analyses. Multiteam collaborations may be preferred when generality of a finding is in question, whereas collaborative analysis may be preferred when data are hard to collect or analytic decisions are flexible. In this way, large-scale collaborative research is no different from traditional research projects; it is up to individual community members to determine what project they feel would advance the field.

Constraints of Application and Impact

The proposed practices will not solve all problems, but that does not mean that education research would not benefit from their wider application. To be successful, such initiatives will need to be relevant and appropriately matched to a field's needs. Moreover, because they are new, bumps in the road should be expected as applicability to education research is assessed. Which projects are best suited for these types of projects is up for debate; it may be best to start with relatively low-hanging opportunities so that all involved can learn the process before advancing to more complicated applications or topics.

As discussed elsewhere (Uhlmann et al., 2019), we envision collaborative research serving as a complementary addition to traditional independent research. Substantial exploratory, descriptive, and confirmatory work would still be conducted independently, although all could also be the focus of collaborative research. For example, large-scale collaborative qualitative projects could assess the replicability of some qualitative work (e.g., Do independent researchers analyzing the same data develop similar themes? If not, what does that say about the findings?).

Importantly, adoption of large-scale collaborative research in education will require a shift in incentives. For example, many of the actions we describe do not map neatly onto traditional metrics used for faculty evaluation (Ortiz, Haviland, & Henriques, 2017). Helping to create the infrastructure to facilitate something like ManyClasses is harder for a promotion committee to judge than number of publications and conference presentations. Additionally, there is currently little promotion-related benefit from being part of a multiteam collaboration that cannot also be had for the cost of a much smaller, individual study. The latter may also take less time and generate more attention for the individual researcher—critical considerations in most hiring and tenure decisions. We do not think these challenges are insurmountable, though it will take creative thinking and support from multiple stakeholders. For example, senior scholars can revise hiring, promotion, and tenure criteria to support collaborative science and open science practices more broadly (e.g., Nosek, 2017). As research norms change, those providing external review letters for tenure and promotion packets will be better able to evaluate collaborative work. Without buy-in from researchers across all career stages, as well as practitioners, journals, and funders, expanding to include more large-scale collaborative research will be quite difficult, as many of the collaborative methods discussed in this article flip the existing research model on its head. Regardless, we believe that the knowledge gained and subsequent returns to students would exceed the additional effort required to implement these approaches.

Conclusion

In education, we want students to transfer what they learn in school to other life situations, to work well with others, and to learn from observing the actions of their peers. As researchers, we should hold ourselves to (at least) the same standard. Other fields are developing large-scale collaborative models to improve research quality; education can benefit from their efforts. We believe greater implementation of large-scale collaboration has the potential to provide valuable information about direction, magnitude, and generalizability of effects in an efficient manner.

ORCID iDs

Matthew C. Makel https://orcid.org/0000-0002-3837-0088
Kendal N. Smith https://orcid.org/0000-0002-2266-7901
Matthew T. McBee https://orcid.org/0000-0001-9122-1658
Scott J. Peters https://orcid.org/0000-0003-2459-3384
Erin M. Miller https://orcid.org/0000-0002-9448-037X

Note

1. For an example of a large-scale collaborative research project involving undergraduate students, see https://osf.io/wfc6u/wiki/home/.

References

Aad, G., Abajyan, T., Abbott, B., Abdallah, J., Abdel Khalek, S., Abdelalim, A. A., . . . Zwalinski, L. (2012). Observation of a new particle in the search for the Standard Model Higgs boson with the ATLAS detector at the LHC. *Physics Letters B*, *716*, 1–29. doi:10.1016/j.physletb.2012.08.020

Abbott, B. P., Abbott, R., Abbott, T. D., Acernese, F., Ackley, K., Adams, C., . . . Affeldt, C. (2017). GW170817: Observation of gravitational waves from a binary neutron star inspiral. *Physical Review Letters*, *119*, Article 161101. doi:10.1103/PhysRevLett.119.161101

Alperin, J. P., Nieves, C. M., Schimanski, L. A., Fischman, G. E., Niles, M. T., & McKiernan, E. C. (2019). Meta-research: How significant are the public dimensions of faculty work in review, promotion and tenure documents? *eLife*, *8*, e42254. doi:10.7554/eLife.42254

