The American Statistician

ISSN: 0003-1305 (Print) 1537-2731 (Online) Journal homepage: https://www.tandfonline.com/loi/utas20

How Effect Size (Practical Significance) Misleads Clinical Practice: The Case for Switching to Practical Benefit to Assess Applied Research Findings

Stanley Pogrow

To cite this article: Stanley Pogrow (2019) How Effect Size (Practical Significance) Misleads Clinical Practice: The Case for Switching to Practical Benefit to Assess Applied Research Findings, The American Statistician, 73:sup1, 223-234, DOI: 10.1080/00031305.2018.1549101

To link to this article: https://doi.org/10.1080/00031305.2018.1549101

THE AMERICAN STATISTICIAN
2019, VOL. 73, NO. S1, 223–234: Statistical Inference in the 21st Century
https://doi.org/10.1080/00031305.2019.1549101

How Effect Size (Practical Significance) Misleads Clinical Practice: The Case for Switching to Practical Benefit to Assess Applied Research Findings

Stanley Pogrow[a,b]

[a]Equity, Leadership, and Instructional Technology, San Francisco State University, San Francisco, CA; [b]Emeritus Professor of Education, University of Arizona, Tucson, AZ

ABSTRACT

Relying on *effect size* as a measure of *practical significance* is turning out to be just as misleading as using *p*-values to determine the effectiveness of interventions for improving clinical practice in complex organizations such as schools. This article explains how effect sizes have misdirected practice in education and other disciplines. Even when effect size is incorporated into RCT research the recommendations of whether interventions are effective are misleading and generally useless to practitioners. As a result, a new criterion of *practical benefit* is recommended for evaluating research findings about the effectiveness of interventions in complex organizations where benchmarks of existing performance exist. Practical benefit exists when the unadjusted performance of an experimental group provides a noticeable advantage over an existing benchmark. Some basic principles for determining practical benefit are provided. Practical benefit is more intuitive and is expected to enable leaders to make more accurate assessments as to whether published research findings are likely to produce noticeable improvements in their organizations. In addition, practical benefit is used routinely as the research criterion for the alternative scientific methodology of improvement science that has an established track record of being a more efficient way to develop new interventions that improve practice dramatically than RCT research. Finally, the problems with practical significance suggest that the research community should seek different inferential methods for research designed to improve clinical performance in complex organizations, as compared to methods for testing theories and medicines.

ARTICLE HISTORY
Received March 2018
Revised October 2018

KEYWORDS
Applied research; Evidence-based decision-making; Improvement science; Interpreting effect sizes; Leadership-decision-making; Organizational effectiveness;

1. Introduction

The widely discussed problems created by relying on *p*-values to determine *statistical significance* have created a wave of advocacy for switching to, or incorporating, *effect sizes* as a measure of the *practical significance* of research findings (Kirk 1996; McCartney and Rosenthal 2000; Hojat and Xu 2004; Sullivan and Feinn 2012). Effect size has the advantages of (a) indicating the amount of difference between groups, and (b) not being affected by the size of the sample. However, I will demonstrate that in the field of education and other disciplines, effect size, typically Cohen's d, has been as problematic as *p*-values in producing misleading findings as to the effectiveness of interventions. In addition, emergent scholarship has documented widespread problems in other disciplines from relying on published effect size results to guide clinical practice.

As a result of the problems with both statistical and practical significance, I recommend switching to an alternative statistical criterion, *practical benefit,* for judging the validity and usefulness of applied research findings for improving the performance of complex organizations and for designing more effective interventions. While the focus of this article is on using practical benefit to inform education leadership decision-making, its use has implications for improving clinical practice in other disciplines. At the same time, the recommendations are not intended to inform research focused on generating or testing theory.

The fundamental question that has guided my work is:

> What kinds of quantitative evidence should leaders seek to feel reasonably confident that if they adopt a given practice based on this evidence that it is likely to improve their schools?

The following discussions will demonstrate why practical benefit, that relies on the actual performance of the experimental group, is a better criterion for answering this question about desired evidence and for supporting leadership decision-making and improving clinical practice in other disciplines, than either practical significance or statistical significance. However, before explaining the nature and advantages of practical benefit, it is necessary to first understand the context of educational research and its similarities and differences to research in other disciplines.

1.1. The Context of Education Research

Experimental education research was largely conducted in laboratory settings till the early 1990s. Criticism of the applicability

CONTACT Stanley Pogrow *spogrow@sfsu.edu, stanpogrow@att.net* San Francisco State University, San Francisco CA 94132; University of Arizona, Tucson, AZ 85721.

of such research to real-world practice led to the major shift of conducting research to test the effectiveness of interventions in school settings. In time, pressure began to mount to increase the scientific rigor of such research by using RCT designs.

However, conducting gold-standard research in the complex, chaotic environment of schools is problematic. Placebos are usually not possible so the control groups tend to represent the existing practice which is usually not well described. Randomization is problematic for the following reasons:

- It is usually impossible to randomize at all the relevant levels of schools, grade levels, teachers, and students,
- Randomization tends not to equate all possible relevant student characteristics,
- Randomization invariably breaks down over time, as teachers leave, parents requests transfers for their children to another class or to a higher performing school, or students leave for other reasons. Student attrition alone over a 3-year period can be 50–70% high-poverty schools.

RCT research is also problematic because there are so many variables that can potentially impact learning that it is impossible to control for most of them. For example, there may be differences in the lead levels of the drinking water between schools, or different room temperatures, etc. Also, generalizing findings across different school districts and states is almost impossible because different tests are used and the contexts of the research vary. In addition, the cost of RCT experiments makes it unlikely that any study will be replicated. The appendix describes how all of the above problems played out in the most famous education experiment, Project STAR in Tennessee to determine the effects of class size reduction.

Given the difficulty of maintaining control of variables in the real world, methodologists have created advanced statistical methods that seek to simulate control of confounding variables or changing circumstances, but these are extremely complex and have their own errors.

The problems with conducting experiments in schools, or any complex organization, as compared to medicine are reflected in Glass's (2016) comparative analysis of meta-analyses in the two disciplines. Gene Glass, who named and pioneered the use of meta-analysis, found that while in medicine the effect sizes of experiments in a given meta-analysis tended to be large and stable across studies, in education meta-analyses the effect sizes are "relatively small" and their variation is "great" within the same meta-analysis. As a result, Glass argues that because of unstable small effect sizes meta-analysis has not been as useful for informing policy in education as in medicine. Indeed, it will be shown that the problem of basing recommendations for clinical practice on small effect sizes is also a problem in other disciplines.

The bright spot in education research has been the establishment of comprehensive databases at all levels. The federal National Assessment of Educational Progress (NAEP) periodically tests national performance by subject area, grade levels, student and district demographics. NAEP results are reported for the nation and for individual states. The federal National Center of Educational Statistics collects longitudinal data on a variety of educational outcomes and processes that are broken out in great detail in terms of student and school characteristics. Federal funding has also strengthened the data collection and statistical reporting of state departments of education on the performance of its schools and districts. In other words, education is now a field rich in benchmarks of current and historical performance.

