Chapter 10
Guiding Principles for Evaluating Evidence in Education Research

Sarah Kay McDonald and Barbara Schneider

Mistaking no answers in practice for no answers in principle is a great source of moral confusion – Sam Harris

Abstract Based on their experiences from their work with two national initiatives designed to reform educational practice in U.S., the authors present seven guiding principles of evidence-based/informed educational policy and research to lay the foundation for making rigorous and comprehensive judgments about what evidence and scientific research designs should be taken into account when scaling-up educational reforms to serve the public good. The authors further provide case examples from US with a clear potential to both utilize and generate evidence in the public interest including educational research studies that seeks to support underrepresented groups in preparing for and achieving successful transitions to postsecondary education and careers, in STEM and other fields. The authors conclude that educational researchers have a critical role to play in providing decision-makers with the tools to judge the evidence to serve public good.

The improvement of the education system has been a constant concern to educators and policymakers both within the U.S. and abroad and it has assumed a position of national and international significance unparalleled in previous decades. Never before have we seen so much attention by governments, philanthropic

S.K. McDonald
National Science Foundation, Arlington, VA, USA

B. Schneider (✉)
College of Education, Michigan State University, East Lansing, MI, USA
e-mail: bschneid@msu.edu

M.Y. Eryaman, B. Schneider (eds.), *Evidence and Public Good in Educational Policy, Research and Practice*, Educational Governance Research 6,
DOI 10.1007/978-3-319-58850-6_10

organizations, and social media directed at the transformation of school organizations, teacher evaluation systems, instruction, and assessments. In the U.S. alone, in 2010, President Obama awarded over $4.5 billion dollars for education reform through the American Recovery and Reinvestment Act. That same year the Bill & Melinda Gates Foundation awarded an additional half a billion dollars to early learning and college-ready education initiatives.[1]

Why is education drawing such attention and resources? Two major problems continue to plague many world-wide educational systems. First, is the continuing achievement gap between more socially advantaged students and those with fewer social and economic resources in elementary, secondary school, and higher education (Duncan and Murnane 2011; Chmielewski 2014). In some countries, these achievement gaps are also confounded by race and ethnicity and immigration status (OECD 2015). For several decades in the U.S. the average performance of white students has surpassed that of blacks and Hispanics.[2] Recent projections indicate that these trends are likely to persist at least in the near future (Reardon 2011).

Second, is differential access to quality schools, postsecondary education, and job training. In the U.S. the number of minorities in low-paying, non-skilled jobs continues to be disproportionately higher than that of whites (U.S. Department of Labor 2011). These trends reflect, in part, the lower numbers of minorities completing postsecondary degrees compared to whites (National Science Foundation 2010). Similar to the U.S., many countries throughout the globe have also been challenged with improving secondary school completion rates and access to higher education and training among all students regardless of their family characteristics. Problems of inequity of educational access and opportunity are also predicted to escalate with the increases in immigrants seeking refugee from political unrest in the Middle East and several African nations (OECD 2015).

Educational developers and researchers have responded to these problems by designing interventions that create new pedagogical tools, instructional content, and assessments to narrow the achievement gap. One area of particular emphasis has been teacher quality including reforms such as alternative routes to teacher certification, merit-pay, and evaluation practices. Other types of reforms for enhancing access include changes in school structure and programs that offer a more successful transition into postsecondary education and the labor market, including national initiatives such as the Knowledge is Power Program (KIPP) and local initiatives such as the Chicago-based Urban Prep Academies.[3] Considerable investments have also been

[1] American Recovery and Reinvestment Act. (Pub.L.11-5); Gates Foundation: http://www.gatesfoundation.org/united-states/Pages/measures-of-effective-teaching-fact-sheet.aspx

[2] Results of the 2009 NAEP for U.S. high school seniors found no significant changes in the gap between white and black students' reading scores from 1992 to 2009, and no significant change between white and black or Hispanic students' mathematics scores from 2005 to 2009 (NCES 2011).

[3] KIPP (http://www.kipp.org/) is "based around high expectations for student achievement; commitment to a college preparatory education by students, parents, and faculty; devotion of time to both educational and extracurricular activities; increased leadership power of school principals; and a focus on results through regular student assessments" (U.S. Department of Education, Institute of Education Sciences, What Works Clearinghouse 2010). Urban Prep is a Chicago-based

made in leveraging the power of technology to support student learning (e.g., through data visualization tools, online learning communities, intelligent tutoring systems, and computer games and virtual environments) and access to postsecondary education.[4] Despite the large number of initiatives being piloted, some have proved disappointing when adopted at scale, while others have had a more successful trajectory.

One major innovation that has been successfully scaled is Success for All (SFA), a comprehensive whole-school reform approach to improvement that incorporates research-based curriculum materials, professional development, assessment and data-monitoring tools, and activities that facilitate family involvement and community support. First implemented in a single school in Baltimore, Maryland, 25 years later the Success for All Foundation serves over 2000 schools in 46 U.S. states and offers assistance to projects in five other countries.[5] In 2010, the Foundation was the recipient of a $50 million grant from the U.S. Department of Education's Investing in Innovation program to scale-up the program to reach over half a million additional elementary school students. Key to the success of SFA has been the robust evidence of its positive impact on student learning. Multiple evaluations have been conducted on SFA including an independent study that showed it met the criteria for the strongest evidence of effectiveness, indicating significant positive effects and replication in multiple contexts including schools likely to adopt and implement SFA (Borman et al. 2003, 2007). Other more recent independent positive evaluations of SFA include an assessment of major comprehensive education reforms by Rowan et al. (2009) and another by MDRC funded by the U.S. federal government showing that SFA was especially effective in schools with students having low preliteracy skills (Quint et al. 2015).

While not without its critics, the SFA program is notable both for its acknowledged impacts and for its commitment to amassing a rich and deep research base that has informed its development and implementation. Few interventions have such a track record of evidence warranting scale-up. Rather, the educational research landscape remains heavily populated by small studies with disparate findings and less rigorous evaluations. This uneven evidential base of research might explain why educational studies have had such a limited role in formulating public policy. Scholars have argued that strong evidence on its own is rarely sufficient to explain how public policy agendas are shaped and enacted (Weiss 1989; Stevenson 2000). Their position has been that research, whether in the U.S or in other countries, rarely provides definitive answers or prescribes specific policies (see, e.g., Weiss 1982; U.K., House of Commons

initiative operating in the only all-male public schools in the state of Illinois to "provide a comprehensive, high-quality college preparatory education that results in graduates succeeding in college" (see http://www.urbanprep.org/about/historvlindex.asp).