Anderson, D., Spybrook, J., & Maynard, R. (2019). REES: A registry of efficacy and effectiveness studies in education. *Educational Researcher*, *48*, 45–50. doi:10.3102/0013189X18810513

Anderson, S. F., & Maxwell, S. E. (2017). Addressing the "Replication crisis": Using original studies to design replication studies with appropriate statistical power. *Multivariate Behavioral Research*, *52*, 305–324. doi:10.1080/00273171.2017.1289361

Anfara, V. A., Evans, K. R., & Lester, J. N. (2013). Restorative justice in education: What we know so far. *Middle School Journal*, *44*(5), 57–63, doi:10.1080/00940771.2013.11461873

Aschwanden, C. (2015). *Science isn't broken: It's just a hell of a lot harder than we give it credit for*. Retrieved from https://fivethirtyeight.com/features/science-isnt-broken/

Augustine, C. H., Engberg, J., Grimm, G. E., Lee, E., Wang, E. L., Christianson, K, & Joseph, A. A. (2018). *Can restorative practices improve school climate and curb suspensions? An evaluation of the impact of restorative practices in a mid-sized urban school district*. Santa Monica, CA: RAND Corporation. Retrieved from https://www.rand.org/pubs/research_reports/RR2840.html

Baker, M. (2016). 1,500 scientists lift the lid on reproducibility. *Nature*, *533*, 452–454. doi:10.1038/533452a

Begley, C. G., & Ellis, L. M. (2012). Raise standards for preclinical cancer research. *Nature*, *483*, 531–533. doi:10.1038/483531a

Brock, T., & McLaughlin, J. (2018, June 28). Building evidence: Changes to the IES goal structure for FY 2019 [Blog post]. Retrieved from https://ies.ed.gov/blogs/research/post/building-evidence-changes-to-the-ies-goal-structure-for-fy-2019

Camerer, C. F., Dreber, A., Holzmeister, F., Ho, T.-H., Huber, J., Johannesson, M., . . . Wu, H. (2018). Evaluating the replicability of social science experiments in Nature and Science between 2010 and 2015. *Nature Human Behaviour*, *2*, 637–644. doi:10.1038/s41562-018-0399-z

Chartier, C. (2017, August 26). *Building a CERN for psychological science*. Retrieved from https://christopherchartier.com/2017/08/26/building-a-cern-for-psychological-science/

Chartier, C. R., Riegleman, A., & McCarthy, R. J. (2018). StudySwap: A platform for interlab replication, collaboration, and resource exchange. *Advances in Methods and Practices in Psychological Science*, *1*, 574–579. doi:10.1177/2515245918808767

Chase, L. J., & Chase, R. B. (1976). A statistical power analysis of applied psychological research. *Journal of Applied Psychology*, *61*, 234–237. doi:10.1037/0021-9010.61.2.234

Chatrchyan, S., Khachatryan, V., Sirunyan, A. M., Tumasyan, A., Adam, W., Aguilo, E., . . . Fabjan, C. (2012). Observation of a new boson at a mass of 125 GeV with the CMS experiment at the LHC. *Physics Letters B*, *716*, 30–61. doi:10.1016/j.physletb.2012.08.021

Chhin, C. S., Taylor, K. A., & Wei, W. S. (2018). Supporting a culture of replication: An examination of education and special education research grants funded by the Institute of Education Sciences. *Educational Researcher*, *47*, 594–605. doi:10.3102%2F0013189X18788047

Christakis, D. A., Zimmerman, F. J., DiGiuseppe, D. L., & McCarty, C. A. (2004). Early television exposure and subsequent attentional problems in children. *Pediatrics*, *113*, 708–713. doi:10.1542/peds.113.4.708

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Lawrence Erlbaum.

