January 2018

Meeting Our Standards for Educational Justice: Doing Our Best with the Evidence

By Kathryn Joyce and Nancy Cartwright

Recently, the U.S. has shifted from conceiving of equal educational opportunities as equal resources, or inputs, for all students to calling for adequacy standards, a threshold level of outcomes that all students should achieve. We shall discuss strategies to help educators choose programs and policies that can help in achieving these standards. This shift raises significant normative issues, many of which continue to attract considerable attention. Is justice served by de-emphasizing the *distribution* of educational goods, whether resources or achievements, and focusing on thresholds for everyone? Does justice indeed demand equality of opportunity in education? Is that end reasonably captured by equal achievement of threshold educational standards across gender, race, income, and ethnic background? Are the standards chosen up to the job? Important as they are, these are not our issues. We shall here take as given the current assumption in the U.S. that justice requires ensuring that every student meets some threshold educational standards and we shall not debate whether the standards chosen are good ones to achieve the aims of educational justice. Our concern is with what educators can do to meet these standards, potentially narrowing the considerable and persisting gap between students who achieve—or exceed—adequacy and students who fail to meet even less rigorous standards. Settling on what justice ultimately requires is surely important. But equally, to move toward greater educational justice in present circumstances we must figure out how to meet the concrete goals set in aid of justice. That is the topic we address.

In recent decades, policy-makers have embraced an evidence-based approach to address low levels of academic attainment among disadvantaged students. The No Child Left Behind Act (NCLB) implemented 'evidence-based policy' (EBP) in 2001 with the explicit aim of narrowing socioeconomic- and race-based achievement gaps and supplying equal opportunity.[1] NCLB requires collecting data on student performance and holds educators accountable for meeting expectations. It employs EBP alongside its accountability measures because it (rightly or wrongly) attributes achievement gaps largely to poor quality teaching and curriculum in schools serving disadvantaged students. By widely disseminating research about the efficacy of interventions, EBP offers high-quality, data-driven methods to educators across districts. NCLB and its successors aim to improve opportunity for and achievement among disadvantaged students by requiring low-performing schools that receive federal funding to use evidence-based practices. For EBP to raise educational

quality, educators must be able to use the available evidence effectively within their schools. We here provide guidance for how to do that.

Although it seems that using scientific evidence in deliberating about strategies to help disadvantaged students should lead to improvements, the attention, effort, and expenditure paid to figuring out 'what works' in education has not produced the desired results. There are many reasons EBP in education disappoints, including failure to uptake and poor implementation. But even when educators attempt to integrate evidence into their practices by using the research available, the results are often dissatisfying. The primary reason probably lies in the difficulty of the task itself: There is no short road to better outcomes, more equally distributed. Moreover, the root sources of students' problems often lie outside the control of educators. However, we argue that there is also a problem of method. Researchers conduct studies to identify efficacious educational interventions. Educators are then tasked with choosing and implementing them. But research doesn't wear its implications on its face. Educators must figure out what the evidence means for their situation and make predictions about the outcomes of interventions prior to implementation. Proponents of EBP pay little attention to this crucial task.

The EBP movement in education has invested heavily in designing better, more cost-effective randomized controlled trials (RCTs), vetting studies, and disseminating results through various databases that organize and evaluate research for educators. Though RCTs are advertised as showing what *works*, what they actually show is that the intervention tested *worked somewhere*—in some particular setting(s), in some particular population(s), at some particular time(s). Since it has worked somewhere (or, in a number of somewheres), an RCT-backed intervention can serve as a starting point for considering solutions to a problem, but educators need to know much more to determine that it *will work* in their local context. Unfortunately, there has been far less effort devoted to figuring out how to fill the gap between what *has worked* and what *will work* than to ensuring the quality of the studies that show what has worked. This is in part because this kind of information is much more difficult to assess rigorously and so what information there is, or might be gathered, is less often reviewed and disseminated. Educators are left on their own to make these evaluations, without the benefit of what information is available, albeit information that is not so rigorously vetted for its accuracy – an exemplary case of the best being the enemy of the good. This shortfall, we think, has contributed to the dismal results we have seen despite significant investment in educational research over the past two decades: Hard-won research results are not being used to

best advantage. We aim to address this gap by examining what educators need to know and to do to make more reliable predictions about the outcomes of available interventions.

We want to underline that there are no rigorous procedures to follow. It is a matter of due diligence in gathering information, of critical deliberation, and of judgment. Educators need to estimate how local conditions will affect outcomes. This is especially important where student outcomes are steadfastly below the adequacy threshold because local factors that impede achievement may impact candidate interventions. Recognizing this can have implications further back in the research chain as well, by, for example, encouraging researchers to investigate causal mechanisms. We urge that far more effort be invested in developing methods for learning what general conditions affect effectiveness and for reviewing and disseminating our best bets about it, even though we think that a catalogue of features that tend in general to moderate effectiveness will never be sufficient for accurate prediction.

I. Preliminaries

Recent philosophical work observes that EBP in education lacks an adequate conception of evidence (Cartwright, 2013; Kvernbekk 2011, 2016; Phillips 2007). Lacking a clear sense of what evidence for a claim is generates unwarranted policy predictions. Discussions within educational research and policy tend to treat 'evidence' as a basic notion, but philosophical treatments reveal its complexity (Reiss, 2014). For educators to use EBP successfully, we must clarify what it means to say that a policy is 'evidence-based.' Sometimes 'evidence-based' is taken to mean that a conclusion (in this case, a conclusion about the effectiveness of an educational intervention) is *derived* or *inferred* from scientific studies. But studies do not imply policies. Rather, studies test hypotheses. If the intervention yields a positive result—it produces a positive effect size—in well-conducted RCTs, the studies *support* the claim that it worked in the study settings. That, then, can count as a *fact*. But when is this fact *evidence* supporting the effectiveness of the intervention in a specific setting? According to the theory we urge (the *argument theory*), facts produced by RCTs—or any other facts—count as evidence for a hypothesis when they serve as premises in a good argument with that hypothesis as the conclusion (Cartwright and Hardie 2012; Cartwright 2013).