[4] See, Dynarski and Scott-Clayton (2007) and Hoxby (2007). Other examples of online resources on the college selection and application processes in the U.S. include the National Center for Education Statistics College Navigator (http://nces.ed.gov/collegenavigator) and the American Council on Education, Lumina Foundation for Education, and Ad Council's KnowHow2GO (http://www.knowhow2go.org/).

[5] See the Success for All Foundation's 'Our Story', retrieved February 22, 2011 from http://www.successforall.org/About/story.html

2006). Instead, research often plays a 'framing' function, shaping discourse, conceptualizations, and the ways problems and potential solutions are formulated.

Times have changed, however, and whereas policy makers may once have discounted educational research, that does not seem to be the case today. Policymakers now value reforms like SFA that produce statistically sound results that can be used to inform educational decisions. In the U.S. this press for evidence accountability encompasses the entire educational system from the federal government to local school districts. The most obvious example of this was the enactment of the No Child Left Behind Act (NCLB) (Public Law 107–110), with its reliance on data to sanction schools based on their lack of academic performance. State and local school districts were mandated to collect, validate, and transmit massive amounts of student, school, and teacher performance data on the effectiveness of their educational systems.

NCLB had a rocky road of implementation, caught in a net of local and state dissatisfaction and bipartisan political conflict all of which delayed reauthorization of the next bill for over a decade. Finally, in 2015, a new federal education bill the, Every Student Succeeds Act (Pub. L. 114–95), was ratified. While permitting states more flexibility in determining standards for measuring school and student performance, the general public and its legislatures, continued to press for testing, reporting, and accountability on the progress of all students and their schools. This emphasis on testing and accountability, although somewhat more relaxed than the previous legislation, corresponds to a more world-wide movement to measure the status and improvement of student learning and teacher and school effectiveness.

This trend toward amassing data for purposes of decision making has been augmented by a number of activities, one of which is the development of research organizations and associations designed to highlight experimental and quasi-experimental studies and methods. Some of these organizations include the Society for Research on Educational Effectiveness (SREE, https://www.sree.org), the What Works Clearinghouse in the U.S., and the Campbell Collaboration (which includes health, social sciences and education), all of which compile lists of robust studies that rely on evidence for decision-making.[6] Older, more established education associations both in the U.S. and around the world are also revamping and professionalizing their organizations to reflect these new demands for rigorous education research. Organizations such as the American Educational Research Association (AERA, https://www.aera.org) have and continue to be committed to these goals and exercise leadership in these areas, including assisting in the formation of the World Education Research Association (WERA, https://www.wera.org), an international society with a similar purpose.

[6] The What Works Clearinghouse is an initiative of the U.S. Department of Education's Institute of Education Sciences which 'develops and implements standards for reviewing and synthesizing education research' (http://ies.ed.gov/ncee/wwc/aboutus). The Campbell Collaboration is an 'international research network that produces systematic reviews of the effects of social interventions' (http://www.campbellcollaboration.org/aboutus/index.php). The Society for Research on Educational Effectiveness seeks to advance and disseminate research on the causal effects of education interventions, programs, and policy (http://www.sree.org/pages/mission.php).

Even though there has been a general sentiment for more rigorous research within the education community, there has been considerable attention regarding the methodology and criteria for determining what works and what does not (National Research Council 2002; Walters et al. 2008), with some critics arguing against standards for evaluating educational programs and practices. Policymakers have strongly pressed for only making investments in education reforms, particularly those with public resources, on robust evidence. However, the field's ability to produce such an evidence base seems incompatible with many reform timelines. One exception to speed the process of evidence-informed reform is being tested at The Carnegie Foundation for Teaching and Learning.

Spearheaded by its President, Anthony Bryk, the Foundation is working on implementing reforms using the modified 90-day cycle for researching and assessing innovative ideas employed by the Institute for Healthcare Improvement (see Bryk 2015). Bryk began by using this model to explore whether math-intensive programs can move students in community colleges out of developmental math courses (Yamada and Bryk 2016) and has now applied the model to other reforms that can be quickly implemented in educational systems. The intent of Bryk's plan is to re-engineer educational research to one that promotes an improvement science that addresses the complexity and variability in school performance within a shorter more productive time frame (Bryk 2015).

One of the most beneficial outcomes of efforts to truncate the research and development cycle may be embracing more realistic expectations regarding the roles educational research can and should play for informing reform. This chapter is designed to define some of the principles for making sound judgments about research quality and what evidence should be taken into account in making decisions regarding educational practices and policies, especially for those interventions designed for scale-up. At issue is not just the strength of evidence that can be attributed to specific interventions (determining what works), but establishing the contexts (e.g., classroom, school, neighborhood) and populations (e.g., demographic characteristics) for which it is likely to work equally well (e.g., generalizability of effects). The principles here reflect current work being conducted by social scientists working in diverse national and international settings and our work with two U.S. national initiatives designed to articulate what considerations need to be taken into account when bringing promising interventions to scale (Schneider and McDonald 2007; Milesi et al. 2014). Principles are merely touchstones; even if scientifically grounded, their use is subject to the will of decision makers. Our intent is simply to lay the foundation for making sound judgments about the nature of evidence that should be taken into account when scaling-up educational reforms.

Principle 1: Gauging the Impact on Learning

One of the first issues to consider in weighing the value of evidence is its potential impact on advancing knowledge of learning and instruction. Whether studying pedagogy, redesigns of school organizations, or new technologies, the fundamental

issue is if the intervention impacts learning outcomes. It is important to consider the theory upon which the intervention is based, how it has been tested over time, and how it affects different populations in diverse settings. One example that meets these criteria is the Carnegie Learning Cognitive Tutor®, developed by John R. Anderson and colleagues.