Given (a) the existence of a wide array of benchmarks, (b) the difficulty of truly establishing causation in field-based research in complex organizations, (c) the desire of leaders for replicated noticeable improvements in research, a different statistical criterion, *practical benefit*, is recommended that compares the unadjusted actual performance of an experimental group to an existing benchmark. Such an approach for assessing research findings and the design of new interventions will make it easier for leaders to better determine whether a given intervention is likely to improve their schools to a noticeable extent. Practical benefit also offers the potential to use research more efficiently to discover better interventions, more quickly, and at a lower cost—than from RCT research.

2. How the Use of Effect Size Misdirects Practice in Education

Effect sizes provide the potential to be more relevant to improving clinical practice than *p*-values as leaders do care about magnitudes of benefits. However, that does not necessarily mean that researchers care about the same level and types of magnitudes that leaders and other clinicians do. What is clear is that education has institutionalized the use of effect sizes within RCT research as the key criterion for identifying effective practices with the establishment of the federal *What Works Clearinghouse* (WWC).

2.1. Has Increasing the Scientific Rigor for Identifying Effective Practices in Education Produced Valid Results?

The goal of the WWC is to bring the most rigorous scientific criteria to bear to inform educators as to what works. Rigorous science has come to be defined as the use of RCT with an effect size of at least 0.25. WWCs methodologies have become the standard for how the US Department of Education determines which research proposals to fund. In addition, the WWC list of practices it has certified as having strong evidence to support their use has become more than just a good housekeeping seal of approval. WWC certification also increases a program's chances of obtaining federal funds to disseminate its use. In addition, there is advocacy that the federal government require that low-performing schools use their federal funds only to implement practices approved by the WWC. So, the stakes for being certified as effective by WWC are quite high.

However, such institutionalization of methodology, combined with political pressures for schools to adopt certified interventions, will be very damaging if the interventions that have been certified are NOT actually effective. The result would be analogous to what would happen if the FDA approved an ineffectual or harmful drug.

Unfortunately, there is growing evidence that relying on *RCT*s and effect sizes are producing misleading findings that

conclude that interventions are effective when in fact they are not. As previously discussed, it is highly questionable whether the static and expensive nature of RCT research can capture the constantly shifting interactive dynamics and chaos of complex organizations such as schools, or that causation can ever be established in such an environment. Indeed, Ginsburg and Smith (2016) examined the evidence for all of the 18 math programs certified by the WWC as having evidence of effectiveness based on a total of 27 approved RCT studies. Ginsburg and Smith (2016, p. 44) found 12 potential threats to the usefulness of these studies and concluded "…none of the RCT's provides useful information for consumers wishing to make informed judgments about what mathematics curriculum to purchase." The questionable utility of RCT methods for identifying clinical practices that are effective in the real world does not just seem to be just a problem in education. Section 4 of this paper shows how a critique of relying on RCT to identify effective practices in another complex organization—hospitals—appeared in the Journal of the American Medical Association (Berwick 2008).

However, regardless of whether RCT or less rigorous methodologies are used, relying on effect sizes as the statistical criterion for determining the practical significance of findings is turning out to be highly misleading—as problematic as relying on *p*-values.

2.2. Problems With Using Effect Size to Determine the Effectiveness of Interventions

The biggest problem is that the magnitude of the effect size needed to claim that a finding has practical significance has been set way too low by the research community. Setting the desired magnitude of effect size so low enables researchers to claim that they found practical significance when there is no actual difference of real-world importance. The minimum standard for the effect size needed to claim practical significance is generally set at 0.2 based on Cohen's (1988) categories of small (0.2), medium (0.5), and large (0.8). However, researchers ignore that Cohen referred to an effect size of 0.2 as "difficult to detect." No leader should consider adopting an intervention based on such a result—Leaders seek *noticeable* improvement. Even worse, education researchers such as Borman, Grigg, and Hanselman (2016) are now seeking to reduce that cutoff to 0.1, or half of "difficult to detect," and some (Deke, Wei, and Kautz 2017) are even advocating for effect size as low as 0.03. Perhaps the silliest rationalization for the importance of a small effect size in the education research literature is the argument that an effect size of 0.18 is important because it is twice as good as other interventions tested which only produced an effect size of 0.09. *The correct conclusion is that neither is likely to produce any benefit in terms of real-world school improvement*.

Even the federal WWC over-estimates the real-world importance of small effect sizes. It is minimum effect size cutoff of 0.25 is simply too small given that Cohen (1988) estimated that effects become noticeable only at an effect size of 0.5.

Small effect sizes also appear to be a reason why so much of influential published experimental research cannot be replicated in psychology (Open Science Collaboration 2015), and oncology (Begley and Ellis 2012), and why laboratory research results often do not predict actual outcomes in psychiatry (Kraemer 2016). The best summary of the problem of exaggerating the importance of small effect sizes was the observation by Ioannidis (2005), co-founder of the Meta-Research Innovation Center at the Stanford University, that the smaller the effect sizes in any scientific research the less likely it is that the research findings are true. We seem to have traded *p*-hacking for effect size-hacking. In both cases, the importance of a small difference is exaggerated, and such exaggeration misdirects practice in education and other disciplines.

The second problem is that even where large effect sizes are reported they can mask the fact that the experimental students did terribly, and/or the actual benefits were so tiny as to not be of any practical benefit. There are generally two ways that effect sizes overstate results. The first way is that relative differences can ignore the key, most relevant, actual context information. The process of ignoring actual context is best illustrated is by the following scenario of a couple living in the upper Midwest during a particularly bad cold spell deciding where to vacation in January where the goal is to find a place where they can relax on a warm beach and get a tan:

Wife:	I cannot wait for our vacation in January. Let's go somewhere warm.
Husband:	Definitely.
Wife:	Where should we go?
Husband:	I just read that Greenland is warmer than Antarctica in January.
Wife:	That sounds great.
Husband:	Even better, due to climate warming Greenland will be warmer this year than last. Plus, it has 27,394 miles of coastline, so it will be no problem finding beaches.
Wife:	That's great. It will be wonderful to go somewhere where we can leave our winter clothes behind.

Their decision is certainly evidence-based—but it is clearly a lousy decision. Why was this couple's *evidence-based decision* so bad? It was bad because they relied *solely* on relative data. They needed a key piece of absolute data about the context, which in this case was the actual temperature in Greenland in January. The actual temperature is −8°C with zero hours of sunshine. This couple is far more likely to die from hypothermia in Greenland in January than to get a tan. With the right actual data, the couple would arrive at the correct decision to reject vacationing in both Antarctica and Greenland as it would be warmer to stay at home or to seek other options.

The above example illustrates why you cannot make intelligent decisions relying *solely* on relative data that relate external outcomes to each other, no matter how compelling that evidence appears to be. You always need some absolute data. The key piece of absolute data needed to judge the quality of an education program is the answer to the obvious question: *How did the students in the experimental intervention actually perform?* For a program to be judged successful, we would expect students to do reasonably well on an absolute basis. Of course, reasonable people can debate as to what an expectation of students "doing well" is. The problem is that researchers are not being asked to provide this most basic information in published research and are able to get away with just reporting

the statistical or practical significance of the relative result. Not having to report actual outcomes leads to the following question: In what percentage of research claiming to have demonstrated the effectiveness of interventions for helping children born into poverty did the experimental students actually do well—or actually do terribly? We simply do not know! If it is the latter, education is wasting billions of dollars and decades in applying evidence-based practices that inadvertently maintain gross inequities.