Good arguments are sound—they include only trustworthy premises that, taken together, imply the conclusion. A premise is trustworthy when it is supported by good reasons, which can include empirical research results, observation, and credible theory at various levels. What counts as a 'good' reason or 'credible' theory depends on particularities of the case. We can say at least that

good reasons are *relevant* to the conclusion and *independently trustworthy*. Thus, premises must be supported *by* evidence to serve *as* evidence in a good argument. We can say that a claim about policy effectiveness is evidence-based, then, when it is supported by a well-evidenced argument. In EBP, 'evidence-based' usually indicates that one or more of the premises in the argument is backed by scientific research. Usually this premise claims that the policy worked in some specific study sites. Our concern is with the other premises needed to complete the argument and with what supports them. In what follows, we use the argument theory of evidence to explore what educators need for warranted predictions.

Before moving on, it is worth noting that many thinkers criticize the shift to EBP in education. Critics point out that educational contexts are value-laden, so values inevitably shape policies and practices (Biesta, 2007; Smeyers and Dapaepe, 2006). This leads some to dismiss educational research altogether and others to minimize its role (see Walters et al., 2009). Those who accept EBP often criticize narrow reliance on RCTs, arguing that other kinds of research and normative theory can usefully inform policy (Bridges et al., 2008). On their view, research can be relevant to policy even if it does not provide evidence of an intervention's effectiveness. Typically, they argue that while RCTs may be well-equipped to answer questions of effectiveness, alternative research types and theory can inform policy by helping policy-makers understand contexts, problems, and available options.

We agree with these basic claims but not with all the lessons drawn from them. In defending the role of philosophical and normative contributions, critics tend to de-emphasize predictions about effectiveness for education policy. If educational justice is a chief aim of EBP, though, this trend is misguided. We do not deny that norms should be central to deliberations about which policies to choose and how to implement them; they may even trump all other considerations. But predicting effectiveness also matters if educators are to enable more students to meet adequacy standards. To do so, educators must focus on causation so that they choose policies that can produce the relevant outcomes for their students. Research on how different students learn and which policies can cause desired effects under certain conditions—including value-laden conditions—cannot be excluded from their deliberations.

A great deal of detailed local knowledge is essential for reliable prediction. But research can play a big role too. Research can uncover important features that a situation must have if a particular causal pathway is to be possible; it can identify factors that tend to promote the effectiveness of a policy and factors that tend to impede it; it can supply new concepts to describe local phenomena;

and it can offer theoretical perspectives that help in designing implementation. Because the stakes of getting it right are high, especially for students performing below the threshold, we think educators should spend *more* effort engaging with scientific research and considering how the causal evidence it generates can best be used to achieve the goals set for educational justice.

II. Deliberating within Educational Contexts

We start by examining the structure of the educators' choice-situations to identify what belongs in good arguments for education policy. Educators work within a particular context defined by many features, including location, institution and student histories, available resources, and historical and current policies and practices. They work with a particular set of students who differ along several dimensions from each other and from peer groups in other locations. Students show different levels of learning readiness, motivation, and aptitude for various subjects. Within their setting with their students, teachers and administrators must plan curricula so their students can meet adequacy standards. These standards focus on individual outcomes, leaving the means for reaching them to districts and schools. For example, California stipulates that by eleventh grade students should be able to write argumentative, narrative, and informative essays. Accomplishing this requires educators to make a host of decisions about how to teach these skills to their students. To make good decisions, they must consider a myriad of factors: how students currently perform relative to the standards, why they perform as they do, school resources available currently and in the foreseeable future, and which strategies have been effective and ineffective historically—to name a few. With all this in view, educators must predict what will work best for their students in their setting, with the complicated network of factors there that will affect those outcomes.

Put more technically, educators must draw *singular* causal conclusions: conclusions about a causal connection in a specific individual case, like this school, this class, or this student. It is often argued that to warrant such claims we need to compare a case where the intervention was used and a case—identical in all aspects relevant to the putative effect—where the intervention was not used (see Menzies, 2014). Whether or not a comparative methodology is ideal for establishing causal claims, given the complexity of educational settings and differences among students, a twin case for comparison is generally—if not always—unavailable.[2] But, this does not imply that educators cannot make good arguments to support singular causal claims. We regularly draw reliable causal conclusions without assistance from established counterfactual cases. Courts rely on juries to consider evidence and draw conclusions about whether a defendant committed the crime without

consulting a case that excludes the defendant's activities but is otherwise identical. We find bugs in software or identify the source of mechanical failures even where nothing like the bug or failure has happened before. Even more commonly, we confidently infer who made a mess of the kitchen or borrowed something without asking. Educators must draw casual conclusions under similar circumstances—they must predict and bet on what will happen to *this* student or *this* class in *this* setting without a comparison case.

Singular causal conclusions can be true or false and the reasoning for them can be better or worse. The strength of the warrant depends in part on the strength of the evidence used to reach the conclusion. Educators find guidance in research that purports to show 'what works,' which sounds very general. But rigorous studies identifying that a causal relationship between an intervention and an effect held in one, or even several, settings show only that the intervention *can* cause the effect under some circumstances. A claim about what happened in other cases is only one part of an argument for a conclusion about a different case.

The first step that we urge educators to take in thinking through what might work in their setting is 'ex ante' causal analysis—analysis of what currently contributes to the undesirable outcomes they are experiencing in order to understand the problem they are trying to solve. Seemingly similar problems can have different underlying causes. For example, reading below grade level might stem from lack of books and reading at home for students in one school and from instruction techniques in another. Tutoring can improve reading. So, if low reading skills stem primarily from lack of books and reading at home, in-school, small-group tutoring might work for these students. By contrast, if low reading skills is primarily caused by poor teaching, tutoring might not be effective. Locating the cause of the problem directs us to a different set of possible interventions. Perhaps the teachers perform poorly because they are inexperienced. If so, a good course of action may be to provide training. If instead the teachers are skilled but overworked, both training and tutoring may be bad options. Hiring teaching assistants or reducing class sizes might be more effective.

Second, educators need to identify *support factors* for the intervention to produce the targeted outcomes. Causes rarely act alone. Although it is common to say that interventions *cause* or *produce* effects, this usually means that the intervention *contributes*—along with a set of generally unmentioned support factors—to the effect. As with other causes, an educational intervention can only produce the effect if the requisite support factors are present. Consider quizzing on material

students will be tested on later as a strategy for promoting retention and improving performances on tests. What must be in place if quizzing is to work?