For decades, psychological experiments have generated data about humans' attention to and perceptions of their external environment, including reasoning, memory, problem solving, and decision-making. Anderson integrated these ideas into a single unified theory of cognition which models how humans perceive, organize, think about, and act upon knowledge.[7] This blueprint of human information processing suggested opportunities to stimulate learning through intelligent computer-based tutoring systems. Critical to the model is the notion that knowledge is strengthened with use. This is the theory upon which he developed a tutoring system that focuses on active engagement with and use of knowledge (see Ritter et al. 2007a). Initial field tests suggested that the tutors were more successful with some teachers than others, a finding that led the investigators to focus more closely on the enacted curriculum (i.e., what was actually occurring in classrooms). Consequently, the team expanded on its work to develop a curriculum that could be embedded within the tutor.

Over time, Carnegie Learning's Cognitive Tutors have been tested in studies using some of the most rigorous designs supporting causal inference, with numerous student and teacher populations and outcome measures. The methodological approach here is a randomized control trial in which the treatment condition is measured against a control condition taking into account potential assignment counterfactual conditions (Holland 1986; Imbens and Rubin 2010; Rubin 2005). Positive impacts of the tutor on students' mathematics learning and achievement have been found in numerous middle-school, secondary school, and higher education settings in California, Colorado, Florida, Oklahoma, Ohio, Pennsylvania, Texas, Washington, and Wisconsin. Controlled comparison field trials (utilizing matched control groups and quasi-experimental designs) and other robust statistical analyses demonstrate significant improvements in student learning attributable to the Cognitive Tutor of student learning (e.g. SAT, Iowa Algebra Aptitude Test, and problem situations and multiple representations tests). On the positive side, an independent evaluation that met the What Works Clearinghouse evidence standards found significant increases in first semester grades and other learning measures including scores on the ETS Algebra I end-of-course exam.[8] But even with these successful evaluations, a U.S. Department of Education study found no significant differences between the Cognitive Tutor versus a control condition (see Campuzano et al. 2009). Should we discount this evidence or recognize that there will be instances where results will not replicate?

[7] Anderson's original Adaptive Control of Thought (ACT) theory of human cognition was first described in Anderson, 1976; elaborated in 1983; and refined into the ACT-R (Adaptive Control of Thought-Rational) theory for understanding and stimulating cognition, 1993, which is the foundation of the Cognitive Tutor software.

[8] For additional information see Ritter et al. (2007a, b). For a review of this study, see the WWC July 2009 Intervention Report on the Cognitive Tutor® Algebra I available online at http://ies.ed.gov/ncee/wwc/pdf/wwccogtutor072809.pdf

Reproducibility of studies, especially ones like this with multiple conditions and unusual contextual factors, including implementation procedures are part of conducting work in classrooms *not laboratories*. There are no silver bullets for improving *all students'* mathematical learning at this time. Nevertheless, we should continue to investigate different designs especially those that take advantage of emerging technologies. The important message here is the value of solid theoretically driven interventions that allow for strategic iterative evaluations which identify factors that influence their success and the contextual conditions that undermine their effectiveness.

Principle 2: Knowing What to Measure

Having established a study's potential to improve our knowledge base regarding learning, it is important to consider how the outcomes of interest should be measured. At issue is whether the metrics proposed are calibrated to detect meaningful change. From the investigator's perspective, key considerations include: how well the metrics capture constructs of interest; whether the process of assigning values to measure change is sufficiently transparent to enable replication; and whether the costs of developing, collecting, coding, and analyzing proposed metrics will yield information of commensurate value. From the perspective of the decision maker, the key criterion is whether what is being measured is the relevant outcome for observing, assessing, and enabling a policy change.

An example of educational research that underscores the importance of employing assessments to detect specific changes in learning is the BioKIDS: Kids' Inquiry of Diverse Species intervention developed by Nancy Songer and colleagues. Like the Cognitive Tutor, BioKIDS integrated new curricular units with innovative technologies (in this case, handheld devices for students' use). Focusing on elementary and middle school students in high-poverty urban classrooms, BioKIDS fostered the development of inquiry thinking skills while providing instruction in life science content. Using their schoolyard environments, students explored biodiversity, tracking animals and logging data on personal digital assistants (PDAs). The students' observational data were explored through a carefully scaffolded series of activities designed to foster inquiry-based science learning.[9]

The Songer team recognized the inadequacy of standard science assessments to detect the outcomes targeted by the BioKIDS intervention. Evaluating students' ability to engage in complex reasoning about scientific ideas required alternative forms of assessment. Developing an assessment that identified and calibrated students' reasoning capacity became central to measuring the impacts of the intervention. The BioKIDS team partnered with researchers on the Principled Assessment Design for Inquiry (PADI) project to develop high quality assessments of science

[9] For additional information on BioKIDS see the project's web site at http://www.biokids.umich.edu/

inquiry aligned with the goals of the intervention and informed by emergent thinking regarding the science and design of assessment.[10]

With the new metric, Songer's team disentangled "students' content knowledge from their complex reasoning abilities," vital for developing students' capacity not only to master content knowledge but also to interpret data and formulate scientific explanations. More generally, empirical evaluations of the BioKIDS intervention and its assessment system enhanced the development of both curricular units and the assessments, while demonstrating statistically significant and substantively meaningful improvements in student achievement (see e.g., Songer et al. 2009, 2007; BioKIDS, University of Michigan 2005). Impressive as student standardized achievement tests were, Songer singled-out the insensitivity of standardized tests to evaluate complex thinking about science' as "perhaps the most important aspect of this work" (Songer et al. 2009: 628).

Importantly, the challenges of assessing rich and multi-faceted effects of interventions that seek to improve content knowledge and deeper thinking skills are not unique to BioKIDS. Standardized tests are often poorly aligned with innovative curricula and are insensitive to changes new interventions seek to foster (see e.g., Pellegrino et al. 2001, 2014) For this reason, it is unwise to dismiss interventions incapable of producing higher scores on existing metrics; instead, it is important to ask whether existing metrics are misaligned with the interventions designed to attain them. Critical questioning of metrics is a natural component of any improvement process. Defaulting to traditional measures is unlikely to prove helpful in advancing new knowledge and skill sets. Weighing evidence, then, it is always important to ask "are we measuring what we ought to measure?", and to consider when it may be necessary to augment the assessment repertoire with new metrics for gauging impacts on learning.