Coles, N. A., March, D. S., Marmolejo-Ramos, F., Banaruee, H., Butcher, N., Cavallet, M., . . . Marozzi, M. (2019). The many smiles collaboration: A multi-lab foundational test of the facial feedback hypothesis. doi:10.31234/osf.io/cvpuw

Cook, B. G., Lloyd, J. W., Mellor, D., Nosek, B. A., & Therrien, W. J. (2018). Promoting open science to increase the trustworthiness of evidence in special education. *Exceptional Children*, *85*, 104–118. doi:10.1177/0014402918793138

Crede, M., Tynan, M. C., & Harms, P. D. (2017). Much ado about grit: A meta-analytic synthesis of the grit literature. *Journal of Personality and Social Psychology*, *113*, 492–511. doi:10.1037/pspp0000102

Crocker, L. C., & Algina, J. (1986). *Introduction to classical and modern test theory*. Belmont, CA: Wadsworth.

de Vries, Y. A., Roest, A. M., de Jonge, P., Cuijpers, P., Munafò, M. R., & Bastiaansen, J. A. (2018). The cumulative effect of reporting and citation biases on the apparent efficacy of treatments: The case of depression. *Psychological Medicine*, *48*, 2453–2455. doi:10.1017/S0033291718001873

Delgado, A. J., Wardlow, L., McKnight, K., & O'Malley, K. (2015). Educational technology: A review of the integration, resources, and effectiveness of technology in K-12 classrooms. *Journal of Information Technology Education: Research*, *14*, 397–416. doi:10.28945/2298

DeWitt, P. (2017, June 28). Misinterpreting the growth mindset: Why we're doing students a disservice [Blog post]. Retrieved from https://blogs.edweek.org/edweek/finding_common_ground/2017/06/misinterpreting_the_growth_mindset_why_were_doing_students_a_disservice.html

Dweck, C. S. (2019). *APS-David Myers distinguished lecture on the science and craft of teaching psychological science: Leading students toward contribution to society*. Paper presented at the meeting of the Association for Psychological Science, Washington, DC. Retrieved from https://www.psychologicalscience.org/members/teaching/aps-david-myers

Eden, M. (2019, January 14). *Restorative justice isn't working, but that's not what the media is reporting*. Retrieved from https://

fordhaminstitute.org/national/commentary/restorative-justice-isnt-working-thats-not-what-media-reporting

Elliot, A. (2019, May). *Competition and achievement outcomes: A hierarchical motivational analysis*. Paper presented at the 31st APS annual convention program, Washington, DC. Retrieved from https://www.psychologicalscience.org/convention/pdf/aps19/APS_2019_Program_Book.pdf

Fiedler, K., & Schwarz, N. (2015). Questionable research practices revisited. *Social Psychological & Personality Science*, *7*, 45–52. doi:10.1177/1948550615612150

Flake, J. K., & Fried, E. I. (2019). Measurement schmeasurement: Questionable measurement practices and how to avoid them. *Advances in Methods and Practices in Psychological Science*. doi:10.31234/osf.io/hs7wm

Fraser, H., Parker, T., Nakagawa, S., Barnett, A., & Fidler, F. (2018). Questionable research practices in ecology and evolution. *PLoS One*, *13*, e0200303. doi:10.1371/journal.pone.0200303

Fyfe, E., de Leeuw, J. R., Carvalho, P. F., Goldstone, R., & Motz, B. (2019). *ManyClasses 1: Assessing the generalizable effect of immediate versus delayed feedback across many college classes*. doi:10.31234/osf.io/4mvyh

Galloway, A. M. (2003). *Improving reading comprehension through metacognitive strategy instruction: Evaluating the evidence of effectiveness of the reciprocal teaching procedure* (Doctoral dissertation). Available from ETD collection for University of Nebraska–Lincoln. (UMI No. AAI3092542)

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

Gelman, A., & Carlin, J. (2014). Beyond power calculations: Assessing Type S (sign) and Type M (magnitude) errors. *Perspectives on Psychological Science*, *9*, 641–651. doi:10.1177/1745691614551642

Gelman, A., & Loken, E. (2013). *The garden of forking paths: Why multiple comparisons can be a problem, even when there is no "fishing expedition" or "p-hacking" and the research hypothesis was posited ahead of time*. Unpublished manuscript. Retrieved from http://www.stat.columbia.edu/~gelman/research/unpublished/p_hacking.pdf

