Avoid evidence-based practices and the What Works Clearinghouse — unless . . .

Stanley Pogrow
October 30, 2023

What scientific evidence can you trust to actually benefit your schools?

Getty Images

At some point you probably sought the best scientific evidence to find a practice that was likely to improve your classroom or school. You may have gone to the federal What Works Clearinghouse (WWC) and adopted one of the interventions it validated as having evidence of effectiveness. You then proudly and confidently implemented it.

And then, after a year or two . . . you probably discovered that your classrooms/schools were no better off. How could that be?

The research community would tell you that because the evidence was scientifically valid, you must have done a poor job of implementing the practice. You feel chastised and that you let your students and colleagues down. However, you also know in your heart and your brain that you implemented the program properly.

What is going on? Are you missing something? Are you lacking as an educator? Isn't everyone else using the practice experiencing success? The answer to these questions is *no*!

You are *not* the problem! The problem was that the evidence-based practice probably was never actually effective, and others probably had the same disappointing experience you did. The problem is that you had the *wrong* evidence generated by the *wrong* scientific process.

The Every Student Succeeds Act requires that educators use Title I funding for practices in the top tier of the What Works Clearinghouse (WWC), meaning that at least one research finding shows strong evidence of effectiveness. The vast majority of these scientifically validated practices are not as effective as researchers claim, and focusing on them is causing educators to misdirect their efforts at improvement.

AT A GLANCE

- Educators are urged to use evidence-based practices to improve instruction but often end up disappointed at the results.
- Common strategies researchers use to demonstrate scientific evidence of success, such as statistical significance or effect size, can make results seem more dramatic effective than they are — thereby misleading practice.
- Educators looking for evidence-based practices should ask where students ended up at the end of the intervention and how big the difference was between students in the experimental and control group.
- The research community, including the What Works Clearinghouse, needs to employ more authentic measures that document actual success in real-world situations.

The doofus approach to using research evidence

To understand how applying current forms of research and evidence leads to bad decisions, consider the following examples.

A getaway gone wrong

A husband and wife, shivering somewhere in the Midwest in December, are planning a January getaway to a warm beach.

> WIFE: I cannot wait for our vacation in January. Let's go somewhere warm.
>
> 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.

While this couple made an evidence-based decision, they are clearly in for a freezing surprise. They are more likely to die from hypothermia in Greenland than they are to get a tan.

The lesson? You cannot just rely on comparison data when choosing between options. You have to know the actual end result of the options you're considering. If the couple in question checked on the actual temperature, they would have discovered that the temperature in Greenland in January averages -8 degrees Celsius with zero hours of sunshine.

An improvement that's not so significant

You want to be environmentally conscious, and you read a research article touting a new car model that gets "significantly better" mileage than your car. You rush out to buy the car. However, you are shocked to discover that the mileage you get is not noticeably better. You are furious at having wasted your money and confront the researcher. The researcher shows you all the inscrutable statistics and even points out the asterisked note that $p < .05$, proving that the improvement was *statistically significant*.

If you had investigated further, you might have discovered that the statistically significant difference was only one mile per gallon, generated when driving at a constant 50 miles per hour on flat ground for 10 minutes. Had you known this, you never would have switched cars.

The lessons? Scientific criteria can make small differences appear to be major discoveries, but the size of the benefit matters, and small differences are irrelevant to practitioners. To have confidence that it's worth adopting a new practice, you should seek *big* benefits. Evidence for those benefits should be generated under conditions of actual practice. And, most important, you cannot make an intelligent decision without knowing the size of the benefit.

How we get duped by the evidence

It's easy to write off the individuals in the above examples as doofuses, but when we decide to adopt an evidence-based practice, we usually end up making the same types of bad decisions. However, we can become smarter by applying the above lessons to ask the right questions about the scientific evidence used to identify effective practices. But first, let's examine how educational research places us into these examples.

Scientific criteria can make small differences appear to be major discoveries, but the size of the benefit matters, and small differences are irrelevant to practitioners.