Principle 3: Employing Standards of Scientific Design

There are many types of study designs, all of which have important roles to play in understanding educational phenomena. In deciding among them a key consideration is how confident the investigator needs to be in examining the nature of relationships she posits or observes among educational outcomes and other variables of interest. Important differences in individual research objectives notwithstanding, any study which aims to generate evidence to inform educational policy or practice fundamentally strives to illuminate potentially causal connections. How secure we need to be in our assessments of these connections varies at different stages in the

[10]The Principled Assessment Designs for Inquiry (PADI) project builds on developments in measurement theory, technology, cognitive psychology, and science inquiry, implementing the evidence-centered assessment design (ECD) framework (see http://padi.sri.com). For additional information on the BioKIDS/PADI collaboration and details of the assessment system, see Songer et al. (2009), and Gotwals and Songer (2006).

research and development cycle. The first stage of the research cycle is to provide proof of concept for innovations. Initial proof of concept tests may tolerate some ambiguity, but by the time we move to the next stage of the experimental cycle (establishing efficacy trials), gaps in logic models cannot be overlooked. By the time one is testing a fully scaled intervention with an effectiveness trial, the design should provide solid evidence of cause and effect.

Scientific design standards are invaluable for constructing investigations that yield evidence for eventually meeting requirements for scale-up. Properly applied, they increase the likelihood that robust and credible evidence rather than compelling stories will provide the foundations for policy initiatives. Likelihood is not, however, certainty; even the best designs may yield evidence of questionable value – for example, when plagued by circumstances (such as attrition) beyond the investigator's control, or when concerns with establishing the cause of an effect overwhelm attention to moderators which may condition and constrain impact.

An example of a program of educational research that over a decade employed a wide range of robust designs to establish causal connections was conducted by Barbara Foorman and colleagues. Working in Texas and Florida, Foorman developed, piloted, refined, tested, and scaled two evidence-based reading interventions. The first intervention was designed for teachers to establish appropriate learning objectives for each student and provide individualized instruction enabling students to read at or above grade level. Targeting children in the primary grades, they developed the Texas Primary Reading Inventory (TPRl) to align with new state standards and research evidence on the development of reading skills. The second intervention was the Florida Assessments for Instruction in Reading (FAIR) to assist teachers in their instructional decision-making. Both TPRl and FAIR use diagnostic, classroom-based assessments to identify those students at risk of developing reading problems with more intensive, targeted diagnostic inventories.

Each of these interventions uses technology (e.g., in the case of TPRI, internet and handheld devices; in the case of FAIR, computer adaptive testing) that provides ancillary supports to assist teachers in adapting and targeting instruction that focuses on skills the students have not yet mastered. Both of these interventions have been tested with rigorous validity and reliability evaluations of the assessment instruments and their impact for supporting assessment-driven instruction. On the basis of this evidence each has been scaled for use with students and teachers across the state. In Texas, TPRI is used with students in Kindergarten through the third grade; in Florida, FAIR is used at no charge in public schools with students in grades K-12.[11]

While both the TPRI and FAIR evolved through a careful progression from development to evaluations establishing effectiveness and achieved widespread adoption, each was further developed with ongoing testing of the assessments and the targeted

[11] For additional information on the TPRI see Foorman et al. (1998) and Foorman et al. (2007); and the web site at http://www.childrensleaminginstitute.org/ourprograms/program-overview/TPRI/. For information on FAIR see Foorman and Petscher (2010) and Foorman et al. (2009); and the web site at http://www.fcrr.org/fair/index.shtm

instruction they facilitate. A 2008–2009 development study was designed to assess and improve the validity and reliability of the entire TPRI (CLI/TIMES 2014: 4) based on material tested with approximately 3000 students. Similarly, investigators at the Florida Center for Reading Research continued to leverage data from FAIR to explore and develop activities that enhanced reading skills (see Foorman and Petscher 2010), and conduct research on the development and evidence from the assessment system, including causal effects of individualized instruction.[12]

The TPRI and FAIR initiatives highlight the iterative refinement of effective interventions, the partnerships required to enact robust designs in the classroom, and the importance of continued R&D commitments long after efficacy and effectiveness is established. Exemplary interventions moved to scale should not be regarded as sacrosanct but instead as appropriate responses to particular problems in given situations which, given the ever-evolving standards for instruction and expectations regarding student achievement, will continue to shift over time. From an evidentiary perspective, scale-up signals confidence that robust evidence of meaningful change warrants widespread adoption. Scalable interventions are not, however, dead-end products of an R&D process from which further movement is neither possible nor desirable. Continual examination of exemplary interventions is vital to ensure their continued viability.

This is the case for interventions warranted by the sequential 'proof-of-concept to efficacy to effectiveness trial' experimental model of evidence generation, but also for those whose positive effects are established in other ways. Consider the secondary analyses that provide the evidence warranting various grade retention and remedial instruction policies. Analyses of administrative records can yield incontrovertible evidence of the benefits of ending social promotion policies, but periodic re-analyses to establish the veracity of these conclusions can change as new student populations move through the education system. In thinking about the standards of scientific design necessary to warrant the adoption of new educational policies and practices, it is critical to remember that science must evolve if only to ensure static outcomes in dynamic contexts.

Principle 4: Recognizing Magnitudes of Change

Even when designs support causal inference, care needs to be exercised in interpreting their import. Critical is distinguishing statistically significant from substantively meaningful changes. When findings are statistically significant, we can be confident (within specified boundaries, e.g., 95% of the time) that observed results are not likely due to chance. However, statistical significance is not always substantively meaningful, signaling important differences meriting attention or action.

[12] For a complete listing of current research projects being conducted by research faculty at the Florida Center for Reading Research, see http://www.fcrr.org/centerResearch/centerResearch.shtm

Some results (e.g., an increase in scores on a test of student achievement following exposure to an intervention) provide clear indications of changes which are meaningful and worth replication. In such cases, the metrics employed to measure the results are unambiguously aligned with our educational objectives. Unfortunately, not all primary effects (e.g., changes in test scores) are inherently meaningful, and there are wide variations in metrics and measurement scales. To address these difficulties in interpreting primary findings, researchers increasingly report the size of an effect (i.e., change attributed to an intervention) not only in absolute terms (e.g., the number of points scored on a test of basic skills) but also on a common scale which facilitates comparisons of outcomes (see, e.g., Hedges 1981).