The second way that reports of larger effect sizes often can mislead and hide the fact that the experimental students did terribly is the use of questionable adjustments to the *means*. The most typical adjustments in education are to use covaried *means*. However, covarying *means* creates several problems. First of all, since we are talking about small differences, it is easy to cherry pick samples in such a way so as to make it appear that the initial *mean* of the experimental group is lower on some measure in order to boost the final *mean* of the experimental group using an analysis of covariance and thereby produce a positive effect size. A second type of adjustment, normalization of *means* occurs if different tests are used. The relative results are then based on normalized covaried *means*. However, such results are abstractions with no real-world meaning that can be understood by practitioners and policy makers.

So, researchers have come up with a way to kill two birds with one stone. They convert the abstract numbers into ones with real world meaning by creating hypothetical extrapolations that make small differences in the normalized covaried *means* appear important. In education, the most common extrapolations are to convert adjusted differences between groups into (a) extra days of learning, or (b) advantages for the experimental group on a nationally normed test. However, these extrapolations are usually grossly invalid and are designed to make the relative results seem important.

For example, if an intervention in a high poverty school produces an effect size of 0.2, the researchers will note that this result is equivalent to increasing scores of the experimental students on a nationally normed test from the 50th to the 58th percentile. Such an increase indeed seems impressive! Of course, such extrapolation ignores that these schools are way below national norms which is why the improvement effort was conducted in the first place. Extrapolating results from distributions of low performing schools onto national norms is deceptive and invalid. In addition, since the actual results were usually not from a nationally normed test—if the experimental students had actually taken the test the results were extrapolated to, they might have scored at the 28th percentile as opposed to the 58th.

Such manipulated extrapolation is akin to telling our hapless vacationers that the difference between the temperatures in Antarctica and Greenland in January is equivalent to raising the temperature in Miami in January from 76° to 85°. Such a hypothetical extrapolation of temperature makes Greenland seem warm! At the end of the day the reality is that the average temperature in Greenland in January is not 80°—it is still −8.

As a result of the above problems, the statistical criterion of effect size appears to be as misleading for identifying effective practices as relying on *p*-values. The following examples show how manipulated effect size results have misdirected practice in education on a large scale.

2.3. Two Examples of how Effect Size Results in Influential Research Misdirected Clinical Practice in Education on a Large Scale

Example 1. The effectiveness of *Success for All* (SFA)

SFA became the leading reform for improving the performance of failing high poverty inner city elementary schools. It is an intensive reading intervention for grades k-5. SFA built its reputation on the basis of research published in the top research journals apparently showing it to be uniquely effective in improving the performance of high poverty urban schools. Between 1990 and 2008, there were eight articles in the prestigious journals of the American Educational Research Association (AERA) by the co-developer and/or researchers who were, or had been, associated with him that documented the success of *Success for All* in a series of urban districts (Slavin 1990; Madden et al. 1993; Ross et al. 1995; Borman and Hewes 2002; Borman et al. 2005, 2007). These publications were the tip of the iceberg of research articles demonstrating the success of the program.

SFA research took full advantage of the awesome power of modern statistical analysis. The state-of-the-art analyses established the superior relative average performance of the experimental SFA students over students in comparison schools. However, SFA research, amidst all the numbers in all the published articles in the highly ranked journals *never* (to my knowledge) revealed how the SFA students actually performed on the national measures used in the studies. It turns out that SFA students were actually doing terribly.

The most famous experiment with SFA occurred in high-poverty schools in Baltimore in the early 90s. SFA research findings claimed success based on large effect sizes. However, an independent reanalysis of the data by Venezky (1998) showed the students doing poorly. Pogrow (1998, 1999) elaborated Venezky's findings and showed how after five years in SFA the experimental students entered the sixth-grade reading almost 3 years below grade level. This deficit is actually an over-estimate of the experimental students' performance as Venezky reported that (a) the sample was limited to only stable students (who tend to score higher) and did not include the students who were in the program for less than 5 years, and (b) the special needs students disappeared from the post-test experimental sample. The latter, combined with the analysis of covariance, boosted the relative performance of the experimental group. In addition, SFA was one of the most expensive interventions available and more money was spent on the experimental schools, and the students spent substantially more time reading. If these variables two variables had been covaried, the comparison schools would most likely have been ended up with a higher adjusted *mean*.

The SFA students in the Baltimore experiment in fact did terribly. No educator would consider the actual results to be a success or even an acceptable result. The poor performance of the experimental group cited above was confirmed by the independent study that the district's research office conducted (Ruffini 1992). As a result, the district dropped the program.

Subsequent evidence has shown that SFA underperforms other available options. Dade County School District in Florida made a large commitment to SFA. Urdegar (2000) showed that SFA schools in Dade County were doing poorly and that similar

schools in the district using their own homegrown interventions were doing better at a lower cost. In addition, the only experiment that compared SFA to other "effective" reading programs while controlling for the amount of time spent reading, found that SFA was doing so poorly that most of the schools that had been randomly assigned to implement SFA dropped it after the first year for poor performance and refused to continue using it in the second year of the experiment (Burdumy et al. 2009).

Pogrow (2000a, 2000b, 2002) reported consistent failure across the U.S., with disillusioned schools dropping the program. Yet even as such contrary evidence appeared, and schools dropped the program, the government continued to provide even larger grants to disseminate it—and the SFA researchers continued to cite Baltimore as a "success" in their successful funding proposal for a new $50 million dissemination grant almost two decades later.

In other words, there was a dichotomy between the published *effect size* findings in the top journals and what was actually happening in practice. Under any common-sense metric that practitioners who the research was supposed to "inform" would use, the program was clearly failing. None of the peer-review panels of the research journals or the government panels reviewing proposals thought to ask the one question amidst all the data that were of most interest to practitioners and the public at large; that is, *How did the SFA students actually perform?* Even worse, despite the demonstration of failure over several decades, the federal WWC continues to list SFA as one of the programs with the strongest evidence of success. WWC's support of SFA is the equivalent of the federal government recommending that school leaders should vacation in Greenland during winter recess to get a tan. More importantly, WWC's support means that schools seeking to accelerate the achievement of their low-income students will continue to be pressured to adopt an intervention that has a long track record of failure.

Example 2. The CREDO study of the effectiveness of charter schools

The widely cited CREDO study at the Stanford University compared the effectiveness of charter schools to traditional public schools. The study concluded that charter schools produce an extra 14 days a year of learning a year for Black students as compared to traditional public schools. Fourteen extra days of learning a year seems to be a substantial advantage for charter schools. However, the actual *effect size* was .02, or a tenth of "difficult" to detect." How did CREDO (2013) extrapolate .02 to 14 days? CREDO (2013, p. 13) noted that their findings "…are only an estimate and should be used as a general guide rather than as empirical transformations." CREDO is essentially admitting that there is no real empirical basis for their published hypothetical extrapolation. Maul and McClelland (2013) noted that CREDO's conversion of *effect size* into days of learning was "insufficiently justified" and that there really was not a substantial difference between the performances of the two types of schools. My own thought experiment concluded that the tiny effect size of .02 was at best equivalent to an advantage of 2 hours a year of extra learning. Two hours is a tiny difference that is dwarfed by the errors involved in the analyses. In other words, the correct conclusion should have been that *there was no practical difference* between the two types of schools. Alas, advocates for charter schools cited the CREDO conclusion which helped spur increased support for such schools nationally, while unfairly damaging the reputation of traditional public schools.