- Quizzes are delivered
- Material on quizzes corresponds with test material
- Teachers have time to create tests well in advance and corresponding quizzes
- Students understand the material on the quizzes
- Time to study or practice
- Appropriate study space
- Correct-answer feedback
- Student motivation

These support factors are essential for quizzes to contribute to better retention and testing performance.[3] We can think of them as ingredients in a 'causal cake' and arrange them graphically:

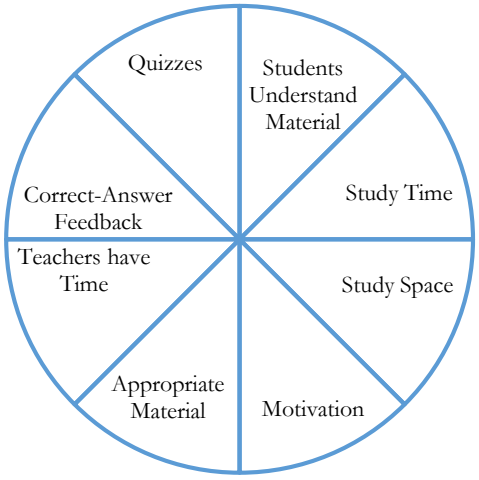

Figure 1: Causal Cake

Causal cakes are diagrammatic representations of support factors that are conjointly sufficient for producing an effect.[4] 'Cake' is an apt metaphor: Making a cake requires all the necessary ingredients. If we can produce a cake without some ingredient, it is unnecessary. The cake itself is sufficient to contribute to the effect, but it is not necessary because we could improve test scores by implementing other policies—represented by different cakes.[5] By 'sufficient' we mean that it is highly probable that a contribution to the effect will occur if the whole cake is in place (though note that this contribution can be offset by negative contributions from other cakes).

So, for quizzing to produce a positive effect, its support factors must be present to a sufficient degree. Imagine that students did not understand the material introduced because of poor instruction. They may do poorly on quizzes, bringing down their grade. The time spent quizzing might be better spent teaching the material prior to the test in that case. Or, a teacher could solve this problem by providing correct answers but not grading the quizzes. Notice that doing so might

interfere with another support factor if students are only motivated to perform well on quizzes that affect their grades. One must ensure that efforts to supply some support factors do not detract from others.

We cannot include everything that matters in the cake. We assume some things that are likely to be ingredients in any causal cake for an educational intervention—that teachers and students have a certain level of rapport, or that students attend school regularly. To be sure, the quality of teaching and the individual student-teacher interaction is likely to affect any intervention. Moreover, we cannot anticipate all support factors that bear on the outcome. They are not static and some are not discernible in advance. Presence or absence of other factors within—or outside of—school can support or undermine interventions. Even when implementing quizzing with all support factors in place, the outcome is uncertain and the size of the effect is unknown. Other school or classroom policies might bear on the results in unexpected ways. Quizzing could disrupt learning, for instance, if it caused unproductive anxiety. Although it is not possible to be sure about all the support factors and their effects, thinking about them is important. For example, if success from quizzing relies on students having a study period during the school day, then it is advantageous to make sure new policies will not remove the study period without finding some substitute.

Take another example. The flipped classroom is a blended-learning model designed to individualize instruction by having students watch lectures at home so teachers can use class time for assignments that might otherwise be homework. Studies show positive outcomes and many teachers report impressive results (Esperanza et al., 2016; Means et al., 2013). Others report dismal results, even if they have carefully followed instructions for implementation. Considering support factors should help educators predict whether a flipped classroom could improve learning outcomes for them. For the flipped classroom to work, students need access to resources outside of class—computers with the internet, software, DVD players, and a quiet workspace, to name just a few. Teachers must be effective lecturers and excel at teaching in a project-based classroom. Students must be able to grasp information without asking questions while learning the material at home. Time outside of school is crucial. What happens in other classes also matters—if many teachers or whole schools adopt this model, students would need several hours each night to prepare for their classes.

Third, educators must consider whether aspects of their environment external to their school might obstruct or dilute the outcome—events like loss of a community partner or changes in public transportation. Socio-economic factors that can impact children's abilities to learn are

important aspects of the environment (Booth and Crouter, 2008; García and Weiss, 2017; Rothstein, 2009; Timar, 2012). Some factors that put children at risk of underperforming in school due to low learning readiness stem from differences in family and community resources. Primary risk factors include stress within a family, poor social and economic conditions, and poor or unstable health. Socio-economic factors also matter for estimating things like parental involvement, how much time students have for homework, and rates of student turnover. Although educators cannot control these, they are crucial support factors—both positive and negative—determining which interventions will work for which students.

Consider programs that provide free breakfast to increase learning readiness. Environmental factors matter to how to structure the program most effectively. If breakfast is served thirty minutes before school begins, students must be able to arrive early enough to attend, which depends on how they get to school. If most use public transportation and have the choice between arriving an hour or more before school opens or just in time for classes, holding the program thirty minutes before school could reduce its effectiveness. Alternatively, the school could deliver breakfast to students who qualify during their first class. But this too may be undermined by environmental factors. For example, if students come from a wide range of socio-economic backgrounds, delivering breakfast may stigmatize recipients or trigger other social ramifications that are less likely if breakfast were available to every student regardless of need or if students had similar backgrounds.

Fourth, educators must consider how social factors like local norms and values within the school and wider community serve as positive support factors or detract from an intervention's effectiveness. For example, local norms or values might make sufficient uptake unlikely. Consider Amanda Datnow's (1998) analysis of restructuring efforts at what she calls 'Central High School' in 1992. In response to the lack of academic and social guidance students received at home, a group of teachers proposed to provide academic, career, and personal support to all students. They based their plan on analysis of the problems interfering with their students' success and research about effective interventions. Their proposal included a thorough assessment of what we call support factors and what they would need to secure them: They had plans for learning centers, individualized programs, study skills classes, mentors to assist with motivation and goal setting, and a partnership with a local university for teacher and student training. The district, school board, university, and teachers' union endorsed the proposal. The state accepted and funded the program, providing 1.3 million dollars over five years.

The restructuring failed despite funding, time, support from administrators and the community, a clear vision, and many teachers on board. Datnow attributes the failure to conflicting ideologies and power dynamics among teachers. Although most teachers were on board, resistance from a small group of experienced, entrenched male teachers undermined the entire project. The resistant teachers refused requests to change their teaching styles. They refused to play the role of mentor and were unsympathetic to students who did not perform well or lacked motivation, regardless of their disadvantage. Whereas those in support of the proposal viewed all students as capable of succeeding in school and deserving of support and opportunities, those against it held that some students were just more capable than others and directed their efforts toward students with talent and motivation. The small group of resistant teachers insisted on traditional beliefs and approaches to education and associated the new proposal with femininity and feminism.