Such 'effect size' metrics are invaluable in assessing the practical import of changes that follow exposure to interventions. Yet even when confidence is high that observed changes following implementation of an intervention are both real (statistically significant) and substantively meaningful (in absolute or effect size terms), questions often remain regarding the implications of study findings for particular individuals in specific contexts. For example, an intervention that boosts academic achievement in mathematics by a third of a grade level may produce important benefits for students near the middle of a test-score distribution, yet have far less import for students at the bottom of the distribution. When average growth is 1 year of schooling, it is vital to consider whether an intervention is likely to help a student who starts the school-year more than a year behind her grade-level peers. Given how much of the variation in academic performance is accounted for by external factors outside the classroom, it is important to establish parameters within which it is reasonable to expect a single teacher to help raise student performance over the course of an academic year. Even evidence of large effects may not be sufficient to warrant support for an intervention in all circumstances or contexts.

The importance of context and its impact on magnitude is particularly evident with respect to efforts to improve student achievement by reducing class size. Tennessee was one of the first states to undertake a statewide class-size reduction initiative, the Student/Teacher Achievement Ratio (STAR) project. Implemented in 1985, the STAR project was designed to study the effects of reduced class sizes on kindergarten through third grade. Students were randomly assigned to one of three classroom size conditions (a 'small' class of 13–17 students per teacher, a 'regular class' of 22–25 pupils, and a 'regular-with-aide' class of 22–25 students with a full-time teacher's aide), and remained in the same classroom size from kindergarten through third grade. Data were collected from 79 schools and over 7000 students throughout the state, with outcome data including the Stanford Achievement Test (SAT), the Basic Skills First (BSF) performance tests (starting in first grade), and the SCAMIN self-concept and motivation scales (see Word et al. 1990).

Overall results from the STAR program showed that students uniformly benefited from smaller classes, scoring significantly higher on standardized tests of reading and math across grades and regardless of whether the small classes were in urban, suburban, or rural schools. Students in small classes outperformed students in classrooms with full-time teacher aides, the only exception being when aides were in regular first grade classrooms. Despite some concerns regarding student

attrition and movement between classrooms, and the inability to generalize results to very small or ethnically diverse schools, the experimental results of Project STAR held up under considerable scrutiny (Schanzenbach 2006).

So impressive were the results from the STAR program that the research was used to justify a similar effort in California. In the mid-1990s elementary schools in California averaged 29 students per classroom, the highest in the country. Regional economic prosperity provided tax revenues, over $1 billion per year that allowed bringing all K-3 classroom sizes down to 20 or fewer students. However, when class size reduction was implemented in California the outcome was quite different from that experienced in Tennessee.

The 1996 California class size reduction initiative affected over 1.6 million public school students in kindergarten through the third grade (see Bohrnstedt et al. 2000). This ambitious reform was carefully chronicled and evaluated by a research consortium whose members included the American Institutes for Research (AIR), RAND, WestEd, Policy Analysis for California Education (PACE), and EdSource. Key outcomes assessed in this 4-year, non-experimental evaluation of the California program included not only impacts on student achievement but also the quality of the state's teaching corps (Bohrnstedt and Stecher 2002). Since there was no random assignment of students to classrooms and the program was being implemented statewide, analyses of achievement gains relied on controlling for student and school characteristics and tracking cohorts of students with varying exposures to class size reduction.

Despite these methodological limitations, based on analyses of state data supplemented with information (including internal evaluation reports and specially-prepared student and teacher data sets) from school districts, the evaluators ultimately concluded that the relationship of the program to student achievement was inconclusive and attribution of gains in scores to the program was not warranted. One possible reason for this contrary finding is that rapid statewide implementation greatly increased the demand for teachers the year before the program was implemented. The demand for new teachers was met, in part, by hiring teachers not yet fully credentialed. In addition, most California districts also lacked sufficient funds to fully implement the program, often leading to a reallocation of resources from other programs and services.

The California experience suggests that policies that work in one place may not work in another, and moving to a statewide reduction in class size may have been premature. Importantly, recommendations arising from the California experience underscored the need to consider potential unanticipated consequences, contextual differences, and local adaptations that may be necessary to successfully bring to scale interventions that previously had produced meaningful change. The Tennessee STAR class size reduction project embraced scientific research principles, in both its design and its evaluation, and achieved impressive, substantively meaningful results. Results of a similar magnitude were not achieved, however, when an, on the face of it, quite similar reform was implemented in another context. The student populations were similar (K-3 public elementary school students) but critically the instructional work force with whom these students now had the opportunity to come

into closer daily contact was not. Tennessee's and California's different experiences with class size reduction policies underscore the need when making judgments about evidence that is statistically significant and substantively meaningful, that salient contextual factors in this case the quality and experience of the teacher can make major differences in results.

Principle 5: Judging the Evidence for Scale-Up

Questions about context are central to efforts to 'scale-up' interventions, extending the reach of policies and taking promising practices to larger diverse populations. Since the late 1990s, the scale-up model's stage-wise progression from innovation and proof of concept to widespread implementation of effective interventions has attained considerable traction in the U.S. among both policymakers and researchers as a framework for accumulating evidence in support of reform. Scale-up has become the implicit end-game of many R&D initiatives, the ultimate goal of a research and development process that begins with proving the concept behind an intervention, moves on to establish efficacy in ideal then document effectiveness in 'real world' contexts, all the while accumulating a body of knowledge as the foundation for judgments regarding the possibility (or undesirability) of scaling things that 'work' (with one population, in one context to others). Increasingly it has also become an explicit standard guiding research funding decisions. Embraced by governmental and philanthropic organizations alike, the scale-up heuristic underscores key differences in the aims and strategies of generating evidence to inform educational reform, providing a framework that guides study design and focuses attention on the types of evidence it is reasonable to demand before implementing largescale systematic reforms.

Importantly, with this emphasis on the pathways to devising largescale solutions, the question shifted from the straightforward (if not always straightforward to answer) 'what works?' to the more nuanced 'what works when, for whom, under what conditions?' Answers to these more finely-grained questions are critical if both human capital and financial resources are to be targeted efficiently and effectively to improve educational outcomes. But to answer them often requires substantial resources and a shortened timeline to implementation. Leveraging the wealth of administrative and accountability data can be a seedbed for designing and implementing future reforms. Properly mined, such data hold the potential to identify teachers, classrooms, schools and districts which, on the face of it, appear to be 'over-performing' (e.g., in comparison to population norms). Such outliers can then be examined more closely to see if their success are identifiable and potentially replicable in other settings.

