Making Education Research Relevant

How researchers can give teachers more choices

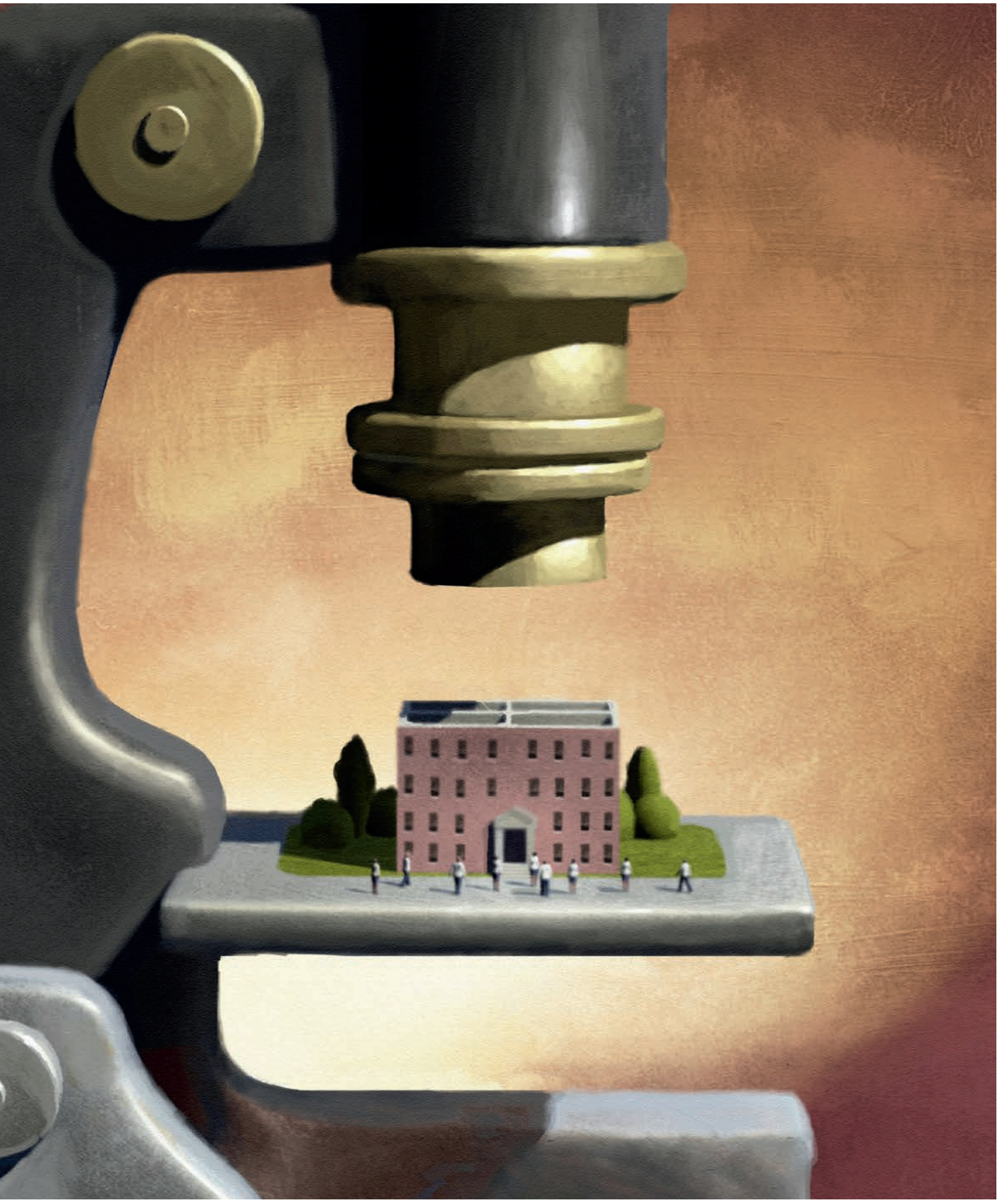
IN THIS JOURNAL, AS IN OTHERS, scientific evidence is regularly invoked in defense of one classroom practice or another. And on occasion, scientific evidence features prominently in federal education policy. It had a star turn in the 2002 No Child Left Behind Act, which used the phrase "scientifically based research" more than 50 times, and an encore in the 2015 Every Student Succeeds Act, which requires that schools implement "evidence-based interventions" and set tiers of academic rigor to identify programs by their proven effectiveness.