2.4. Maintaining the Hype of the Importance of Small Effect Sizes

The majority of education researchers, practitioners, or journalists do not understand how to interpret the magnitude of effect sizes. In 2013, MDRC released its evaluation of SFA's effectiveness on its latest federal dissemination grant it had received several decades after the initial Baltimore experiments described earlier. The study (Quint et al. 2013) concluded that the results were promising. The findings were based on effect sizes of −0.01 and 0.18 in two different reading skills at the kindergarten level. This conclusion was parroted in an article in the widely read *Education Week* with the headline: "School Improvement Model Shows Promise in First i3 Evaluation (Sparks 2013)." The reporter then goes on:

> One of the biggest early bets in the U.S. Department of Education's Investing in Innovation program seems to be paying off: Success for All, a literacy-related, whole-school improvement model, shows signs of changing teaching practice and boosting students' early reading skills after a year in schools…

Aside from the question as to why you would consider results for a 25-year-old program promising, the fact that the early grades are the ones where it is easiest to show large *effect sizes*, the reality is that these results are terrible. A meta-analysis of the effects of intensive reading interventions in the earliest grades by Scammacca et al. (2007) found effect sizes at the ranging from 0.34 to 0.56 across five different reading skills. (A less direct comparison of meta-analysis effects of reading interventions by Hattie (2009) found effect sizes ranging from 0.60 to 0.67.) The correct interpretation of the effect sizes reported by MDRC is that these results are among the worst ever recorded for an intensive reading intervention at the early grades.

So, once again effect size findings are being put forth in a misleading fashion and parroted by the media. No one in the research community or media is raising obvious questions about such findings. Why not?

A key argument to accept small effect size results as a finding of practical significance is that rigorous experiments conducted in schools rarely finds large effect sizes. Researchers therefore argue that it is not fair to expect research to produce larger effect sizes. However, such logic makes no sense from the perspective of the leaders for whom the findings are intended. Why should leaders be expected, or possibly even mandated by the federal government, to adopt an intervention that is not likely to produce noticeable benefits simply because that is the best that researchers can do? Such logic around the interpretation of statistical criteria preserves a system that operates for the benefits of researchers with no evidence that it actually produces benefits for leaders and their schools. Of course, it may be that it is impossible to produce noticeable improvements in the many problems that continue to persist in education. However,

an alternative explanation for the ineffectual research results is that the educational research community is using the wrong scientific model of research for developing and testing interventions. (An alternative scientific method that other disciplines have used to produce major clinical improvements to seemingly intractable problems will be explored in Section 4 of this paper.) At the very least, sufficient evidence has been presented to conclude that the use of effect sizes has been just as misleading as *p-values* in both education and other disciplines. (The use of odds-ratios appears to be equally problematic in exaggerating the importance of outcomes in education and medicine, but that is a discussion for another paper.)

2.5. Where Do We Go From Here?

The magnitude of benefits from a new intervention, or what Ziliak and McClosky (2004, 2008) call "oomph," is clearly important. Leaders seek easy-to-understand and intuitive evidence that an intervention has produced "oomph" elsewhere and that such outcomes have been replicated in a variety of settings—ideally ones with a similar context to their own. Most of all they want to know how students in the intervention actually performed. However, RCT research and *effect size* reporting generally do not provide information on the actual (as opposed to relative) performance of the experimental students.

It is therefore critical to develop an alternative evidentiary criterion that (a) is more intuitive, (b) provides leaders with the type of information they seek, and (c) leaders can apply to their unique context. Such methods need to be less mathematically complex, and better able to distinguish which interventions are truly likely to be effective. As a result, I developed the alternative evidentiary criteria for informing leadership decision-making of practical benefit. These criteria are designed to provide leaders with an indication of whether adopting the findings and recommendations of specific studies are likely to benefit their particular schools in a clearly noticeable way.

3. Switching to Practical Benefit

The first time I presented the concept of practical benefit to a few statisticians over lunch they simply stared at me for several minutes without saying anything. When I broke the uncomfortable silence and asked them why they were not reacting, their response was that the idea was contrary to everything they had been taught, everything they believed in, and everything they taught their students. Anyway, here is that idea.

3.1. Determining the Practical Benefit of a Research Finding for Improving One's Schools

Practical benefit recognizes that the data that leaders most want to know is (a) how the experimental students actually did, and (b) whether the experimental students' actual results are noticeably better than *their students*' existing results. Knowing how an external control group did is of no interest to them—nor should it be. The relevant comparison is how the external experimental group (only) actually performed relative to their own students. The experimental students' performance is then compared to an existing benchmark of current performance in their own organization and possibly statewide or nationally.

In other words, when considering the likelihood that a published research finding will produce a noticeable benefit in one's own schools, practical benefit ignores how the external control group performed. That means that leaders can ignore all the statistical analyses trying to determine relative performance, and just focus on the unadjusted mean/median of the experimental group, and how it compares to their schools' performance. The only other relevant information in any study is the context in which the research was conducted. Every other stat or technical term in the research can be ignored. Using this method of focusing only on actual performance of the experimental group educators can peruse a stack of the most sophisticated quantitative research articles and determine in which, if any, the performance of the experimental group exceeds their existing benchmarks. Leaders can then focus on considering the interventions in those articles that report high actual performance.

A major argument against ignoring the control group results is that they are important for understanding the context of the research. That is probably true in highly controlled laboratory research. That is not true for research in complex organizations. The few variables that the study makes an effort to control are a minority of the many other interactive variables that exist. Therefore, while context is critical, a much better way to understand the context of the research is via a mixed-methods approach where there is a careful qualitative description. Alas, in most published quantitative research that I see there is little more than cursory discussion of context. (The consequences of ignoring context will be explored in the next section.)

But how does one measure "noticeable benefit" or "oomph?" Ziliak and McClosky (2004, p. 531) define oomph as "A big change, important for the science." I will paraphrase that and define practical benefit as a big change, important for the profession and/or individual leader. "Big" can be determined by human judgment. At the very least "oomph" is a clearly noticeable improvement or other benefit that does not require precise statistical criteria to discern.

While relying on human judgment to determine the effectiveness of an intervention is obviously imprecise and subject to its own biases, Ziliak and McClosky argue that creating the potential for big benefits is more important than being able to precisely predict the level of benefit. When a big improvement occurs, people can generally recognize it.

In addition, relying on human judgment to estimate whether the level of performance of an experimental group in a published research article is likely to produce noticeable benefits in one's schools is consistent with management theory which uses a goal setting process. Goal setting is also usually based on human judgment and aspiration.