Gray, M., Sirinides, P. M, Fink, R, Flack, A., DuBois, T., Morrison, K., & Hill, K. (2017). *Discipline in context: Suspension, climate, and PBIS in the School District of Philadelphia* (CPRE Research Reports). Retrieved from http://repository.upenn.edu/cpre_researchreports/106

Greenberg, E. (2018). *New measures of student poverty: Replacing free and reduced-price lunch status based on household forms with direct certification*. Retrieved from https://www.urban.org/research/publication/new-measures-student-poverty

Harwell, M. (2019). Don't expect too much: The limited usefulness of common SES measures. *Journal of Experimental Education*, *87*, 353–366, doi:10.1080/00220973.2018.1465382

Harwell, M., & LeBeau, B. (2010). Student eligibility for a free lunch as an SES measure in education research. *Educational Researcher*, *39*, 120–131. doi:10.3102/0013189X10362578

Hedges, L. V., & Schauer, J. M. (2019). More than one replication study is needed for unambiguous tests of replication. Advance online publication. *Journal of Educational and Behavioral Statistics*. doi:10.3102/1076998619852953

Hussey, I., Hughes, S., Lai, C. K., Ebersole, C. R., Axt, J. R., & Nosek, B. A. (2018). *Attitudes 2.0: A large dataset for investigating relations among implicit and explicit attitudes and identity* [Data file and code book]. Retrieved from https://osf.io/pcjwf

Institute for Education Sciences. (2018). *Request for applications: Education research grants* (CDFA Number 84.305A). Retrieved from https://ies.ed.gov/funding/pdf/2019_84305A.pdf

Institute for Education Sciences & National Science Foundation. (2018). *Companion guidelines on replication & reproducibility in education research*. Retrieved from https://ies.ed.gov/pdf/CompanionGuidelinesReplicationReproducibility.pdf

International Human Genome Sequencing Consortium. (2001). Initial sequencing and analysis of the human genome. *Nature*, *409*, 860–921. doi:10.1038/35057062

Ioannidis, J. P. A. (2005). Why most published research findings are false. *PLoS Medicine*, *2*, e214. doi:10.1371/journal.pmed.0020124

Ioannidis, J. P. A. (2014). How to make more published research true. *PLoS Medicine*, *11*, e1001747. doi:10.1371/journal.pmed.1001747

John, L. K., Loewenstein, G., & Prelec, D. (2012). Measuring the prevalence of questionable research practices with incentives for truth telling. *Psychological Science*, *23*, 524–532. doi:10.1177/0956797611430953

Jussim, L., Crawford, J. T., Anglin, S. M., Stevens, S. T., & Duarte, J. L. (2016). Interpretations and methods: Towards a more effectively self-correcting social psychology. *Journal of Experimental Social Psychology*, *66*, 116–133. doi:10.1016/j.jesp.2015.10.003

Kahneman, D. (2003). Experiences in collaborative research. *American Psychologist*, *58*, 723–730. doi:10.1037/0003-066X.58.9.723

Kaimal, G., & Jordan, W. J. (2016). Do incentive-based programs improve teacher quality and student achievement? An analysis of implementation in 12 urban charter schools. *Teachers College Record*, *118*, Article ID 20450.

Kerr, N. L., Ao, X., Hogg, M. A., & Zhang, J. (2018). Addressing replicability concerns via adversarial collaboration: Discovering hidden moderators of the minimal intergroup discrimination effect. *Journal of Experimental Social Psychology*, *78*, 66–76. doi:10.1016/j.jesp.2018.05.001

Klein, R. A., Ratliff, K. A., Vianello, M., Adams, R. B., Bahnik, S., Bernstein, M. J., . . . Nosek, B. A. (2014). Investigating variability in replicability: A "many labs" replication project. *Social Psychology*, *45*, 142–152. doi:10.1027/1864-9335/a000178

Lakens, D., Hilgard, J., & Staaks, J. (2016). On the reproducibility of meta-analyses: Six practical recommendations. *BMC Psychology*, *4*, Article No. 24. doi:10.1186/s40359-016-0126-3

Latham, G. P., Erez, M., & Locke, E. A. (1988). Resolving scientific disputes by the joint design of crucial experiments by the antagonists: Application to the Erez-Latham dispute regarding participation in goal setting. *Journal of Applied Psychology*, *73*, 753–772. doi:10.1037/0021-9010.73.4.753