Controlling for causality

Researchers and the WWC consider the best scientific evidence to be that which shows a practice causes a beneficial outcome. Typically, studies that generate such evidence involve comparing an experimental and control group that are identical in all respects — except that one is testing a new practice. Scientists control for all other confounding factors by randomly assigning individuals or schools to groups, which is the gold standard, or by statistically accounting for any initial differences. The problem is that when you control all the other factors you are creating an artificial environment. An educator is no more likely to have such control than the typical driver is to drive at a steady 50 miles per hour for 10 minutes.

Statistical significance

However, the problem is not just the methodology. We also have to determine whether the superior performance of the group engaging in the new practice is sufficient to conclude that the practice is effective. Historically, the final arbiter has been the statistical significance of the benefit. That is where you see $p < .05$ or $p <. 01$ indicating a result's "significance."

However, statistical significance does not indicate the size of the benefit, and in a large sample a tiny difference can show up as statistically significant — similar to the earlier example of the "significant" evidence of the supposedly more fuel-efficient car. Plus, researchers will sometimes make slight adjustments to get their results over the hump and indicate significance — a process referred to as p-hacking (Ioannidis, 2005). It is simply too easy to use statistical significance to make education practices appear to be effective without any practical benefit. Given this long-standing problem, the prestigious American Statistical Association has called for researchers to stop using it to determine the effectiveness of interventions across all disciplines (Wasserstein, Schirm, & Lazar, 2019).

Effect sizes

Stephen Ziliak and Deirdre McCloskey's (2004) theory of economic significance posits that imprecise findings of big benefits are more important in the real world than precise predictions of

small benefits. Fortunately, there is a way to determine the size of the benefit from the experimental intervention — effect size. But how big should the effect size be to conclude that an intervention is effective? The most widely cited thresholds are those derived by Jacob Cohen (1988), who characterized an effect size of .2 as being small, .5 as being medium, and .8 as being large. Researchers then started to use .2 as the minimum for indicating evidence of an important benefit. (WWC uses .25 as an indication of a substantively important benefit.) However, Cohen characterized .2 as a "difficult to detect" benefit. It is only when the effect size reaches .5 that Cohen characterized the benefit as being noticeable.

Clearly, practitioners are not interested in a benefit that is "difficult to detect." This low threshold makes it easier for researchers to make claims of effectiveness, but it doesn't do much for practitioners.

Even worse, the vast majority of research evidence behind evidence-based practices cannot even meet the inadequate .2 threshold. Hughues Lortie-Forgeus and Matthew Inglis (2019) found that, among 141 government-funded large-scale gold-standard educational research studies in the U.K. and the U.S., the average effect size was .06 — a third of "difficult to detect." Matthew Kraft (2020) concluded that median effects in math cluster tightly between 0.04 and 0.09 across all grades — a quarter to a half of "difficult to detect."

Hacking effect sizes

Rather than using these consistently disappointing results to spur them to rethink their methodologies, the research community has developed a series of strategies to make these microscopic effects seem more notable than they are — a practice I have referred to as ES-hacking (Pogrow, 2019). For example, some have argued that the practitioner community should accept such findings because they are typical within education (see, for example, Lipsey et al., 2012; Kraft, 2020). This is essentially the research community saying about practitioners, "let them eat cake." Another strategy is to provide a relative comparison of results: Practice A, which has an effect size of .2, is twice as effective as Practice B, which only has an effect size of .1. The correct interpretation is that neither practice is effective.

However, the most misleading strategy used in "high-quality" research to make tiny effect sizes appear important is to create seemingly impressive "equivalent" outcomes. For example, a researcher who found an effect size of .2 may claim that it is equivalent to increasing reading scores from the 50th to the 58th percentile. Wow! That is indeed impressive. Alas, they will then fail to show whether the experimental students actually ended up reading at the 58th or the 18th percentile or how much they actually improved. It may be that researchers hide the actual results because revealing them would be embarrassing — or at least not noteworthy enough to mention — or

because the numbers have been so statistically manipulated that it is impossible to know what they actually mean.

We can see this practice of equivalency claims in action in a *Kappan* article on how to interpret research, in which David Steiner (2021) claimed that an effect size of .2 at the high schools is akin to an extra year of learning. Similarly, a 2013 study by the Center for Research on Education Outcomes (CREDO) at Stanford University concluded that Black and Hispanic students made 14 days of additional learning per year in charter schools as compared to traditional public schools. Here, the effect size was even smaller: .02, or a tenth of "difficult to detect." Was it 14 days of extra learning or 14 seconds? Without any indication of what the students' actual learning levels were, these are just disembodied numbers.