Yet teachers, for the most part, ignore these studies. Why?

There's research about that, too. First, teachers may view research as somewhat removed from the classroom, with further translation needed for the practice to be ready to implement in a live setting. Second, teachers may judge a practice to be classroom-ready in general but delay implementation because their particular students and setting seem significantly different from the research context. Third, teachers may resist trying something new for reasons unrelated to its effectiveness—because it seems excessively demanding, for example, or because it conflicts with deeply held values or beliefs about what works in the classroom.

by DANIEL T. WILLINGHAM and DAVID B. DANIEL

STUART McREATH



Finally, teachers may be unaware of the latest research because they only rarely read it.

No matter the reason, it seems many teachers don't think education research is directly useful to them. We think these teachers have it right. And we think the problem lies with researchers, not teachers.

The first three obstacles listed above—two concerning applicability of research and one concerning perceived constraints research puts on practice—are products of the methods researchers use. Research seems irrelevant to practitioners because it does not pose questions that address their needs. Teachers feel constrained by research because they feel pressured to use research-approved methods, and research creates clear winners and losers among practices that may be appropriate in some contexts but not others.

The root of these issues lies in two standard features of most studies: how researchers choose control groups and their focus on finding statistically significant differences. The norm in education research is that, for a finding to be publishable, the outcomes of students receiving an intervention must be noticeably different from the outcomes of an otherwise similar "control" group that did not receive the intervention. To show that an intervention "works," you must show that it makes a positive difference relative to the control. But are such comparisons realistic, reasonable, or even helpful for teachers?

In the real world, classroom teachers—not to mention school and system leaders—are choosing among several possible interventions or courses of action.

No—but they could be. Here's how.

Better Than Nothing Is Not Enough

Let's consider the hypothetical case of CM1, a new method of classroom management meant to reduce the frequency of suspensions. Suppose we recruit eight schools to join an experiment to assess the effectiveness of CM1. We randomly assign teachers in half of the participating classrooms to implement it. We could then compare the rate of suspensions from students in those classrooms to the rate observed in the classrooms that are not implementing CM1. This type of comparison is called "business as usual," because we compare CM1 to whatever the comparison classrooms are already doing. A similar choice would be to compare the rate of suspensions before CM1 is implemented to the rate after it's implemented within the same schools. This "pre-post" design is comparable to the business-as-usual design, but each school serves as its own control.

If suspension rates are lower with CM1, we can conclude that it "worked." But with a business-as-usual control group this conclusion is weak, essentially that "something is better than nothing." Even that may be too optimistic. We might be observing a placebo effect—that is, students behaved differently only because they knew they were being observed, or because something in their classroom changed. Or maybe CM1 isn't especially effective, just better than whatever the teachers were doing before, which might have been actively harmful.

We can draw a somewhat stronger conclusion if we use an "active control," which means that control classrooms also adopt a new method of classroom management, but one that researchers don't expect will affect suspension rates. Active-control designs make researchers more confident that, if a difference in suspension rates is observed, it's really CM1 that's responsible, because both CM1 classrooms and control classrooms are doing something new. This model means we need not worry about placebo effects or that CM1 merely prevented ineffective practices. However, even the best-case scenario produces a weak conclusion, because the control method was predicted not to work. It's still "something is better than nothing."

Still another type of comparison tests an intervention that's known to be effective against a newer version of the same intervention. The goal, obviously, is to test whether the new version represents an improvement.

The three research designs we've considered answer questions that will often be of interest only to researchers, namely, whether CM1 "works" or, in the case of the old versus new version comparison, whether CM1 has been improved. When "works" is synonymous with "better than nothing," the answer can be important for distinguishing among theories and hence is of interest to researchers. But is this question relevant to teachers? Practitioners are not interested in theories and so would not ask, "Is this program better than nothing?" They would ask something more like, "What's the best way to reduce suspensions?"

The answer "CM1 is better than nothing" is useful to them if no other interventions have been tested. But in the real world, classroom teachers—not to mention

school and system leaders—are choosing among several possible interventions or courses of action. What about other methods of classroom management intended to reduce suspensions? If, say, hypothetical classroom-management program competitors CM2 and CM3 have each been shown to be better than nothing, practitioners would prefer that researchers compare CM1 to CM2 and CM3 rather than compare it to doing nothing at all. Is one much better than the others? Or are all about equally effective, and it's up to practitioners to pick whichever one they prefer?