Li, W., & Konstantopoulos, S. (2017). Does class-size reduction close the achievement gap? Evidence from TIMSS 2011. *School Effectiveness and School Improvement*, *28*, 292–313. doi:10.1080/09243453.2017.1280062

Lilienfeld, S. O. (2017). Psychology's replication crisis and the grant culture: Righting the ship. *Perspectives on Psychological Science*, *12*, 660–664. doi:10.1177/1745691616687745

Lindsay, D. S. (2015). Replication in psychological science. *Psychological Science*, *26*, 1827–1832. doi:10.1177/0956797615616374

Lortie-Forgues, H., & Ingles, M. (2019). Most rigorous large-scale educational RCTs are uninformative: Should we be concerned? *Educational Researcher*, *48*, 158–166. doi:10.3102/0013189X19832850

Maassen, E., van Assen, M. A. L. M., Nuijten, M. B., Olsson-Collentine, A., & Wicherts, J. M. (2019). *Investigating the reproducibility of meta-analyses in psychology*. doi:10.31234/osf.io/g5ryh

Makel, M. C., & Plucker, J. A. (2014). Facts are more important than novelty: Replication in the education sciences. *Educational Researcher*, *43*, 304–316. doi:10.3102/0013189X14545513

McBee, M. T., Makel, M. C., Peters, S. J., & Matthews, M. S. (2018). A call for open science in giftedness research. *Gifted Child Quarterly*, *62*, 374–388. doi:10.1177/0016986218784178

Mellers, B., Hertwig, R., & Kahneman, D. (2001). Do frequency representations eliminate conjunction effects? An exercise in adversarial collaboration. *Psychological Science*, *12*, 269–275. doi:10.1111/1467-9280.00350

Miller, P. (Director). (1988). *Saturday Night Live* [Television series]. New York, NY: National Broadcasting.

Missett, T. C., & Foster, L. H. (2015). Searching for evidence-based practice: A survey of empirical studies on curricular interventions measuring and reporting fidelity of implementation published during 2004-2013. *Journal of Advanced Academics*, *26*, 96–111. doi:10.1177/1932202X15577206

Morgan, S., & Winship, C. (2014). *Counterfactuals and causal inference: Methods and principles for social research* (2nd ed., Analytical Methods for Social Research). Cambridge, MA: Cambridge University Press. doi:10.1017/CBO9781107587991

Morrison, B., & Vaandering, D. (2012). Restorative justice: Pedagogy, praxis, and discipline. *Journal of School Violence*, *11*, 138–155. doi:10.1080/15388220.2011.653322

Moshontz, H., Campbell, L., Ebersole, C. R., IJzerman, H., Urry, H. L., Forscher, P. S., . . . Chartier, C. R. (2018). The psychological science accelerator: Advancing psychology through a distributed collaborative network. *Advances in Methods and Practices in Psychological Science*, *1*, 501–515. doi:10.1177/2515245918797607

Neir, J. A., & Campbell, S. D. (2013). Two outsiders' view on feminism and evolutionary psychology: An opportune time for adversarial collaboration. *Sex Roles*, *69*, 503–506. doi:10.1007/s11199-012-0154-2

Nicol, A. A. M., Owens, S. M., Le Coze, S. S. C. L., MacIntyre, A., & Eastwood, C. (2018). Comparison of high-technology active learning and low-technology active learning classrooms. *Active Learning in Higher Education*, *19*, 253–265. doi:10.1177/1469787417731176

Nosek, B. (2017). *Are reproducibility and open science starting to matter in tenure and promotion review?* Retrieved from https://cos.io/blog/are-reproducibility-and-open-science-starting-matter-tenure-and-promotion-review/

Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science*, *349*, aac4716. doi:10.1126/science.aac4716

Ortiz, A. M., Haviland, D., & Henriques, L. (2017, November 2). Documenting your career for success. *Inside Higher Education*. Retrieved from https://www.insidehighered.com/advice/2017/11/02/how-create-strong-reappointment-tenure-or-promotion-file

Pashler, H., & Harris, C. R. (2012). Is the replicability crisis overblown? Three arguments examined. *Perspectives on Psychological Science*, *7*, 531–536. doi:10.1177/1745691612463401

Pashler, H., McDaniel, M., Rohrer, D., & Bjork, R. (2009). Learning styles: Concepts and evidence. *Psychological Science in the Public Interest*, *9*, 105–119. doi:10.1111/j.1539-6053.2009.01038.x

Pavelka, S. (2013). Practices and policies for implementing restorative justice within schools. *Prevention Researcher*, *20*(1), 15–17.