This example illustrates how aspects of the environment and social factors might obstruct an outcome even if other conditions are met. Several levels of the environment interacted to cause an obstruction. Within the school, a small set of teachers were able to exert significant power over the administration. Some of their power came from relationships within the community. They were not easily sanctioned or removed for refusing to abide by new school policies. In fact, they forced a key administrator to resign. According to Datnow, their power primarily stemmed from traditional, dominant ideology. Instead of making arguments about different approaches to education, the resistant teachers relied on gender norms and stereotypes, presenting the issue as a choice between traditional values and progressive feminist values. They targeted the female teachers driving the reform, making sexist and derogatory comments about them personally. Such tactics would not work without a background of gender inequality.

Critics of EBP in education use this sort of case to argue that educational research is trivial for creating effective policy within educational contexts comprised of values, norms, and power relations. Instead of rejecting EBP, we respond by emphasizing context as a source of evidence for good policy predictions. Indeed, the five issues we highlight concern varying aspects of context. Had the contextual challenges at Central High been visible at the proposal stage, the educators could have made efforts to restore authority to the administration, for example. In reply, critics might say that it is undesirable to interfere with local contexts, advising policies that fit with Central High's values, norms, and power structure. But, stressing, as we do, that the efficacy and desirability of policies depend on context, does not imply favoring policies that can work in the existing context over adjusting context to fit policy. It would be unfair to maintain the status quo at Central High just

because it coheres with local culture, and dominant ideology more broadly, when good arguments indicate that changes could improve achievement among disadvantaged students. Of course, deciding when to change the context to support an intervention requires assessing whether doing so is good overall. At Central High, for instance, restructuring the power dynamics among faculty and administrators may improve the context independent of policies it enables.

Finally, for reliable predictions, educators should consider *concatenation*. That is, whether the local arrangements allow the process to go through start to finish. Interventions involve teams of causal factors operating through involved and, often, long sequences. If local arrangements disrupt some of the intermediary steps or the underlying structure is not geared to support the required causal pathway start to finish, the intervention will not produce the effect. This may occur, for instance, when an intervention depends on irreplaceable resources. Educators should identify which resources are irreplaceable and do their best to determine whether they will be there throughout the causal process. For example, if an intervention relies on one enthusiastic teacher, the likelihood of their participating through the end of the program matters. Or, if the intervention has multiple phases but funding is only in place for the first, educators must make a judgment about how likely it is that they will secure funding for the later phases.

Often, obtaining the relevant evidence for thinking through these five issues is difficult. Though it is not always clear what to look for or how to acquire the information needed, knowledge of the local context is essential. This puts educators on the ground in the best epistemic position to obtain some of the crucial evidence. As such, reliable policy predictions require collaboration with educators. Top-down policies imposed externally can be a threat to this, especially if they are applied to settings with relevant differences.

Figuring out what can work, while essential, is only one part of crafting good policy. Educators must answer many other questions before deciding what to do. They must weigh the costs and benefits of intervening. Based on their aims, values, and limitations, they must decide which interventions are options for them. Which would they be willing and able to use if they have a good argument that the intervention is likely to work for them? These considerations are essential. We focus on deciding what will work because doing so is crucial for meeting the standards set in aid of educational justice. But we want to reiterate that values and local norms are important for both tasks—figuring out what will work and which policy is the best choice overall.

Although warranted predictions require multiple well-evidenced premises, educational research largely focuses on the causal relationship between intervention and outcome in study

settings. It is often presented as showing 'what works' on its own, but characterizing educational research this way is misleading. Having discussed some premises educators need for reaching causal conclusions in their individual cases, we will now consider the role evidence from educational research—namely RCTs—can play in arguments for singular causal claims.

III. How RCTs Can (and Cannot) Inform Education Policy

To facilitate EBP in education, various governmental agencies like the U.S. Department of Education's What Works Clearinghouse (WWC), non-profit organizations, and universities have created extensive databases organizing and evaluating studies devised to demonstrate the efficacy of educational interventions and strategies. When the studies meet their criteria, the WWC declares that they work. For example, *Coping Power* is designed to help students with social and emotional deficiencies cultivate necessary skills for their transition to middle school. It is classified as having strong evidence of a positive effect on external behavior based on multiple RCTs.[6] Similar resources are available from the Coalition for Evidence-Based Policy, which collaborates with policy-makers and published a guide to choosing and implementing evidence-based interventions for the U.S. Department of Education. Many universities also supply websites that advertise evidence-based programs directly to educators.[7]

Attempting to follow evidence-based medicine and a widespread movement for EBP more generally, much educational research relies on RCTs which are widely considered gold standard evidence for 'what works'. Only RCTs can meet the WWC's 'Group Design Standards without Reservation' and so be advertised as the 'most credible.' Quasi-experiments that demonstrate some baseline equivalence between groups (so called 'wannabe RCTs') can only meet 'Group Design Standards with Reservation.' The WWC does not include studies that fall short of this threshold, nor do they consider studies using different methodologies.[8] But 'it works' is an incomplete claim. Does it mean *always* works, *generally* works, taking it to work *can be treated as the default assumption,* it works *under these specific assumptions,* it works *somewhere*? It is only the last of these for which RCTs can provide direct evidence.[9] An RCT by itself can only support claims about the population enrolled in the study and for that population it only *estimates* the intervention's *average* effect *relative to the intervention used in the comparison group.*

First, consider the problem of the average. RCTs can in the ideal give an unbiased *estimate* of the *average* effect within the population enrolled in the experiment. *Estimate*: How close the estimate is likely to be to the true average depends in part on how big the study population is but can also be

affected by asymmetries in the distribution of outcomes (Deaton and Cartwright, 2016). *Average*: What one can conclude from a higher average across a group of individuals with the intervention than without is that at least some of the individuals in the intervention group improved. We don't know which ones, though. How the average is made up matters because educators need to make decisions about distributing benefits and burdens across students. The policy could cause harmful effects as well as beneficial—it might impede performance for several individuals though on average the effect is positive. Perhaps the policy helps higher-performing students but harms students performing at a lower level, which is detrimental to the aim of bringing all students up to an adequacy threshold. Or, it might help low-performing students at the expense of higher-performing students. Risking such costs may be acceptable in education, depending on our principle of distribution. Some prioritarians about educational resources might advocate interventions that *worsen* performance for students who have surpassed the threshold if there is strong evidence that it will improve outcomes for lower-performing students. Claims about what has worked on average, don't provide enough information for educators to weigh the costs of implementing an intervention against the benefits.[10]

The second problem is 'transfer' of results. We know that student populations vary widely. In what new settings can we expect similar results to those found in the study population? The WWC responds to this problem in two ways. First, by allowing users to filter according to student demographics; second, by requiring that RCTs show positive results in a variety of settings to merit their strongest mark. Neither of these provide sufficient information to make reliable comparisons between the individuals in question and the study populations, especially without well-warranted information of what factors about student demographics matter.