AP PHOTO / RON EDMONDS

The 2002 No Child Left Behind Act used the phrase "scientifically based research" more than 50 times.

Best Practices—But for Whom?

If we set a goal of finding the best way to reduce suspensions, and there are no successful interventions known, comparing CM1 to business as usual makes sense. However, if there are successful interventions known, researchers should compare CM1 to what is currently thought to be the most successful intervention. We might think of this as the strong definition of the term "best practices." It indicates that there is one champion method, a single preeminent way of reducing suspensions, and the goal of research is to find it.

But that's generally not how the world works and indeed, "What's the best way to reduce suspensions?" is probably not *exactly* what an educator would ask. Rather, they would ask, "What's the best way to reduce suspensions at my school, with the particular students, faculty, and administrators found here, and with our peculiar set of assets and liabilities, and without negatively impacting other important instructional goals?"

CM1 may be terrific when it comes to reducing student suspensions, but it may also be expensive, demanding of administrators' time, or workable only with very experienced teachers or with homogenous student bodies. And maybe CM2 is also terrific, especially for inexperienced teachers, and CM3 is helpful when working with diverse students. Research certainly shows such variability across contexts for some interventions, and teachers know it. As we've noted, one reason teachers don't tend to use research is because they assume that whatever positive impact researchers found would not necessarily be the

same for their particular students in their particular school.

If a universal champion "best practice" really emerges, improbable as that seems, it would be useful to know, of course. But teachers would benefit most not by researchers' identifying one program as *the* best, but by their identifying or broadening a range of effective interventions from which teachers can then choose. Research can support that goal, but it requires a change in what we take to be an interesting conclusion. Instead of deeming a study interesting if the intervention is better than the comparison group, teachers would be interested in knowing whether a new intervention is *at least as good as* the best intervention. That would allow them to choose among interventions, all of which are known to be effective, based on which one they believe best fits their unique needs.

Null (and Void) Hypothesis

But that's not the goal of research studies. Researchers are looking for differences, not sameness, and the bigger the difference, the better. Teachers might be interested in knowing that CM1's impact is no different than that of another proven classroom-management method, but researchers would not. Researchers call this a null effect, and they are taught that this conclusion is difficult to interpret. Traditionally, research journals have not even published null findings, based on the assumption that they are not of interest.

Consider this from a researcher's point of view. Suppose a school leader implements CM1 because the leader thinks it reduces suspensions. There are 299 suspensions in the school that year, whereas in the previous year there had been 300. Did CM1 help? A researcher would say one can't conclude that it did, because the number of suspensions will vary a bit from year to year just by chance. However, if the difference were much larger—say there were 100 fewer suspensions after CM1 were put in place—then the researcher would say that was too large to be a fluke. A "statistically significant difference" is one that would be very unlikely to have occurred by chance.

This logic undergirds nearly all behavioral research, and it leads to an obsession with difference. Saying "I compared X and Y, and I cannot conclude they are different" because the outcomes were similar may be uninteresting to researchers, but it is potentially very interesting to practitioners looking to address a particular challenge. They would be glad to know that a new intervention is at least as good as a proven one.

Null effects matter for another reason. Interventions often spring from laboratory findings. For example, researchers have found that memory is more enduring if study sessions are spread out over time rather than crammed into a short time period. We should not assume that observing that effect in the highly controlled environment of the laboratory means that we're guaranteed to observe it in the less controlled environment of the classroom. If spacing out study sessions doesn't work any better in schools than cram sessions, that's a null effect, but it's one that's important to know.

Researchers are right that null effects are not straightforward to interpret. Maybe the intervention *can* work in schools, but the experimenters didn't translate it to the classroom in the right way. Or they may have done the translation the right way, but the experiment the wrong way. Nevertheless, null effects are vital to tally and include in a broader evaluation of the potential of the intervention. Researchers can make null effects more readily interpretable through changes in research design, especially by increasing the number of people in the study.