Determining how likely the performance of the experimental group in a published study is likely to replicate in one's own schools also requires considering the context of the study. Examples of key context variables include (a) the types of students in the sample, (b) whether the duration of the intervention is a few days or an entire school year, and (c) whether the outcome measure is standardized. If there is a way to match up the results

for the sample or subsample in the studies with the characteristics of one's schools that increases the likelihood that the results will replicate. The final decision is whether resources and expertise needed to implement the intervention are available to the leader.

Is it possible to produce "oomph?" Yes! What does it look like? Here are two examples.

- The Carnegie Foundation established an initiative to try and improve the dismal record wherein half the students who enter community college fail to pass developmental math and therefore never earn any community college credits before giving up. This level of failure may be the biggest dropout rate in American education. Carnegie shifted the developmental sequence away from high school Algebra to statistics with its Statway program. The result was an increase in the percentage of students who passed developmental math and subsequently earned community college math credit from 15% after 2 years, to 50% after only one year.
- Studies of federal efforts to support students born into poverty (Title I students) consistently finds that students make gains in grades k-3, and then fall back after that. My own Higher Order Thinking Skills (HOTS) project provided general thinking development help in grades 4–8 instead of remediation. This program was adopted in approximately 2600 schools around the US and served 1/2 million students. Results consistently showed Title I students making three times the growth in reading comprehension and twice in math, and close to 15% of these low-performing students made honor roll in the first year of the program.[1]

In both cases the benefits relative to benchmarks are obviously substantial, and do not require calculations of p-values or effect sizes to justify their practical benefit.

In summation, determining the practical benefit of the findings of a published research study involves the following steps:

(a) Extract the average actual unadjusted performance of only the experimental group from the mass of data, and ignore all external relative results, adjusted results, and extrapolated results,
(b) Check the context of the research to determine how relevant the study is to your schools, and which findings are the most relevant—though leaders may choose to implement the study's approach even if their context is different if the benefits are substantial, much as my development work applied techniques from private schools to inner city schools, and
(c) Use human judgment to decide whether the actual results of the experimental group (and of the key subgroups) are sufficiently better than how your students are currently doing to warrant the time and money to invest in adopting the intervention.

3.2. Advantages of the Criterion of Practical Benefit

By focusing ONLY on how the experimental group did on an unadjusted basis and using human judgment to determine whether the experimental students did or did not do sufficiently better than your students, you bypass virtually all the complex statistics. In addition, the actual performance of the experimental group is the measure that leaders are most interested in knowing and most care about. There is no reason to use p-values or calculate effect sizes or any other external measure of relative differences such as odds-ratios. As a result, the use of practical benefit bypasses all the problems described above that have misdirected practice. There is reason to believe that the simple and intuitive measure of practical benefit will yield better decisions for improving clinical practice than results produced by traditional inferential analysis—for mission critical problems in complex organizations where benchmarks exist.

Practical benefit also makes the most complex quantitative research in the top journals accessible to practitioners. Practitioners can critically examine such research and reach their own conclusions as to its practical benefit by simply looking at the unadjusted *means* and standard deviations of the experimental group, and then possibly also medians—and then compare the results and context to the benchmark of how their schools are performing.

3.3. Limits on the Applicability of Practical Benefit

3.3.1. Limit 1—What if Research does not Report Unadjusted Means/Medians?

While I am not aware of journal policies in other fields, AERA does not require authors to report unadjusted results in its journals. To put it simply, you cannot rely on adjusted means/medians for decision-making purposes for the reasons already discussed. Therefore, you cannot really determine the practical benefit of the findings. The best you can do is look at the magnitude of effect sizes and seek studies that produce large ones and view that as an indication that the findings are worth looking into. How big should the effect size be?

Given that Hattie's (2009) meta-analysis of all meta-analyses in education found an overall effect size of 0.4, and Cohen (1988) concluded that effects become noticeable at 0.5, expecting an effect size of at least 0.45 seems to be a reasonable starting point for deciding that a research finding merits consideration. Expecting a minimum effect size of 0.45 is particularly true for research after the earliest grades and research conducted in schools over a period of at least several months with a widely used measure.[2]

For nonexperimental research, the recommended minimums for correlation coefficients to be considered as potentially important for leadership decision-making are $r > 0.39$, or $r < -0.39$. An r of 0.39 means that the variation in one variable is associated with 15% of the change in the other. A good argument can be made that this admittedly arbitrary minimum should be set higher. However, the ability to increase the predictability

[1] The HOTS program ran for 24 years and was discontinued 5 years ago This section discusses what was learnt as a researcher from the iterative development, evaluation, and large-scale dissemination of this program.

[2] Effect sizes tend to be higher for short-term laboratory research, research conducted at the earliest grades, and research with nonstandard measuring tools—particularly for measures developed by the researcher.

of another outcome by 15% is probably sufficient incentive for most leaders to consider that approach. For regression the suggested cutoff is whether the unique contribution of an actionable variable in a regression model increases R^2 by 15%.

There are three main problems with such cutoffs.

- These criteria will not prevent *p*- or effect size-hacking, however, the larger effect size minimum values might possibly make it more difficult to hide such behavior,
- Leaders may often NOT be able to find research in a particular area of interest that meet these criteria. Of course, knowing that research cannot help in a given situation is better than pursuing a course of action thinking that it has strong evidentiary support when it does not. Indeed, there is a long history of national fads in education that were supposedly evidence-based, and
- The biggest problem is that there are times when a small effect size can be important, depending on the context of the research. Sometimes a small effect size can mask a high potential clinical approach.

An example of the latter is the extensive research literature that consistently found *negative* effect sizes for the approach used in the previously mentioned HOTS program. *Fortunately, I was not aware of this research.* My ignorance enabled me to produce a major success in developing the thinking skills of students born into poverty that translated into major gains in reading and math (Pogrow 2005) by using a method that research conclusively showed did not work. How is this possible?

One of the most agreed-upon findings in the psychology research literature is that if you want to increase students' problem-solving ability in a content area such as math, you should provide more and better practice in solving thoughtful math problems—a *thinking in content* approach. A thinking in content approach is universally considered superior to a *general thinking development* approach. The latter tries to develop students' general thinking skills, such as metacognition, that might simultaneously improve problem solving in a variety of content areas. As a result, thinking in content is the basis of all reform movements in education that seek to increase the problem-solving ability of disadvantaged students in math, social studies, science, and reading comprehension. Of course, my "ignorant" approach was to improve the academic performance of children born into poverty in grades 4–8 solely by developing their general thinking over a year or two without providing any additional help in content specific skills.

However, when I shared the major gains we were seeing with my psychology friends they invariably told me that the approach could not possibly be working given that research consistently found negative effects from general thinking interventions. As a result, I decided to look into the apparent conflict between the large-scale results we were seeing and what research had universally concluded.

The extensive body of research on the effectiveness of thinking in content methods was well done and had indeed consistently found that general thinking did not work (Willingham 2007). However, the clear disconnect with the published research findings and the success we were producing resulted from the fact that the contexts in the published research had nothing to do with the context I was working in. The initial experiments were carried out with graduate students in math and physics. Clearly, students would not be at such an advanced level if they did not already possess a high level of general thinking ability. It is also clear that further skill development at such a high level of learning requires that students learn to solve ever more advanced math and physics problems—that is, thinking in content. What does the methods used with graduate students have to do with a fifth grader in Harlem reading 3 years below grade level? Nothing! The more recent research supporting the thinking in content approach with students in grades 4–8 was conducted in largely high-income schools, and the experiments ran from 1 to 3 days. Clearly, if you want students to learn how to solve a specific problem in 1 and 2 days they will clearly do better when they are taught the specific method for solving that problem as opposed to providing general thinking activities. As a result, no one had ever tried to provide intensive, daily, general thinking activities to low-income students—which in my ignorance I did successfully.