Pigott, T. D., Valentine, J. C., Polanin, J. R., Williams, R. T., & Canada, D. D. (2013). Outcome-reporting bias in education research. *Educational Researcher*, *42*, 424–432. doi:10.3102/0013189X13507104

Polanin, J. R., Tanner-Smith, E. E., & Hennessy, E. A. (2016). Estimating the difference between published and unpublished effect sizes: A meta-review. *Review of Educational Research*, *86*, 207–236. doi:10.3102/0034654315582067

Rohrer, J. (2018). Thinking clearly about correlations and causation: Graphical causal models for observational data. *Advances in Methods and Practices for Psychological Research*, *1*, 27–42. doi:10.1177/2515245917745629

Rosenshine, B., & Meister, C. (1994). Reciprocal teaching: A review of the research. *Review of Educational Research*, *4*, 479–530. doi:10.3102/00346543064004479

Schochet, P. Z. (2008). Statistical power for random assignment evaluations of education programs. *Journal of Educational and Behavioral Statistics*, *33*, 62–87. doi:10.3102/1076998607302714

Sedlmeier, P., & Gigerenzer, G. (1989). Do studies of statistical power have an effect on the power of studies? *Psychological Bulletin*, *105*, 309–316. doi:10.1037/0033-2909.105.2.309

Silberzahn, R., Uhlmann, E. L., Martin, D. P., Anselmi, P., Aust, F., Awtrey, E., . . . Carlsson, R. (2018). Many analysts, one data set: Making transparent how variations in analytic choices affect results. *Advances in Methods and Practices in Psychological Science*, *1*, 337–356. doi:10.1177/2515245917747646

Simmons, J., Nelson, L. D., & Simonsohn, U. (2011). False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science*, *22*, 1359–1366. doi:10.1177/0956797611417632

Simons, S. J., Shoda, Y., & Lindsay, D. S. (2017). Constraints on Generality (COG): A proposed addition to all empirical papers. *Perspectives on Psychological Science*, *12*, 1123–1128. doi:10.1177/1745691617708630

Simonsohn, U., Simmons, J. P., & Nelson, L. D. (2015, November 24). *Specification curve: Descriptive and inferential statistics on all reasonable specifications*. doi:10.2139/ssrn.2694998

Sisk, V. F., Burgoyne, A. P., Sun, J., Butler, J. L., & Macnamara, B. N. (2018). To what extent and under which circumstances are growth mindsets important to academic achievement? Two meta-analyses. *Psychological Science*, *29*, 549–571. doi:10.1177/0956797617739704

Slavin, R. E. (1987). Ability grouping and student achievement in elementary schools: A best-evidence synthesis. *Review of Educational Research*, *57*, 293–336. doi:10.3102/00346543057003293

Slavin, R. E. (1990). Achievement effects of ability grouping in secondary schools: A best-evidence synthesis. *Review of Educational Research*, *60*, 471–499.

Song, S. Y., & Swearer, S. S. (2016). The cart before the horse: The challenge and promise of restorative justice consultation in schools. *Journal of Educational and Psychological Consultation*, *26*, 313–324. doi:10.1080/10474412.2016.1246972

Steegen, S., Tuerlinckx, F., Gelman, A., & Vanpaemel, W. (2016). Increasing transparency through a multiverse analysis. *Perspectives on Psychological Science*, *11*, 702–712. doi:10.1177/1745691616658637

Steenbergen-Hu, S., Makel, M. C., & Olszewski-Kubilius, P. (2016). What one hundred years of research says about the effects of ability grouping and acceleration on K-12 students' academic achievement: Findings from two second-order meta-analyses. *Review of Educational Research*, *86*, 849–899. doi:10.3102/0034654316675417