In short, the use of equivalencies to make microscopic effects seem impressive only benefits the research community — and misleads practitioners. This is *not* evidence. Matthew Baird and John Pane (2019) have criticized converting effect size results into equivalent amounts of extra learning as producing "implausible" results. Unfortunately, CREDO continues to use this implausible methodology. In their 2023 report, the researchers concluded that Black and Hispanic students in charter schools now made the equivalent of 35 and 30 additional days of learning respectively in reading per year as compared to students in traditional public schools. While the pro-charter funders of this research and *The Wall Street Journal* trumpet this finding, the correct conclusion is that there is no difference on average between traditional public schools and charters. We need to find ways to create more excellent schools of both types.

What this means for policy and practice

Inconveniently, if you just look at actual effect sizes, *only 3%* of the practices that the WWC rated as having the highest tiers of evidence achieved an average effect size of .5, thus meeting Cohen's (1988) standard of having found a noticeable benefit (D'Agostino & Pogrow, 2023).

How has the research community reacted to these disappointing and implausible results? "Let them eat even more cake." The WWC recently reverted to relying on statistical significance. This is going backward to create the illusion of effectiveness where there is none and to increase the number of evidence-based practices it can certify.

This is not just an abstract, academic discussion. Research evidence has been exaggerating the actual effectiveness of interventions and thereby misdirecting practice for decades. For example, consider the actual effectiveness of the program widely considered by the research community to be the poster child for the successful use of research evidence to validate the effectiveness of a program and improve schools: Success for All.

If you have been an education practitioner for an extended period of time, you probably know of a high-poverty school or district that adopted the Success for All reform for accelerating K-6 reading achievement. Launched in 1987, it became widely used because of its supposedly strong research evidence and track record of effectiveness in high-poverty schools — including evidence published in the top research journals. However, from 2000 to 2002, *Phi Delta Kappan* bucked conventional wisdom and published three articles in which I debunked the effectiveness of Success for All. The final article was titled "Success for All Is a Failure" (Pogrow, 2002). I came to that conclusion because of what I had found in approximately 260 Success for All schools across the country, including those where national demonstration projects and published research claimed success. Instead of success, the consistent reality was that the program had actually failed to produce gains or acceptable progress. Despite the tremendous expenditures of money and effort, the Success for All schools were not even doing as well as the other Title I schools using locally developed interventions. As a result, virtually all the districts were dropping the program.

This finding was subsequently replicated in other independent studies, which found equally dismal results (Boulay et al., 2018; James-Burdumy et al., 2009). In a rigorous, independent, evaluation of the effects of a $50 million federal grant Success for All received to expand the program to an additional 1,100 randomly selected schools, Beth Boulay and colleagues (2018) found that Success for All had produced no impact on student achievement and that a surprising number of schools had dropped the program shortly after adopting it. Richard Venezky's (1998) independent analysis of the raw data for the most important experiment, Baltimore Public Schools, found that the Success for All students were actually doing terribly and that there had been some cherry-picking in the sampling. The dichotomy between the published claims of effectiveness and its actual failure is similar to what the couple found when choosing Greenland over Antarctica as a warm vacation spot in January.

Despite all this contrarian evidence, the WWC still lists Success for All as having strong evidence of effectiveness, and progressives seeking to increase equity continue to advocate for its use. For example, *New York Times* columnist Nicholas Kristof endorsed the program in February 2023, even urging readers to contribute to its foundation. I know of only one progressive who actually visited Success for All schools — Jonathan Kozol. In a series of articles in *Phi Delta Kappan*, Kozol (2005, 2006) described a learning environment that was "Skinnerian," using methods akin to teaching dogs to do tricks and that open no doors to understanding. The program, Kozol (2006) said, is "a bottom feeder. It goes to where the misery and hypersegregation are the most extreme. It is an apartheid course of study" (p. 626). How is this social justice? Will progressives and the research community ever acknowledge that they were duped and thereby unwittingly supported a major miscarriage of social justice that still continues decades in?