The point of my experience with the HOTS program is that depending on the context of the research, a small or negative effect size can mask a high potential intervention in a different context. This example also illustrates that it is always important to thoughtfully examine the context of research instead of just relying on some outcome number or numerical cutoff score.

3.3.2. Limit 2—Macro-Level Decision making at the Federal Level

The problems of establishing appropriate statistical criteria to measure effectiveness get more complicated when government seeks to impose improvement goals on all schools within an accountability framework. How much improvement should the government require? How should it be measured? The various efforts to date to set standards for improvement have generally not been very successful.[3]

A better approach is for the federal education department to support a better scientific method for developing interventions that produce replicable "oomph." Several successful applications, Statway and HOTS, were highlighted which met the criterion of practical benefit at scale and "oomph."

4. Practical Benefit and Improvement Science

While education was trying to do its best to emulate medical research as a sign of its growing professionalization, unbeknownst to it there were elements of medical practice that did not use RCT research. While gold-standard experiments are the universal method for testing drugs, there are branches of healthcare and medical practice that do not rely on experiments—

[3] The first time the federal government tried to set improvement standards it required all states to set improvement targets. Most states set the targets embarrassingly low, and this provision was dropped. The No Child Left Behind law passed in 2002 used a complex algorithm to establish annual improvement goals for each school in the nation to bring all students up to grade level in reading and math within 10 years. There were strong accountability penalties for schools that did not meet their annual progress goal. However, this improvement goal was too ambitious and made a mess. In 2015 Congress reversed course and eliminated federally mandated progress goals.

particularly for solving time-critical problems in complex organizations such as hospitals. A notable example of applying alternative research methods were successful efforts to improve the quality of health care in hospitals. The inadequacy of RCT research for improving hospital care is highlighted in the following quote by Berwick (2008) in the *Journal of the American Medical Association*:

> Changes in the current approach to evidence in health care would help accelerate the improvement of systems of practice…Educators and medical journals will have to recognize that, by itself, the usual… experimental paradigm is not up to this task. It is possible to rely on other methods without sacrificing rigor. Many assessment techniques developed in engineering and used in quality improvement…have more power to inform about mechanisms and contexts than do RCTs, as do ethnography, anthropology, and other qualitative methods. For these specific applications, these methods are not compromises in learning how to improve; they are superior.

Berwick's methods came to be known as improvement science and is based on the rapid prototyping of approaches to improve clinical outcomes, and the ones that demonstrate initial practical benefit are iterated to try and adapt to a wide variety of contexts. The goal is to produce clearly noticeable improvements against existing benchmarks (i.e., practical benefit) and collaboratively share the results of iterations in different contexts.

The methods of improvement science are on the surface a somewhat haphazard and imprecise method to design innovation. However, improvement science can produce dramatic, large-scale, clinical improvements. Gawande (2007) notes that the medical field of obstetrics has the best record of increasing the number of lives saved and that such improvement was produced without conducting formal experiments.

Elements of improvement science are now even being extended to the testing of new medicines. Kolata (2015) described how *precision medicine*, the newest federal initiative seeking to discover new and more powerful new cancer drugs, is moving away from conducting experiments in favor of seeking large benefits from the rapid testing of many different compounds for specific cancers. Instead of comparing the results to a formal control group, the results are compared to the known benchmark of the response rates of patients to currently available treatments. Kolata (2015, p. 2 of download) notes that "unlike previous efforts that looked for small differences between a new treatment and an older one…researchers are gambling on finding huge effects." Scientists are finding a patient response rate to the new drugs of 50–60% as compared to existing treatments that give a response rate of only 10–20%.

These alternative methods of scientific discovery that seek large improvements over existing benchmarks are essentially using practical benefit for demonstrating evidence of effectiveness. Improvement science networks are now springing up in medicine and other fields to try and solve heretofore-intractable problems using the criterion of practical benefit.

Improvement science is equally applicable to solving problems in education. The Carnegie Foundation has been taking the lead in working with school districts on how to use improvement science methods, and in creating networks and conferences to share results. Indeed, the dramatic results of the Statway and HOTS programs described earlier could not have been developed through rigorous experiments. There are simply too many variables and parameters to manipulate. The only way to find some near optimal combination of implementation parameters is to embrace the reality that there is tremendous variation in school contexts and learn from the beginning how early prototypes are working or not working in various contexts—and then make needed adjustments on the fly. Such adjustments cannot be made without researchers becoming intimately knowledgeable about the many facets of the different contexts that the intervention needs to be able to adapt to. As previously noted, such a focus on context has been noticeably lacking in more gold-standard forms of research, and such absence has led to misunderstanding of the applicability of research findings.

The methods of improvement science are in contrast to RCT research. RCT methodology pretends that it can control just a few of the interactive factors and that the findings will therefore generalize because of some elements of randomness and a significant p-value or nondetectable effect size. Such research has little actual understanding of the variations of contexts that exist and their potential impact on external validity. Nor will federal efforts to scale interventions that meet WWC standards to other contexts work because the standards are inadequate for reasons discussed earlier and the initial results probably had no "oomph."

Improvement science, on the other hand, relies on quick, flexible adaptation using practical benefit to determine which iterations of the intervention are working and which are not. If it is not initially effective it is critical to quickly figure out why not and make needed adjustments. Such flexibility and adaptability are critical because developing an intervention that provides practical benefit at scale requires developing a set of very precise parameters for all the key design and implementation factors. Improvement science eschews seeking causation to produce replicable, consistent improvement across contexts.[4]

Of course, it is not easy to produce such outcomes. Kerwin and Thornton (2018) describe how producing an effective intervention requires determining the near optimal mix of design parameters, and how minor changes in any parameter can substantially reduce effectiveness. They show how an early literacy intervention in northern Uganda found major gains in reading and writing. However, when relatively small programmatic changes were made the intervention was no longer effective. Kerwin and Thornton (2018, p.33) further note that:

> Evidence on the sensitivity of program results to implementation details is scarce…
> Thus even successful educational interventions implemented as a pilot may be completely uninformative to the results of a scaled-up version of the program: it is hard to know whether a seemingly small change can cause a large difference in a program's impacts, and there are innumerable such changes that can and will occur…Programs designed to

[4] Producing consistent improvements across contexts minimizes Ziliak and McClosky's notion of loss, that is, reduces the probability that adopting the intervention will do damage, as part of maximizing "oomph."

exploit complementarities—rather than to isolate the effects of individual inputs—are likely to be more effective at improving learning.

This is one of the few studies that captures how precise the values of the many different parameters need to be set if the intervention is to be effective at scale which is consistent with my development experience in developing the HOTS intervention. For example, we quickly discovered through informal trials wherein we varied the intensity of the program that providing the service less than 4 days a week produced little benefit.