Secondary analyses of major national datasets can also be invaluable in suggesting and monitoring the effects of strategies for implementing sound educational practices at scale. An example is research conducted by Richard Ingersoll to establish the prevalence and correlates of out-of-field teaching in U.S. public elementary

and secondary schools. Drawing on personal insights and experience as a secondary school teacher in Canada and the U.S., Ingersoll (1998) observed first-hand meaningful differences in student performance when teachers were assigned to offer instruction in subjects in which they were not specifically trained. Beginning with the U.S. Schools and Staffing Survey (SASS) that surveyed teachers, principals, and district administrators to comprehensively learn the characteristics of the instructional workforce; conditions in schools; and other related issues, he analyzed this administration survey data from several decades.[13] Ingersoll and colleagues found substantial proportions of high school teachers taught classes for which they were not adequately qualified, a problem exacerbated by teacher turnover. Subsequent analyses continued to document meaningfully high levels of outof-field teaching, leading Ingersoll to characterize the problem nearly a decade later as "chronic and widespread" (Ingersoll 2004: 14).

The data on the prevalence of out-of-field teaching (and subsequent replications of Ingersoll's findings) began to shape discourse and strategies for addressing the larger issue of what it takes to ensure equal access to high quality instruction (see, e.g., Ingersoll 1999). Particularly powerful was the inclusion in the No Child Left Behind Act of 2002 (U.S. Pub. L. 107-110) in its definitions of 'highly qualified' public elementary or secondary school teachers specific requirements for demonstrating competence in all academic subjects taught. These requirements included holding advanced degrees and passing state tests or graduate coursework in specific areas. However, knowledge of subject matter does not, of course, guarantee quality teaching, or even qualified teachers (Ingersoll et al. 1995). Such implicit choices and tradeoffs (e.g., devoting resources to placing more qualified teachers in classrooms versus expending the same resources to redress more fundamental socioeconomic inequalities, or calculating the moderating effect of the latter on investments in the former) underscore the important role judgment is likely to continue to play in decisions regarding the desirability of enacting laws and issuing regulations to address perceived shortcomings in the educational system, and reaching conclusions more generally regarding the scalability of interventions.

The intuitive appeal of evidence documenting the prevalence of 'poorly qualified' teachers is considerable; at some level, the evidence of out-of-field teaching has face validity so powerful that protracted testing to confirm this problem seems unwarranted. A counterargument however, could be made that one cannot be assured resources allocated to placing more highly qualified teachers in classrooms will prove more effective than resources devoted to better diagnostic assessments, computerized tutoring, and more offerings in online learning opportunities. Rich longitudinal national and state datasets coupled with sophisticated analytic procedures hold great promise for identifying potentially troubling characteristics of underperforming classrooms, schools, and districts, and for suggesting corrective actions for achieving best practices at scale. Ingersoll's important work on the prevalence of

[13] For a detailed description of the Schools and Staffing Survey, including copies of instrumentation administered in 1987–1988 m 1990–1991, 1993–1994, 1999–2000, 2003–2004, and 2007–2008, see the National Center for Education Statistics online at http://nces.ed.gov/surveys/sass/index.asp

out-of-field teaching, while not causal, presents robust evidence that underlie our judgments regarding which practices are indeed 'best' and strongly related to desired outcomes.

The availability of finely-grained data and efforts to support cultures of data sharing and data linkage suggest we may well be moving towards having the information necessary to document and weigh such tradeoffs, but it is unclear whether other obstacles to evidence-based education will ever be overcome. Reverse engineering exemplary practices already in the field (e.g., as identified through data mining that focuses attention upon districts, schools, and classes in which unusually large achievement gains are made over the course of a school-year) may help short-circuit the time intensive research and development process. But randomized control trials to ensure these outlier effects are replicable may take years to produce results. It is thus unlikely – and indeed would arguably be wrong to insist – that experimental evidence will ever become the sole basis for reform. Innovation and evidence generation will continue to proceed side-by-side, and important education policy decisions will continue to be made absent the most robust evidence scientific education research can provide. Moreover, judgment will always come in to play in weighing evidence. The task for educational researchers is to provide frameworks in which reasonable judgments can be made regarding the risks and likely benefits of supporting change with more and less of an empirical base.

Principle 6: Accumulating Knowledge for Generalizability

It is important in weighing evidence to consider whether or not study findings are applicable to a broader population. If every member of a population were affected equally by an intervention – i.e., if treatment effects were homogeneous – then results of any well-designed study would be generalizable to the population in its entirety. Typically, however, we expect that specific individuals (e.g., students, teachers) and organizations (e.g., schools, districts) will be differentially affected by interventions. Specifically, we expect populations themselves to be heterogeneous and anticipate key characteristics of population elements (e.g., the developmental trajectory of students in a classroom, the experience of instructors teaching in a particular field, the social organization of a school) will moderate interventions' impacts, resulting in heterogeneous intervention effects.

One way to enhance the generalizability of study findings is to address such variations (or covariates) at the design stage, specifying procedures for drawing the sample that will be investigated. For example, individuals might be randomly selected from the population to constitute the study sample and members of the sample might then be randomly assigned to receive or not receive an intervention. Alternatively, when distinct segments of the population share characteristics known (or hypothesized) to affect the outcome of interest and/or the likelihood of having a positive response to an intervention, these subgroups may constitute strata from which sample members may be selected purposively.

Leveraging information regarding subgroup characteristics is valuable not only in designing representative samples but also to an alternative strategy for estimating the generalizability of findings. Specifically, information on covariates and the probability these covariates predict selection into the study sample can be utilized to identify the inferential population to which the sample applies (i.e., the population of which the sample is representative), and to estimate average treatment effects for that subpopulation. In this way, we can be more confident of the broader applicability of findings found in studies of samples which are underrepresented either by design or as a result of implementation problems (such as inability to secure cooperation or attrition).