How to avoid being duped by research evidence

This is not an anti-research article. The use of quantitative (and qualitative) evidence to find effective practices remains a desired goal. In addition, there is a small fraction of evidence-based practices that can be effective in your schools. To find these gems, you can easily apply the lessons from the earlier doofus examples by looking for the following two indicators.

Indicator 1: Where students end up

The most important first step is to find out how the experimental group actually ended up performing (i.e., finding the average temperature in Greenland in January). Then compare the final average result for the experimental group to how your students are already performing. You are looking for the simplest statistic: the average/mean of the actual, final result of the experimental group. The performance of the control group is irrelevant. It's like the temperature in Antarctica, an irrelevant distraction.

Simply find the unadjusted average final status for the experimental group. Do not trust any adjusted or equivalent average/mean. If your school has a high percentage of Latinx students, then look for how the Latinx students ended up performing. Do the same for any other group you're seeking to target for this intervention.

Also, do not trust percentage gains. A practice may have increased the graduation rate by 50%, but that does not necessarily mean that the actual final outcome is better than your existing graduation rate. Focusing only on the actual end result status of the experimental group is how you avoid deciding to take a January vacation in Greenland with summer clothes — or adopting an intervention that is not likely to produce noticeable improvement in your schools.

If you cannot find the end result status, unadjusted, for the experimental group, it is probably not there. It is amazing how often the most academically prestigious research does not provide that result. The same is true for evidence presented by salespeople. Ignore all the asterisks and simply ask: What was the final reading score of the Black males in the sample? If this information is missing, be immediately suspicious! Is the research covering up an actual poor outcome?

Indicator 2: The size of the difference

If you cannot find the end result status of the experimental group but still want to give the evidence the benefit of the doubt, you are left to rely on how big the relative difference was between the experimental and control groups. Remember, when any study claims that the difference was statistically significant or states in the conclusion that an outcome was "significant," this is a technical distinction that is usually not meaningful in practice. You may find yourself in the same position as the car buyer who saw no improvement in real-world driving.

For those studies that rely on effect sizes, ignore any result where the effect size is less than .5. This means the benefit was not noticeable. This indicator does not guarantee that the intervention will

produce noticeable improvement in your schools, but at least you know that it did so in the experimental schools. But always keep in mind that this does not mean that the experimental schools ended up performing better than yours are already doing.

We need different evidence

If we are to reap the benefits from adopting evidence-based practices, we need the research community to reform how it generates evidence.

If we are to reap the benefits from adopting evidence-based practices, we need the research community to reform how it generates evidence. Instead of defaulting to the most academically prestigious methods for judging effectiveness, researchers need to switch to methods that better predict noticeable effectiveness under actual conditions of practice, with confirmation from multiple studies in a variety of school contexts. Until then, the focus on evidence-based practices is generally an inhibitor to improving practice and increasing equity.

So, it is time for the research community to stop:

- Positioning itself as the arbiters of which practices are effective and then telling practitioners that they are at fault when the hypothetical benefits do not appear.
- Exaggerating the benefits of research findings and calling that evidence of effectiveness.
- Hiding how students or schools actually ended up performing and considering practices to be evidence-based even when the students ended up performing terribly.
- Trying to fit a large square peg into a small round hole by imposing the use of gold-standard research methods that are artificial forms of practice. Even Albert Einstein could not make causal methods apply to all aspects of the natural world.

Until the research community starts to offer more authentic evidence of effectiveness at scale in actual practice, the federal government should stop telling schools which practices to adopt. In addition, the current version of the WWC needs to be put on hiatus, reconceptualized, and restarted from scratch. We, the practitioners, should have equal representation in this reconceptualization. Discussions should start by considering what evidence practitioners need to feel confident adopting a practice based on research, and the research community should then apply the best methodologies for producing such data. The same reforms are needed in other entities that identify effective interventions, such as the Evidence-Based Education Center at Johns Hopkins University, and its Best Evidence Encyclopedia.