The WWC search filters help users find studies conducted on populations similar in terms of gender, race, region, classroom type, school type, grade, ethnicity, and urbanicity. How reliable are those filters as guides to the factors that matter? The filters appear to be based only on reasonable assumptions about what sort of factors *might* affect the impact of an intervention.[11] The WWC does not offer evidence for the relevance of the filters themselves. In general, what evidence can be mounted for them does not fit the rigorous standards they demand of evidence that the intervention has worked somewhere. This does not mean the filters are of no use. But they certainly don't provide the sort of detailed contextual information educators need to predict whether the intervention will work in their setting.

One reason is that the categories are too coarse-grained. There are significant differences between students belonging to these categories that are likely relevant. Another is relevance. Review protocols identify which observable characteristics (e.g. minority status) served as entry conditions for the study but acknowledge that differences in unobservable characteristics (e.g. motivation) may remain and the reasons for choosing these characteristics—whether, or why, they are relevant to the outcome—are not given. Moreover, research about whether they are relevant would not meet the WWC's eligibility criteria because it would be considered secondary analysis. But that some particular set of descriptions we settle on pick out features that are relevant to effectiveness, and in which ways, is just as much an issue in need of evidence as whether the intervention is effective in some group. Knowing this sort of coarse-grained information with uncertain relevance about the study population does not get us near where we need to be to make accurate predictions about other contexts. It might even lead educators to rule out interventions that could work for them because they take location or race into account when it is irrelevant.

Consider an example. The WWC highly recommends *Success for All* (*SFA*), a whole-school initiative that includes programs for literacy, social-emotional development, computer-assisted tutoring tools, family support for parents, facilitators who work with faculty, and extensive training for teachers. The components integrate academic curriculum with school culture, family, and community inclusion. The WWC's report sums up each study that meets their standards—a combination of RCTs and quasi-experimental studies. For each, it breaks down the study and comparison samples according to minority status, grade, location, and percentage of students eligible for free or reduced-price school lunch programs. *SFA* is particularly well researched—studies have been conducted in various settings with different demographics over several years. We can surmise from the summaries that location, socio-economic status, and minority status are characteristics included in the equivalence baseline. What we don't know is whether they are relevant to the effectiveness of the program. The categories themselves are broad. There are a range of factors differentiating students with minority status that may impact results. *SFA* is a multi-faceted program, but studies reviewed by the WWC only examined the literacy component. The WWC report specifies that its ratings do not account for variations in how *SFA* was implemented. Schools can implement the program in whole or in part and the various aspects may interact in ways that affect literacy outcomes. Without knowing which other aspects of the program were employed when the interventions worked, educators cannot determine whether they were support factors for the literary component.

In addition to insufficient resources for comparing participants, there is inadequate information for comparing interventions in each group. Recommendations from the WWC and close relatives de-emphasize that results from RCTs and quasi-experimental studies are *relative* (Simpson, 2017). Educational studies never have pure 'placebos' in their control arm— i.e. interventions just like the one under test with respect to effect on the outcome but for the 'active' ingredient in the intervention. Rather, studies in education compare two different genuine interventions. It would, for instance, be unethical to withhold reading instruction entirely in a comparison school. Moreover, comparing reading levels of those who received reading instruction with those who received none would not produce useful results. Educators need to figure out whether an intervention will work better than something else they might do, not whether to teach a skill or subject. Thus, examining what was used in the comparison group is crucial for predicting whether an intervention will produce an effect in a new setting and for estimating the effect size. Unfortunately, intervention reports made for educators generally do not provide sufficient information about comparisons.

To illustrate, return to the intervention report for *SFA*, which describes the intervention and summarizes all studies that meet their standards with and without reservations, but says little about the comparison groups.[12] For some studies, it is simply noted that the comparison group received 'business-as-usual' literacy instruction (see Ross et al., 1998, 1995; Tracy et al., 2014). For another, the comparison group used 'standard' reading programs from mainstream publishers. Only some comparison classes used the same program for the duration of the three-year study—others switched between 'standard' programs which both researchers and evaluators treat as equivalent (see Quint et al., 2015). The most detailed summaries name the programs but provide little explanation of its components or how it was used. Educators trying to predict whether and to what extent *SFA* would improve reading levels in their setting should consider how their current methods compare to those used in the comparison groups. If they are using better methods, they should expect less effect than in the study—perhaps even a negative effect. Vague descriptions do not provide adequate information for this consideration—business-as-usual methods vary widely across schools and districts. In some cases, more detailed information can be found in the study itself. But the WWC and similar sources are supposed to compile and present the relevant facts to educators. By neglecting information about the comparison interventions, they imply that these are not important for deciding what to do.

Furthermore, researchers often do not control for factors present in the intervention group but missing in comparison groups. For example, when learning a new teaching method, teachers generally receive special training. When implementing *SFA,* schools provide several days of training and offer ongoing assistance. Principals and school leaders attend a week-long conference. Teachers in comparison groups rarely receive training or similar professional development opportunities. Without controlling for training and resources, we do not know the extent to which continuing professional development contributed to the positive effect. Providing the same training and resources to teachers—maybe even students—in both groups could reduce such confounding factors.

The WWC organizes interventions according to categories of educational challenges (e.g. behavior, literacy, math), offering interventions without attention to the underlying causes of the challenge. Guides for using research it disseminates do not indicate that underlying causes are relevant. As discussed above, ex ante causal analysis can aid predictions about what will work. Take another familiar example. For students performing poorly on essay writing tasks, research suggests that receiving feedback from the instructor on multiple drafts of their essays improves performance. If poor writing is, at least in part, due to teachers being overburdened by crowded classrooms, adding this intervention is unlikely to improve writing. Even if teachers' workload is not part of the cause, it is relevant to policy predictions because it provides information about the causal pathways available within that school.