There is now sufficient experience with successful application of the principles of practical benefit and improvement science producing major clinical improvements in practice to ask the following empirical question: Are these newer scientific methods more likely to develop and validate interventions with major clinical benefits at scale for complex organizations such as schools and hospitals as compared to the current RCT methods? The answer appears to be "yes," and this has major implications for changing how applied research is conducted, taught, and funded in education—and clinical research in other fields.

5. Conclusions and Recommendations

The prevailing rigorous methodology and statistical criteria used by research journals and government agencies in education research has routinely overstated the actual effectiveness of interventions. Relying on effect sizes to determine the practical significance of interventions in education has been as problematic as using *p*-values to determine statistical significance. Examples have been provided wherein effect size results have misdirected practice in education on a large scale, and evidence has been provided that the same thing has happened in other disciplines where influential published research has not been replicable and/or has misdirected clinical practice partly because of relying on small effect sizes.

In addition, the problem may not just be the statistical criteria that are used, but also the dominant scientific paradigm of relying on RCT research to identify effective interventions that can scale across complex organizations. There is no evidence to date that schools that adopt interventions recommended by the What Works Clearinghouse (WWC) actually improve, and a comprehensive study of its RCT based recommendations on its certified math programs found that the results did not provide useful information to practitioners.

As a result, it is recommended that education switch to a simpler measure of practical benefit that (a) measures the likely benefit of an intervention in simpler, actual outcomes for the experimental group (only), and (b) presents results in the types of measures that leaders can easily understand and that they value as key to their own improvement efforts. Practical benefit also has potential for improving the validity of research to improve clinical practice in other disciplines. It also appears that the alternative scientific approach of improvement science, that relies on looking for patterns of practical benefit in fast prototyping of iterative trials across contexts, has generated major improvements in previously intractable problems of practice in a variety of disciplines, including medical practice and education. Improvement science provides the potential to be a more appropriate methodology for developing interventions that can lead to large-scale improvements in schools than current reliance on gold-standard methods. Several examples of large-scale improvement in education using improvement science and practical benefit were provided. The key question is whether improvement science will be relegated to something that individual practitioners choose to use or whether they will be incorporated as a mainstream methodology for research and development.

Shifting applied education research to incorporate practical benefit and improvement science requires a variety of changes in the nature of knowledge production and editorial policy at the major research journals, as well as in the teaching of quantitative methods to researchers and leaders.

5.1. Recommended Changes in Knowledge Production

5.1.1. American Statistical Association

It may be important to create a dual track of recommendations for reforming inferential statistics—one track for basic research and another for applied research geared toward providing guidance for organizational improvement. Recommendations for reforming applied research should incorporate the principles of practical benefit across the disciplines.

5.1.2. Education

- Research journals need to require that applied research reveal the unadjusted means/medians of results for all groups and subgroups, and require review panels to take such data into account
- Research journals and the US Department of Education need to establish alternative standards for publishing studies and funding the research and dissemination of projects based on the demonstration of consistent practical benefit at scale, and
- Funding needs to be provided to establish the kinds of networks for sharing the results from iterative trials in different contexts similar to what exists in obstetrics.
- The WWC should stop certifying programs as effective, and federal legislation should stop mandating the use of evidence-based programs, until new criteria for identifying effective programs are developed that incorporate practical benefit.

Appendix

A1. Overview of the Project star Experiment

Project STAR was undertaken in 75–79 schools in Tennessee from 1985 to 1989 at a cost of approximately $12 million. Each year in the experimental period, 6000–7000 students in grades K-3 participated in the experiment, for a total of 12,000 students during the entire period.

As students initially entered STAR schools, they were randomly assigned to small classes with 13–17 students, regular classes with 22–25 students without teacher aides, and regular classes with 22–25 students with teacher aides.

Sohn (2010, 2015) identified the following problems with the research:

- Schools volunteered to participate, thus were not randomly selected,

- Kindergarten was not mandatory so the children who enrolled had unique characteristics,
- Students constantly left and entered STAR schools, so (a) the overall attrition rate was almost 50%, and class sizes became distorted from the original design with some small classes ending up with more students than the other categories,
- In the absence of pretest scores, it is not clear whether randomization controlled for initial ability, and randomization did not produce equal levels of low-income students (as measured by free and reduced lunch eligibility) across the treatments,
- Students switched from one type of class to another,
- While teachers were randomly assigned, some received training which may have affected outcomes.

References

Begley, C. G., and Ellis, L. M. (2012), "Drug Development: Raise Standards for Preclinical Cancer Research," *Nature*, 483(7391), 531–533. [225]

Berwick, D.M. (2008), "The Science of Improvement," *JAMA*, 299(10), 1182–1184. [225,231]

Borman, G. D., and Hewes, G. M. (2002), "The Long-term Effects and Cost-Effectiveness of Success for All," *Educational Evaluation and Policy Analysis*, 24, 243–266. [226]

Borman, G. D., Slavin, R. E., Cheung, A. C., Chamberlain, A. M., Madden, N. A., and Chambers, B. (2005), "The National Randomized Field Trial of Success for All: Second-year Outcomes," *American Educational Research Journal*, 42, 673–696. 226

Borman, G. D., Slavin, R. E., Cheung, A. C., Chamberlain, A. M., Madden, N. A., and Chambers, B. (2007), "Final Reading Outcomes of the National Randomized Field Trial of Success for All," *American Educational Research Journal*, 44, 701–731. 226

Borman, G. D., Slavin, R. E., Cheung, A., Chamberlain, A. M., Madden, N. A., and Chambers, B. (2005), "Success for All: First-year Results From the National Randomized Field Trial," *Educational Evaluation and Policy Analysis*, 27, 1–22.

Borman, G. D., Grigg, J., and Hanselman, P. (2016), "An Effort to Close Achievement Gaps at Scale Through Self-affirmation," *Educational Evaluation and Policy Analysis*, 38, 21–42. [225]

Burdumy, J. S., Mansfield, W., Deke, J., Carey, N., Lugo-Gil, J., Hershey, A., and Douglas, A. (2009, June 8), "Effectiveness of Selected Supplemental Reading Comprehension Interventions: Impacts on a First Cohort of Fifth-Grade Students," Mathematica Inc., Presentation at *The Institute of Education Sciences* research conference, available at *http://www.mathematica-mpr.com/~/media/publications/pdfs/education/ies_readcomp_james-burdumy0609.pdf*. [227]

Cohen, J. (1988), *Statistical Power Analysis for the Behavioral Sciences*. Hillsdale, NJ: Lawrence Erlbaum. [225]

CREDO (2013), National Charter School Study, Center for Research on Education Outcomes, Stanford, CA: Stanford University, available at *http://credo.stanford.edu/documents/NCSS%202013%20Final%20Draft.pdf*. [227]

Deke, J., Wei, T, and Kautz, T. (2017), *Asymdystopia: The Threat of Small Biases in Evaluation of Education Interventions That Need to be Powered to Detect Small Impacts*, Washington, DC: Institute of Education Sciences, National Center for Educational Evaluation and Regional Assistance. [225]

Gawande, A. (2007), *Better: A Surgeon's Notes on Performance*, New York: Metropolitan Books. [231]