Sterling, T. D. (1959). Publication decisions and their possible effects on inferences drawn from tests of significance—or vice versa. *Journal of the American Statistical Association*, *54*, 30–34. doi:10.2307/2282137

Suvall, C. (2009). Restorative justice in schools: Learning from Jena High School. *Harvard Civil Rights-Civil Liberties Law Review*, *44*, 547–569. Retrieved from https://harvardcrcl.org/wp-content/uploads/sites/10/2009/07/547-570.pdf

Taylor, K., & Doolittle, E. (2017, March 28). Building evidence: What comes after an efficacy study? [Blog post]. Retrieved from https://nces.ed.gov/blogs/research/post/building-evidence-what-comes-after-an-efficacy-study

Uhlmann, E. L., Chartier, C. R., Ebersole, C. R., Errington, T. M., Kidwell, M., Lai, C. K., . . . Nosek, B. A. (2019). Scientific utopia III: Crowdsourcing science. *Perspectives on Psychological Science*. Advance online publication. doi:10.1177/1745691619850561

van der Zee, T., & Reich, J. (2018). Open education science. *AERA Open, 4*. doi:10.1177/2332858418787466

van Elk, M., Matzke, D., Gronau, Q. F., Guan, M., Vandekerckhove, J., & Wagenmakers, E. J. (2015). Meta-analyses are no substitute for registered replications: A skeptical perspective on religious priming. *Frontiers in Psychology*, *6*, 1365. doi:10.3389/fpsyg.2015.01365

Waterhouse, L. (2006). Multiple intelligences, the Mozart Effect, and emotional intelligence: A critical review. *Educational Psychologist*, *31*, 201–225. doi:10.1207/s15326985ep4104_1

Wearmouth, J., Mckinney, R., & Glynn, T. (2007). Restorative justice in schools: A New Zealand example. *Educational Research*, *49*, 37–49. doi:10.1080/00131880701200740

Westfall, J., & Yarkoni, T. (2016). Statistically controlling for confounding constructs is harder than you think. *PLoS One*, *11*, e0152719. doi:10.1371/journal.pone.0152719

What Works Clearinghouse. (n.d.). *What we do*. Retrieved from https://ies.ed.gov/ncee/wwc/WhatWeDo

Wolf, R., Morrison, J., Slavin, R., & Risman, K. (2019). *Do developer-commissioned evaluations inflate effect sizes?* Presented at the annual meeting of SREE in Washington DC. Retrieved from https://hechingerreport.org/wp-content/uploads/2019/03/developer-abstract.pdf

Woods, D. (2015, September 23). *The class size debate: What evidence means for education policy*. Retrieved from https://gspp.berkeley.edu/research/featured/the-class-size-debate-what-the-evidence-means-for-education-policy

Yong, E. (2018, November). Psychology's replication crisis is running out of excuses. *The Atlantic*. Retrieved from https://www.theatlantic.com/science/archive/2018/11/psychologys-replication-crisis-real/576223/

Zehr, H. (2015). *The little book of restorative justice: Revised and updated*. Intercourse, PA: Good Books.

Authors

MATTHEW C. MAKEL is the director of research and evaluation for Duke University's Talent Identification Program. His research focuses on academic talent development and research methods.

KENDAL N. SMITH is a doctoral candidate at the University of North Texas and the assistant editor of the *Journal of Advanced Academics*. Her work considers research methods in education, as well as intersections between high ability, morality, and epistemic cognition.

MATTHEW T. McBEE is an associate professor of quantitative psychology at East Tennessee State University. His research focuses on computational methods, statistical simulation, and the identification of students for gifted programs.

SCOTT J. PETERS is a professor of educational foundations and the Richard and Veronica Telfer Endowed Faculty Fellow of Education at the University of Wisconsin-Whitewater. His research work focuses on educational assessment, gifted and talented student identification, disproportionality, and educational policy.

ERIN M. MILLER is an associate professor of psychology at Bridgewater College. Dr. Miller conducts research on how beliefs about cognitive and athletic ability affect motivation and decision making.