Despite the lack of relevant evidence RCTs provide, blame for failed attempts often falls to those implementing the intervention. Blogs and other publications for teachers are full of anecdotal reports that recommended strategies did not work. Responses (when there are responses) are often dismissive. Respondents commonly say that strategies won't work for everyone or in every setting. It is up to *the teacher* to decide if a strategy will work in their classroom and to implement it properly. But this is not how recommended interventions are advertised. Recall the flipped classroom. The support factors we identified are critical to its success, but they are hard to see if educators rely on gold standard studies alone.

The WWC's commitment to rigorous evidence, defined in terms of RCTs, leads them to exclude many interventions altogether. Some interventions are more easily tested in RCTs than others, but this does not mean they will be more effective. Some policies may be effective over a longer period than RCTs can track or they may be too complex. The most highly recommended

interventions are software, very specific programs (like *Coping Power*), or a specific aspect of a larger program like the literacy component of *SFA*.

Broader methods or approaches to teaching often come highly recommended from peer-reviewed sources and experienced educators but will not be found on the WWC and similar databases. For example, the Stanford Center for Opportunity Policy in Education (SCOPE) studied student-centered and project-based learning strategies for teaching elements of Common Core at low-performing schools. They highly recommend both for closing the achievement gap. There are many interventions and teaching models that fit under the rubric of student-centered and project-based learning, though. RCTs would have to test a specific species of these to meet WWC's standards.[13]

Some educators think that, because disadvantaged students often face systematic barriers to performing well in school, holistic approaches will be more effective than specific programs or lesson plans. SCOPE's findings in its studies of four schools using approaches focusing on instruction and 'wraparound' services to offset stressful conditions that students experience outside of school support this view. They report that *all* students improved to some extent in the study settings (SCOPE 2014, 2015). To achieve similar results, they advise educators to develop a wide repertoire of strategies. If educators rely on recommendations from databases, they will be encouraged to adopt narrow programs or software targeting a specific outcome rather than an alternative aimed at meeting the broad needs of a particular group of students. Of course, good predictions about whether a holistic approach can be effective in a new setting require the same due diligence in collecting evidence as any other intervention. The point is that experts recommend these holistic strategies for disadvantaged students, so they are worth considering even if not supported by an RCT.

Focusing on narrowly-defined interventions using randomized trials may make for more rigorous studies, but doing so may not produce information that is more useful to educators. Even if the 'high' standards of evidence did increase the chances that results generalize, educators do not need interventions to work *generally*. They need interventions that work for their individual case. If an intervention works for one school only, it is a success for that school.

Conclusion

In the U.S., we continue to fall short of our aims for educational justice. We have argued that, despite slow and uneven progress, using EBP in education can contribute to the justice-

oriented goal of all students meeting threshold standards. To make the most of educational research, one must address the gap between the fact that an intervention has worked and the claim that it will work in a particular setting. Focusing on how educators can bridge this gap, we argue that making warranted policy predictions requires a good argument composed of relevant, well-supported premises. The fact that an intervention has worked somewhere can be a useful premise, but it is far from enough to conclude that it will work in a new context. Unfortunately, such inferences are common among those using EBP in education. We have identified other kinds of relevant premises educators can use to construct sound policy arguments. There is no straightforward formula or guide for gathering the right evidence, but we hope that recognizing the kinds of facts that impact effectiveness can help improve policy deliberations.

Although educators within local contexts are in the best epistemic position to secure evidence for some of these premises, researchers can help. They can investigate *how* interventions worked in the study setting and report on causal components and their support factors. Similarly, researchers can consider which aspects of the arrangements in study settings and features of individuals affect the outcome. Also, they can identify intermediate steps observed during the study that indicate success. Learning more about causal mechanisms of interventions and the conditions enabling their operation will put educators in a better position to make reliable policy predictions. There are steps in this direction, for instance *realist evaluation* that studies the circumstances under which an intervention worked and the study population for whom it worked in addition to the causal efficacy of the intervention itself, which proposes context-mechanism-outcome models. Pursuing this methodology could supply much more of what educators need to make good policy predictions (see Pawson and Tilly, 1997).

Implementation science is another recent trend aiming to bridge the gap between research and practice. It does not focus on causal mechanisms but rather examines methods for transferring and applying 'effective' policies to real-world contexts (Kelly, 2012). We endorse efforts to understand what makes environments hospitable to certain policies and creating assessment tools for policy-makers, but we want to register a caution. Thinking about the success of evidence-based interventions in new settings as a matter of good implementation risks ignoring the prediction phase. Implementation science emphasizes generic measures like appointing implementation teams that can be held accountable and finding better ways of ensuring high fidelity (Carroll, et al. 2007; Dusenbury, et al. 2003; Hasson, 2010). Our arguments indicate that fidelity is not always the best

strategy and we reject the assumption that an evidence-based intervention can work almost anywhere if properly implemented.

Importantly, understanding the causal mechanism underlying interventions is better than ensuring high fidelity because it allows teachers to innovate and adapt interventions. Experienced teachers recognize their strengths and have a sense of what will work for their students. It is generally good for teachers to depart from strict fidelity to adapt interventions accordingly. If they understand the causal mechanism, they can adapt in ways that do not disrupt or dilute the efficacy of interventions they employ. For example, say quizzing improves test scores because frequently calling recently learned facts to mind helps students remember them. Teachers may be able to design quizzes in the form of games or other activities more appealing to their students and allow them to play to their strengths. In one school, a competition for the best quiz scores might motivate students. In another, quizzing might work best if implemented as an interactive game with only a small emphasis on individual performance.

Researchers can also help by developing new concepts and theories. Indeed, educational researchers have developed some concepts important for predicting whether interventions will be effective and achieving greater educational justice. *Learning readiness*—discussed above—is a prime example. Since the concept emerged, researchers have distinguished between different aspects of learning (un)readiness and continue to identify contributing factors, like family activities, values, and parenting styles. Recognizing how learning readiness impedes or promotes learning opens important avenues for further investigation, especially because greater justice in education, as currently conceived in the US, requires helping disadvantaged students reach a threshold of adequate outcomes. Considering this goal, research could focus more on studying *particular* populations rather than using randomization to produce anonymous evidence. Funding could support research that aims to better understand obstructions disadvantaged students encounter and devise strategies to ameliorate them.

Funding Acknowledgements: Funding for this project was provided to Nancy Cartwright and Kathryn Joyce by the Center for Ethics and Education. For Cartwright, this material is based upon research supported by the National Science Foundation under grant no: 1632471 and the European Research Council under the European Union's Horizon 2020 research and innovation program (grant agreement no. 667526 K4U). It is acknowledged that the content of this work reflects only the authors' views and that the ERC is not responsible for any use that may be made of the information it contains.