The Scaling Up SimCalc project conducted by Jeremy Roschelle and colleagues, integrates technology, curriculum, and teacher professional development to support middle school students in learning key mathematical concepts.[14] In the scale-up project, two large-scale randomized controlled trials and a quasi-experiment were conducted with middle-school teachers in Texas. These studies, found statistically significant and meaningful treatment effects on student learning (see Roschelle et al. 2007). As random assignment to treatments was not feasible, the investigators had to seek alternative methods to estimate the generalizability of study findings (Tipton 2011).

Utilizing data on 26 covariates (including school-level achievement, aggregated student and teacher demographics, and school funding and structure), analysts were able to identify a subpopulation characterized by the 78 schools in the study sample – i.e., a population to which the study sample generalizes (see Tipton 2011; and Roschelle et al. 2010b).[15] Subsequent re-analyses of the SimCalc data (Tipton 2011) suggested this line of inquiry proved promising. Both at the design stage and as sampling strategies are implemented and studies unfold, educational research frequently explores impacts of interventions within non-representative samples. We are not advocating that this is the ideal situation, but realize it is one that often occurs in education studies as researchers work toward studying interventions anticipating the likelihood of scale-up.

The SimCalc work illustrates the possibility of generalizing appropriately findings of even those studies which are not at the design stage devised to represent the population of ultimate interest. This is not to say that efforts to conduct studies of the impacts of interventions upon representative samples of populations should be abandoned, but as the example illustrates it may be possible to draw sound conclusions regarding the extendibility or potential broader impacts of a particular set of study findings. These researchers' innovative use of statistical techniques to create

[14] For information about the SimCalc intervention and the scaling-up SimCalc study, see the Kaput Center for Research and Innovation in STEM Education (http://www.kaputcenter.umassd.edu/projects/simcalc), the SRI International Scaling Up SimCalc project website (at http://math.sri.com/index.html), and Roschelle et al. (2010b).

[15] Specifically, using a method and a propensity score sub classification estimator introduced by O'Muircheartaigh and Hedges reduced "bias in the estimate of a population average treatment effect" and identified "the portion of a population for which an experiment can generalize with fewer costs in terms [of] bias, variance, and extrapolation" (Tipton 2011: 4).

their representative population shows great promise for assessing the impact of an intervention and generating broadly generalizable findings (Hedges 2013; O'Muircheartaigh and Hedges 2014; Tipton et al. 2014; Tipton 2014).

This cutting-edge approach leveraged information derived from extant data collections to define a population to which it is reasonable to generalize the SimCalc findings, underscoring the research value of state and federal data systems and supporting a culture of data sharing (with appropriate privacy and confidentiality safeguards).

Administrative data are increasingly being used to assess state level interventions including changes in curricular requirements, teacher effectiveness, and scholarship programs to enable postsecondary attendance. Federal compliance and state data systems not only have key roles to play in administering and ensuring accountability across educational systems, but can also (when shared and linked) be used for a variety of analytic purposes, including deriving and testing hypotheses regarding factors that contribute to and impede instruction, learning, and achievement, and addressing issues such as small sample size, unrepresentative samples (e.g., due to the challenges of recruiting study participants, differential attrition) and other statistical problems that plague educational research. As the SimCalc example shows, working with administrative data can ease the process of generating evidence that warrants the move from intervention development to scale-up. Critically, strengthening the elements of the state and federal data systems, and the mechanisms and cultures for linking these with primary data from studies such as the SimCalc evaluation, provide new opportunities to appropriately contextualize single-study findings, assisting practitioners, policymakers, and educational researchers in making principled judgments regarding the generalizability of their findings.

Principle 7: Conducting Research for the Public Good

An important goal of educational research in an era of evidence-informed decision-making is to promote the utilization of knowledge resulting from scholarly inquiry in support of the public good. Research conducted for the public good tackles issues of broad social interest. Striving to ensure research results in the greatest possible good for the largest number of individuals brings us back full circle to the importance of investigating issues that matter. Issues highly salient to only a small number of individuals merit exploration, but it is critically important for investigators and funders alike to ask themselves at every step of the educational research process 'who benefits from this work?' and 'do the potential implications of the evidence warrant the resources required to support the inquiry?'

A common appeal to motivate interest in educational research is to link education and learning with future economic competitiveness (for the individual and/or nations and society more generally). Examples include educational research that seeks to support underrepresented groups in preparing for and achieving successful transitions to postsecondary education and careers in STEM and other fields. One such

study is an intervention designed to facilitate the successful entry of minority youth into health research careers, Training Early Achievers for Careers in Health (TEACH) research, directed by Vineet Arora M.D. The TEACH intervention was itself the product of research on an important social issue: factors affecting low-income urban high school students' matriculation to college. Informed by extensive analyses of longitudinal observational data and a resulting theory regarding the importance of aligning students' knowledge, attitudes, and behaviors to attain their ambitions (see Schneider and Stevenson 1999), the TEACH program was designed to foster 'aligned ambitions' (educational expectations in sync with occupational aspirations) for Chicago area high school students interested in preparing for health research careers. TEACH enabled students to engage in realistic health career experiences (e.g., internships and opportunities to observe clinical rounds) and to receive mentoring support from a multi-tiered structure of peers that includes high school student peers, undergraduate students, medical school students, and clinical research faculty.[16]

Drawing on lessons learned from the TEACH experience and with evidence of the efficacy of that intervention behind them, in 2009 a team of researchers from Michigan State University's College of Education collaborated with a sample of central Michigan high schools to launch the College Ambition Program (CAP), a school-wide initiative that like TEACH seeks to align ambitions and "give students the support system they need to make it to, and in, postsecondary education" (Schneider 2015). CAP investigators seek evidence on the merits and limits of their intervention striving to make changes for the public good (in this case improving the educational opportunities for low-income and minority children). In practice this means not only employing research designs capable of yielding evidence of meaningful change at the end of the 3-year study period, but ensuring those not selected to be part of the CAP treatment condition are not disadvantaged by serving in the controlled comparison group (for example, a wide range of online resources to support students in planning to attend postsecondary institutions are publicly available through the study website).[17]

Applying These Principles for Educational Research

Another dimension of what it means to conduct research for the common good is to ensure access and improve the communication of research findings. Data upon which analyses are based and the measures employed in collecting them should be seen as public goods, and appropriately documented, archived, and made available for confirmatory or secondary analyses. A commitment to data sharing is critical to

[16] For additional information on the TEACH (Training Early Achievers for Careers in Health) Research program see http://chess.uchicago.edu/TEACH

[17] For additional information on the College Ambition Program and the NSF-supported Transforming Interests in STEM Careers (TISC) study evaluating its impacts see the program website at http://collegeambition.org/

facilitate the replications that increase confidence in findings. It is also vital to leverage investments in often costly primary data collections and encourages careful training in and application of best practices for recording and tracing provenance, and documenting the coding, re-coding, and data transformation decisions to create archival-quality data for secondary study. A corollary to a commitment to data sharing is access. Whether research entails primary data collection or relies on secondary data analyses, investigators have moral and legal obligations to handle (e.g., collect, store, analyze, and report) data responsibly and in accordance with provisions governing the protection of human subjects.