The problems of research evidence misinforming practice are not unique to education. They have infected other disciplines and resulted in what is referred to as “the replication crisis.” Instead of

continuing practice as usual, the education research community should follow the lead of other disciplines, such as obstetrics (Gawande, 2007) and hospital health care (Berwick, 2008), which have shifted to more intuitive and adaptable scientific research methods to identify effective practices. These disciplines realized that identifying practices that actually saved lives was more important than imposing a singular, academically prestigious, view of science. If education followed suit, it would usher in an era of more valid scientific evidence and better practices — ones that can actually improve schools and increase equity.

References

Baird, M.D. & Pane, J.F. (2019). Translating standardized effects of education programs into more interpretable metrics. *Educational Researcher, 48* (4), 217-218.

Berwick, D.M. (2008). The science of improvement. *Journal of the American Medical Association, 299* (10), 1182-1184.

Boulay, B., Goodson, B., Olsen, R., McCormick, R., Frye, M., Gan., K., . . . & Sama, M. (2018). *The Investing in Innovation Fund: Summary of 67 evaluations. Final Report*. U.S. Department of Education, Institute of Education Sciences.

Cohen, J. (1988). *Statistical power analysis for the behavioral science*s. Lawrence Erlbaum.

Center for Research on Education Outcomes. (2013). *National charter school study II*. Stanford University.

Center for Research on Education Outcomes. (2023). *As a matter of fact: The national charter school study III*. Stanford University.

D'Agostino, J. & Pogrow, S. (2023, April 13). *Identifying highly effective interventions* [Conference presentation]. American Educational Research Association Annual Meeting, Chicago, IL.

Gawande, A. (2007). *Better: A surgeon's notes on performance*. Metropolitan Books.

Ioannidis, J.P. (2005). Why most published research findings are false. *PLoS Medicine, 2* (8).

James-Burdumy, J.S., Mansfield, W., Deke, J., Carey, N., Lugo-Gill, J., Hershey, A., Douglas, A., . . . & Pendleton, A. (2009, June 8). *Effectiveness of selected supplemental reading comprehension interventions: Impacts on a first cohort of fifth-grade students*. U.S. Department of Education, Institute of Education Sciences.

Kozol, J. (2005). Confections of apartheid: A stick-and-carrot pedagogy for the children of our inner-city poor. *Phi Delta Kappan, 87* (4), 265-275.

Kozol, J. (2006). Success for All: Trying to make an end run around inequality and segregation. *Phi Delta Kappan, 87* (8), 624-626.

Kraft, M.A. (2020). Interpreting effect sizes of education interventions. *Educational Researcher, 49* (4), 241-253.

Kristof, N. (2023, Feb 11). Two-thirds of kids struggle to read, and we know how to fix it. *New York Time*s.

Lipsey, M.W., Puzio, K., Yun, C., Hebert, M.A., Steinka-Fry, K., Cole, M.W., . . . & Busick, M.D. (2012). *Translating the statistical representation of the effects of education interventions into more readily interpretable forms*. Institute of Education Sciences, U.S. Department of Education.

Lortie-Forgeus, H. & Inglis, M. (2019). Rigorous large-scale educational RCTs are often uninformative: Should we be concerned? *Educational Researcher, 48* (3), 158-166.

Pogrow, S. (2002). Success for All is a failure. *Phi Delta Kappan, 83* (6), 463-468.

Pogrow, S. (2019). How effect size (Practical Significance) misleads clinical practice: The case for switching to practical benefit to assess applied research findings. *The American Statistician, 7*3 (sup1), 223-234.

Steiner, D. (2021). Make sense of the research: A primer for educational leaders. *Phi Delta Kappan, 103* (3), 43-47.

Venezky, R.L. (1998). An alternative perspective on Success for All. In K. Wong (Ed.), *Advances in Educational Policy* (Vol. 4, pp. 145-65). JAI Press.

Wasserstein, R.L., Schirm, A.L., & Lazar, N.A. (2019). Moving to a world beyond “p< 0.05.” *The American Statistician, 73* (sup1), 1-19.

Ziliak, S.T. & McCloskey, D.N. (2004). Size matters: The standard error of regressions in the American Economic Review. *Journal of Socio-Economics, 3*3, 427-456.

This article appears in the November 2023 issue of *Kappan*, Vol. 105, No. 3, p. 42-49.