Personal Acknowledgments: We thank participants who attended a workshop for an early draft of this paper hosted by the Center for Humanities Engaging the Social Sciences at Durham University, especially Erin Nash, Julian Reiss, Katherine Fuhrman, and Brian Earp. We are grateful to Adrian Simpson, Gina Schouten, and Harry Brighouse for extensive feedback and discussion.

[1] NCLB renews the Elementary and Secondary Education Act (1965). Title I, the central provision of NCLB and its predecessors is called 'Improving the Academic Achievement of the Disadvantaged'. In 2009, the Every Student Succeeds Act replaced NCLB, preserving its commitment to EBP.

[2] Even testing an intervention in the same class at different times might not provide an identical case but for the intervention. Students may be affected by a variety of factors from day to day that interact with the effect tested for.

[3] This list is meant as an example. Success with quizzing may require additional or different support factors in different specific settings.

[4] This is a simplistic model. Models can provide more complex causal maps that include other causal cakes that are expected to be in place. We can construct maps that trace the causal process, identifying each step and creating causal cakes for each causal factor. Or, we can construct and compare many causal cakes that are expected to positively or negatively affect an outcome. For more complex modelling see Munro et al. (2016) and Layne et al. (2014).

[5] We can describe causal cakes in terms of INUS conditions—Insufficient but Necessary parts of an Unnecessary but Sufficient condition for producing some outcome. Each ingredient is insufficient but necessary for the cake to produce an effect and the cake itself is unnecessary but sufficient for producing the outcome. See J.L. Mackie (1965) and Cartwright and Hardie (2012).

[6] Note that, while studies may show strong evidence of a positive effect, the strongly evidenced effect could be small. Often, 'strong evidence of a positive effect' is conflated with 'evidence of a *strong positive* effect.' A 'strong positive effect' indicates that an intervention can produce significant improvements. Conflating these two leads to confusion about effect sizes and the significance of findings for educators. See Simpson (2017).

[7] The Coalition for Evidence-Based Policy has been integrated into the Laura and John Arnold Foundation as the Evidence-Based Policy and Innovation Initiative. Johns Hopkins University website: Best Evidence Encyclopedia and Evidence for ESSA (Every Student Succeeds Act) is a prominent example. In addition to these, websites like Education World offer lesson plans and strategies designed for 'Connecting Educators to What Works.'

[8] For some programs, the WWC includes within their detailed intervention reports a list of studies that didn't meet their standards, indicating reasons for exclusion. So, educators using WWC have some access to other studies but only if those correspond to approved research that warrants an intervention report.

[9] Except in the rare case where the population in the trial can be taken to be a representative sample of the target population.

[10] When conducting RCTs, researchers collect microdata which is usually unavailable to others. If it were available, educators could see how each individual student performed on a task prior to the intervention and after. They could also see how the low performing students improved relative to higher performing students. Still, this microdata is insufficient. For any individual who improved, the study cannot guarantee that the intervention caused her improvement. Other factors might be responsible, like a change in family dynamics. Gathering such information is not part of the RCT

design. Even if it were collected, confidentiality rules associated with RCTs would likely disallow sharing individual information.

[11] When applying filters to searches, the WWC offers a 'hint' that says: 'Student, school, and setting characteristics can affect the effectiveness of an intervention.' This is all that it offers regarding the relevance of the search filters.

[12] The WWC's summary of Madden et al. (1993) provides more detail about the comparison group, but it is still insufficient. Some schools in the study implemented *SFA* while comparison schools used the *Macmillan Connections* basal series and tried other broadly specified strategies including reducing class sizes and offering pull-out services for low-performing students. Some researchers focus on comparing methods. Their results may be more useful for educators. For example, see Skindrud et al. (2006).

[13] Meta-analyses are sometimes used to report on general strategies or approaches by combining effect sizes found in individual studies (e.g. one study uses strategy X in social studies, another uses X in math, and another uses X in Language Arts) to show the overall impact of the approach. There are several challenges associated with combining effect sizes that may undermine the validity of conclusions. For a discussion of problems with many meta-analyses, see Simpson (2017) and Stegenga (2011).

References

Biesta G (2007) Why 'What Works' Won't Work: Evidence-Based Practice and the Democratic Deficit in Educational Research. *Educational Theory* 57(1):1-22.

Booth A and Crouter AC (2008) *Disparities in School Readiness.* Taylor & Francis.

Bridges D, Smeyers P, Smith R (2008) *Special Edition on Evidence* in *Journal of Philosophy of Education* 42 (1).

Brighouse H, Ladd H, Loeb S, Swift A (2016) Educational Goods and Values: A Framework for Decision Makers. *Theory and Research in Education* 14(1): 3-21.

Carroll C, Patterson M, Wood S, Booth A, Rick J, Balain S (2007) A Conceptual Framework for Implementation Fidelity. *Implementation Science* 2:40.

Cartwright N (2015) Single Case Causes: What is Evidence and Why. CHESS Working Paper 2015-02.

Cartwright N (2013) Evidence, Argument, and Prediction. In: Karakostas V, Dieks D (eds.) EPSA11 *Perspectives and Foundational Problems in Philosophy of Science.* Springer International Publications, pp. 3-17. The European Philosophy of Science Association Proceedings.

Cartwright N and Cowen N (2015) Making the Most of Evidence: Evidence-Based Policy in the Classroom. CHESS Working Paper 2015-03.

Cartwright N and Cowen N (2014) Making the Most of Evidence in Education: A Guide for Working Out What Works…Here and Now. CHESS Working Paper 2014-03.

Cartwright N and Hardie J (2012) *Evidence-Based Policy: A Practical Guide to Doing it Better.* Oxford: Oxford University Press.

Cartwright N and Stegenga J (2011) A Theory of Evidence for Evidence-Based Policy. In: Dawid P, Twining W, Vasilaki D (eds) *Evidence, Inference, and Enquiry.* Oxford: Oxford University Press, pp. 291-322. Proceedings of the British Academy (71).

Datnow A (1998) *The Gender Politics of Educational Change.* Bristol, PA: Falmer Press.

Datnow A and Park V (2015) Data Use for Equity. *Educational Leadership* 75(5): 48-54.

Datnow A and Castellano M (2000) Teachers' Responses to Success for All: How Beliefs, Experiences, and Adaptations Shape Implementation. *American Educational Research Journal* 37 (3): 775-799.