Publication Bias

How do these phenomena play out in recently published research? To find out, we did some research of our own. We examined a sample of articles reporting intervention studies published from 2014 to 2018 in four journals: *American Education Research Journal, Educational Researcher, Learning and Instruction,* and *Journal of Research in Science Teaching.* Our analysis looked at the type of control group employed and whether the intervention was reported to be significantly different from the control group. We predicted that most published articles employ weak control groups—those allowing the conclusion "better than nothing"—because these offer the greatest chance of observing a significant difference between intervention and control.

Of 304 studies examined, 91 percent were of the "better than nothing" sort: 49 percent employed business-as-usual designs and 42 percent used as the control group an alternative intervention that researchers expected not to influence the outcome. Some 4.5 percent used a control that was a variant of the intervention with the goal of improving it. Another 4.5 percent used a control group that was either known to have a positive effect or was expected to have a beneficial effect based on existing theory.

Coders also noted whether the key comparison—intervention versus control—was reported as a statistically significant difference and whether a particular interaction was emphasized. For example, perhaps the intervention group performed no better than the control group in early grades, but there was a significant difference in later grades. Alternatively, the key conclusion of the report may have been that the

intervention and control group did not differ.

We found that 91 percent of the studies reported that the intervention was significantly different than the control group. Of those that did not, another 4 percent reported a significant interaction—that is, the intervention worked for certain subjects or under certain circumstances. Just 5 percent of studies reported null effects. None of these studies demonstrated that a new intervention is equivalent to another intervention already established as effective.

A More Useful Research Standard

In theory, the goals of education research are to build knowledge and improve decision-making and outcomes for teachers and students. But in practice, education research is shaped by the common practices and priorities of researchers, not teachers or school and system leaders. Most intervention research employs a better-than-nothing control group, and an intervention is deemed worth applying (or, at least, worthy of continued research) only if it makes a measurable and statistically significant difference. The drawback to this pervasive research design is clear: there may well be "research-based" interventions in the marketplace, but educators have no basis on which to compare the alternatives. They have all been shown to be "better"—but better than what, exactly?

> The term "best practices" indicates that there is one champion method, a single preeminent way of reducing suspensions, and the goal of research is to find it.

Imagine instead that the common research design started with whatever trusted intervention is considered the current "gold standard" for the desired outcome and used that as the control group. Imagine too that the criterion of the comparison would be that a new intervention should be at least as good as the gold standard. In time, a group of proven interventions would emerge, roughly equivalent in effectiveness and known to be superior to other interventions not up to the gold standard. As a result, educators would have a range of high-quality interventions to choose from and could select the one that best fits their school context, skills, and personal preference. In addition, choice itself can be an important component of educational effectiveness—interventions with teacher buy-in tend to be more successful, and research has shown that the pervasive adoption of a single intervention that does not suit the broader array of individual differences may lead to less learning.

We see other benefits to adopting this approach as well. We predict that refocusing research on equivalence as the dissemination criterion will spur innovation. "At least as good as" is actually "better than" if the new intervention has fewer side effects, is less expensive, is less time-consuming, or is easier to implement compared to its predecessor. For example, consider electronic textbooks, which are less expensive to disseminate and easier to update. The salient question for educators and policymakers isn't whether they are better than other texts, but whether they are associated with learning outcomes equivalent to those of using traditional, more costly textbooks. The research field's narrow focus on ensuring the intervention is statistically "better than" the control group means that the workaday demands of the intervention in terms of time, money, space, and personnel are not emphasized—in fact, are often not even considered. This disconnect invites skepticism on the part of the teachers charged with implementing supposedly classroom-ready practices.

What will it take to effect this change? We believe researchers are sensitive to the incentives their profession offers. Most education research is conducted in the academy, where the coins of the realm are grants and peer-reviewed publications. There are some encouraging signs that journal editors are taking a greater interest in null effects, such as a recent special issue of *Education Researcher* dedicated to such studies. But change will most likely come about and endure if the foundations and government agencies that fund research make clear that they will view this change in study designs favorably when reviewing proposals. This would encourage journal editors to publish studies with null effects and reject those that use business-as-usual control groups.

Researchers are, in our experience, frustrated and saddened that teachers do not make greater use of research findings in their practices. But nothing will change until the researchers recognize that their standard methodology is useful for answering research questions, but not for improving practice.

Daniel T. Willingham is a professor of psychology at the University of Virginia. David B. Daniel is a professor of psychology at James Madison University.