Ginsburg, A., and Smith, M. S. (2016), *Do Randomized Control Trials Meet the "Gold Standard"? A Study of the Usefulness of RCTs in the What Works Clearinghouse*, Washington, DC: American Enterprise Institute. [225]

Glass, G. V. (2016), "One Hundred Years of Research: Prudent Aspirations," *Educational Researcher*, 45, 69–72. [224]

Hattie, J. A. C. (2009), *Visible Learning: A Synthesis of 800+ Meta-analyses on Achievement*, Abingdon, UK: Routledge. [227,229]

Hojat, M. and Xu, G. (2004), "A Visitor's Guide to Effect Sizes – Statistical Significance Versus Practical (clinical) Importance of Research Findings," *Advances in Health Sciences Education Theory and Practice* 9, 241–249. [223]

Ioannidis, J. P. (2005), "Why Most Published Research Findings are False," *PLoS Med*, 2, available at *http://journals.plos.org/plosmedicine/article/authors?id=10.1371%2Fjournal.pmed.0020124*. [225]

Kerwin, J. and Thornton, R. (2018), "Making the Grade: The Sensitivity of Education Program Effectiveness to Input Choices and Outcome Measures," Paper presented at RISE Annual Conference, Oxford, UK, June 21–22, available at *https://www.riseprogramme.org/sites/www.riseprogramme.org/files/inline-files/Thornton.pdf*. [231]

Kirk, R. E. (1996), "Practical Significance: A Concept Whose Time has Come," *Educational and psychological measurement*, 56, 746–759. [223]

Kraemer, H. C. (2016), "Messages for Clinicians: Moderators and Mediators of Treatment Ooutcome in Randomized Clinical Trials," *American Journal of Psychiatry*, 173(7), 672–679. [225]

Kolata, G. (2015), "A Faster Way to try Many Drugs on Many Cancers," *New York Times*, available at *http://www.nytimes.com/2015/02/26/health/fast-track-attacks-on-cancer-accelerate-hopes.html?_r=0*. [231]

Lipsey, M. W., Puzio, K, Yun, C, Hebert, M. A., Steinka-Fry, K., Cole, M. W., Roberts, M., and Busick, M. D. (2012), *Translating the Statistical Representation of the Effects of Education Interventions Into More Readily Interpretable Forms*, U.S. Department Of Education. Institute of Education Sciences, available at *https://ies.ed.gov/ncser/pubs/20133000/*.

Madden, N. A., Slavin, R. E., Karweit, N. L., Dolan, L. J., and Wasik, B. A. (1993), "Success for All: Longitudinal Effects of a Restructuring Program for Inner-City Elementary Schools," *American Educational Research Journal*, 30, 123–148. [226]

Maul, A., and McClelland, A. (2013), *Review of the National Charter School Study*, National Education Policy Center. Boulder, CO: University of Colorado. [227]

McCartney, K., and Rosenthal, R. (2000), "Effect Size, Practical Importance, and Social Policy for Children," *Child Development*, 71, 173–180. [223]

Open Science Collaboration (2015), "Estimating the Reproducibility of Psychological Science.," *Science*, 349(6251), aac4716-1–aac4716-8. [225]

Plsek, P. E. (1999), "Quality Improvement Methods in Clinical Medicine," *Pediatrics*, 203 –214.

Pogrow, S. (1998), "What is an Exemplary Program and why Should Anyone Care? A Reaction to Slavin and Klein," *Educational Researcher*, 27, 22–29. [226]

——— (1999), "Rejoinder: Consistent Large Gains and High Levels of Achievement are the Best Measures of Program Quality: Author Responds to Slavin," *Educational Researcher*, 28, 24–26, 31. [226]

——— (2000a), "The Unsubstantiated 'Success' of *Success for All*. Implications for Policy, Practice, and the Soul of the Profession," *Phi Delta Kappan*, 81, 596–600. [227]

——— (2000b), "*Success for All* Does not Produce Success for Students," *Phi Delta Kappan*, 82, 67–80. [227]

——— (2002), "*Success for All* is a Failure," *Phi Delta Kappan*, 83, 463–468. [227]

——— (2005), "HOTS Revisited: A Thinking Development Approach to Reducing the Learning Gap After Grade 3," *Phi Delta Kappan*, 64–75. [230]

Quint, J. C., Balu, R., DeLaurentis, M., Rappaport, S., Smith, T. J., and Zhu, P. 2013), *The Success for All Model of School Reform: Early Findings from the Investing in Innovation (i3) Scale-Up. MDRC, available at https://www.mdrc.org/sites/default/files/The_Success_for_All_Model_FR_0.pdf* [227]

Ross, S. M., Smith, L. J., Casey, J., and Slavin, R. E. (1995), Increasing the Academic Success of Disadvantaged Children: An Examination of Alternative Early Intervention Programs, *American Educational Research Journal*, 32, 773–800. [226]

Ruffini, S. et al. (1992), *Assessment of Success for All* [Unpublished research study] Baltimore, MD: Baltimore City Public Schools. [226]

Scammacca, N., Vaughn, S., Roberts, G., Wanzek, J., and Torgesen, J. K. (2007), *Extensive Reading Interventions in Grades k– 3: From Research to Practice*, Portsmouth, NH: RMC Research Corporation, Center on Instruction. [227]

Sohn, K. (2010), "A Skeptic's Guide to Project STAR," *KEDI Journal of Educational Policy*, 7, 257–272. [232]

——— (2015), "Nonrobustness of the Carryover Effects of Small Classes in Project STAR. *Teachers College Record*, 117, 1–26. [232]

Slavin, R. E., Madden, N. A., Karweit, N. L., Livermon, B. J., and Dolan, L. (1990), "Success for All: First-year Outcomes of a Comprehensive

Plan for Reforming Urban Education," *American Educational Research Journal*, 27, 255–278. [226]

Sparks, S. D. (2013, October 30), "School Improvement Model Shows Promise in First i3 Evaluation," *Education Week* (online), available at *http://www.edweek.org/ew/articles/2013/10/30/11successforall.h33.html?qs=sparks+AND+%22Success+for+All%22*. [227]

Sullivan, G. M., and Feinn, R. (2012), "Using Effect Size—or why the P value is not Enough," *Journal of graduate medical education*, 4, 279–282. [223]

Urdegar, S. (2000), *Evaluation of the Success for All Program: 1998-1999,* [Unpublished study]. Dade County: MD, Office of Evaluation and Research, Miami-Dade County Public Schools. [226]

Venezky, R. L. (1998), "An Alternative Perspective on Success for All," in *Advances in Educational Policy*, ed. K. Wong, Vol. 4, Greenwich, CT: JAI Press, pp. 145–165. [226]

Willingham, D. T. (2007). "Critical Thinking: Why is it so Hard to Teach," *American Educator*, Summer, 31, 8–19. [230]

Ziliak, S. T., and McClosky, D. N. (2004), "Size Matters: The Standard Error of Regressions in the American Economic Review," *The Journal of Socio-Economics*, 33, 527–546. [228]

——— (2008), *The Cult of Statistical Significance: How the Standard Error Costs us Jobs, Justice, and Lives*, Ann Arbor, MI: University of Michigan Press. [228]