Davidson KL and Frohbieter G (2011) District Adoption and Implementation of Interim and Benchmark Assessments. Report, Los Angeles, CA: National Center for Research on Evaluation, Standards, and Student Testing.

Deaton A and Cartwright N (2016) Understanding and Misunderstanding Randomized Controlled Trials. The National Bureau of Economic Research Working Paper 22595.

Dusenbury L, Brannigan R, Falco M, Hansen, WB (2003) A Review of Research on Fidelity of Implementation: Implications for Drug Abuse Prevention in School Settings. *Health Education Research* 18 (2): 237-256.

Farkas G and Hibel J (2008) Being Unready for School: Factors Affecting Risk and Resilience. In: Booth A, Crouter AC (eds) *Disparities in School Readiness.* New York, NY: Taylor & Francis.

Friedman-Sokuler N (Manuscript) Empirical Economics, the Gold Standard and Public Policy: The Case of Class Size Reduction.

García E and Weiss E (2017) Education inequalities at the school starting gate: gaps, trends, and strategies to address them. Report, Economic Policy Institute. Available at epi.org/132500.

Hasson H (2010) Systematic Evaluation of Implementation Fidelity of Complex Interventions in Health and Social Care. *Implementation Science* 5:67.

Kelly B (2012) Implementation Science for Psychology in Education. In: Kelly B and Perkins DF (eds) *Handbook of Implementation Science for Psychology in Education*. Cambridge University Press.

Kvernbekk T (2011) The Concept of Evidence in Evidence-Based Practice. *Educational Theory* 61(5):515-532.

Kvernbekk T (2016) *Evidence-Based Practice in Education: Functions of Evidence and Causal Presuppositions*. New York, NY: Routledge.

Layne C, Steinber J, Steinber A (2014) Causal Reasoning Skills Training for Mental Health Practitioners: Promoting Sound Clinical Judgment in Evidence-Based Practice. *Training and Education in Professional Psychology* 8(4): 292-302.

Mackie JL (1965) Causes and Conditions. *American Philosophical Quarterly* 2:245-64.

Madden NA, Slavin R, Karweit N, Dolan L, Wasik BA (1993) Success for All: Longitudinal effects of a restructuring program for inner-city elementary schools. *American Educational Research Journal* 30(1): 123–148.

Menzies P (2014) Counterfactual Theories of Causation. *Stanford Encyclopedia of Philosophy*.

Mosteller F and Boruch R (2002) *Evidence Matters: Randomized Trials in Education Research*. Washington, D.C.: Brookings Institution.

Munro E, Cartwright N, Hardie J, Montuschi E (2016) *Improving Child Safety: Deliberation, Judgement, and Empirical Research*. Center for Humanities Engaging Social Sciences, Durham University.

Park V, Daly AJ, Guerra AW (2013) Strategic Framing: How Leaders Craft the Meaning of Data Use for Equity and Learning. *Educational Policy* 27(4): 645-675.

Pawson R, and Tilley N (1997) *Realistic Evaluation*. Sage Publications.

Phillips DC (2009) Empirical Educational Research: Charting Philosophical Disagreements in an Undisciplined Field. In: Siegel H (ed) *Oxford Handbook of Philosophy of Education*. Oxford: Oxford University Press.

Phillips DC (2007) Adding Complexity: Philosophical Perspectives on the Relationship Between Evidence and Policy. In: Moss P (ed) *Evidence and Decision Making*. Blackwell Press, 376-402.

Quint JC, Zhu P, Balu R, Rappaport S, DeLaurentis M (2015) *Scaling up the Success for All model of school reform: Final report from the Investing in Innovation (i3) Scale-Up*. New York, NY: MDRC

Reiss J (2014) What's Wrong with Our Theories of Evidence? *Theoria* 29(2): 283-306.

Rothstein R (2009) Equalizing Opportunity. *American Educator* 33(2):4-46.

Shavelson RJ and Towne L (eds) (2002) *Scientific research in education*. Committee on Scientific Principles for Education Research. Division on Behavioral and Social Sciences and Education. Washington, DC: National Academy Press.

Simpson A (2017) The misdirection of public policy: comparing and combining standardized effect sizes. *Journal of Education Policy* 32(4): 450-466.

Skindrud K and Gersten R (2006) An evaluation of two contrasting approaches for improving reading achievement in a large urban district. *Elementary School Journal* 106(5): 389–407.

Slavin R (2002) Evidence-based education policies: Transforming educational practice and research. *Educational Researcher* 31(7): 15–21.

Slavin R (2004) Education Research Can and Must Address 'What Works' Questions. *Educational Researcher* 33(1): 27–28.

Smeyers P and Dapaepe M (2006) *Educational Research: Why What Works Doesn't Work*. Netherlands: Springer Publishing

Spillane JP and Miele DB (2007) Evidence in Practice: A Framing of the Terrain. *Yearbook of the National Society for the Study of Education* 106(1): 46-73.

Stanford Center for Opportunity in Education (2014) Student-Centered Schools: Closing the Opportunity Gap. Report, Stanford University. Available at https://edpolicy.stanford.edu.

Stanford Center for Opportunity in Education and Nellie May Foundation (2015) Centered on Results: Assessing the Impact of Student-Centered Learning. Stanford University. Report, Stanford University. Available at: https://edpolicy.stanford.edu.

Stegenga J (2011) Is Meta-Analysis the Platinum Standard of Evidence? *Studies in History and Philosophy of Science Part C* 42(4): 497-507.

Timar T (2012) Reframing Policy and Practice to Close the Achievement Gap. In: Timar T and Maxwell-Jolly J *Narrowing the Achievement Gap.* Cambridge, MA: Harvard University Press.

Tracey L, Chambers B, Slavin RE, Madden NA, Cheung A, Hanley P (2014) *Success for All in England: Results from the third year of a national evaluation.* SAGE Open 4(3): 1–10.

U.S. Department of Education (2003) *Identifying and Implementing Educational Practices Supported by Rigorous Evidence: A User Friendly Guide*. Available at: https://eric.ed.gov

U.S. Department of Education Institute of Education Sciences, What Works Clearinghouse (2016) *What Works Clearinghouse Procedures and Standards Handbook 3.0* Available at: https://ies.ed.gov.

U.S. Department of Education Institute of Education Sciences, What Works Clearinghouse (2017) *Beginning Reading intervention report: Success for All*®. Available at: https://whatworks.ed.gov.