In education, individual studies and larger programs of research are designed not only to generate new evidence on what works to improve instructional practice, educational attainment, and lifelong learning but to inform practice and policy. With these broader goals in mind the criteria we have presented here encourage researchers to consider the intrinsic value of the topic being explored, the capacity to recognize and measure meaningful change, the broader applicability (scalability and generalizability) of findings, and how the research aligns with larger public interest objectives.

Although there are many criteria for assessing the quality of educational research, establishing standards for them is challenging, in part because of the tradeoffs inherent among them. Different stakeholders are likely to attach more or less importance to individual criteria at each stage in the research process. In education as in other fields it is not only the evidence educational science generates but assessments of its quality are often socially constructed and subject to disagreement. Evidence is meant to inform, and some does it better than others. Educational researchers have a critical role to play in providing decision-makers with the tools to judge the evidence before them. Ultimately, however, judgments will need to be made. Our goal is to identify a set of principles for interrogating the quality of evidence especially for studies conducted in the public interest that are designed to inform educational reform.

Acknowledgement This material is based upon work supported by the National Science Foundation under awards: No. DRL-131672 (CAP), No. OISE-1545684 (PIRE), and No. DRL-0815295 (ARC).

References

American Educational Research Association (AERA). (2016). Retrieved from homepage: https://www.aera.org

American Recovery and Reinvestment Act of 2009 (ARRA). (2009, February 19). Pub. L. No. 111–5, 123 Stat. 115, 516.

BioKIDS: Kids' Inquiry of Diverse Species. (2005). Retrieved from http://www.biokids.umich.edu

Bohrnstedt, G. W., & Stecher, B. M. (2002). *What we have learned about class size reduction in California*. Sacramento: California Department of Education.

Bohrnstedt, G., Stecher, B., & Wiley, E. (2000). The California class size reduction evaluation: Lessons learned. In *How small classes help teachers do their best* (pp. 201–226). Philadelphia: Temple University Center for Research in Human Development and Education.

Borman, G. D., Hewes, G. M., Overman, L. T., & Brown, S. (2003). Comprehensive school reform and achievement: A meta-analysis. *Review of Educational Research, 73*(2), 125–230. doi:10.3102/00346543073002125.

Borman, G., Slavin, R. E., Cheung, A., Chamberlain, A., Madden, N. A., & Chambers, B. (2007). Final reading outcomes of the national randomized field trial of success for all. *American Educational Research Journal, 44*(3), 701–731. doi:10.3102/0002831207306743.

Bryk, A. S. (2015). Accelerating how we learn to improve. *Educational Researcher, 44*(9), 467–478.

Campbell Collaboration: Vision, Mission, and Key Principles. (2016). Retrieved from https://www.campbellcollaboration.org/vision-mission-and-principle/explore/our-key-principles

Campuzano, L., Dynarski, M., Agodini, R., & Rall, K. (2009). *Effectiveness of reading and mathematics software products: Findings from two student cohorts*. NCEE 2009–4041. National Center for Education Evaluation and Regional Assistance.

Children's Learning Institute (CLI) at The University of Texas-Houston Health Science Center and the Texas Institute for Measurement, Evaluation, and Statistics (TIMES) *Technical Report TPRI* (2010–2014 Edition). Retrieved from: http://tpri.org/resources/documents/20102014TechnicalReport.pdf

Children's Learning Institute: TPRI Early Reading Assessment. (2015). Retrieved from https://www.childrenslearninginstitute.org/resources/tpri-early-reading-assessment

Chmielewski, A. K. (2014). An international comparison of achievement inequality in within-and between-school tracking systems. *American Journal of Education, 120*(3), 293–324.

College Ambition Program. (2016). Retrieved from homepage: http://collegeambition.org

Duncan, G. J., & Murnane, R. J. (Eds.). (2011). *Whither opportunity?: Rising inequality, schools, and children's life chances*. New York: Russell Sage Foundation.

Dynarski, S., & Scott-Clayton, J. E. (2007). The feasibility of streamlining aid for college using the tax system. In *National Tax Association papers and proceedings* (vol. 99, pp. 250–262).

Every Student Succeeds Act of 2015, S. 1177, 114th Cong. (2015). Washington, DC: US Department of Education. Public Law 114–95.

Florida Center for Reading Research. (2008). Retrieved from homepage: http://www.fcrr.org/centerResearch/centerResearch.shtm

Foorman, B. R., & Petscher, Y. (2010). Development of spelling and differential relations to text reading in grades 3–12. *Assessment for Effective Intervention, 36*(1), 7–20.

Foorman, B. R., Fletcher, J. M., Frances, D. J., Carlson, C. D., Chen, D., & Mouzaki, A. (1998). *Technical report: Texas primary reading inventory* (1998th ed.). Houston: Center for Academic and Reading Skills and the University of Houston.

Foorman, B., Santi, K., & Berger, L. (2007). Scaling assessment-driven instruction using the internet and handheld computers. In B. Schneider & S. K. McDonald (Eds.), *Scale-up in education* (pp. 68–90). Plymouth: Rowman & Littlefield Publishers.

Foorman, B., Torgesen, J., Crawford, E., & Petscher, Y. (2009). Assessments to guide reading instruction in K-12: Decisions supported by the new Florida system. *Perspectives on Language and Literacy, 35*(5), 13–19.

Gates Foundation. (2016). Retrieved from homepage: http://www.gatesfoundation.org

Gotwals, A. W., & Songer, N. B. (2006). Measuring students' scientific content and inquiry reasoning. In *Proceedings of the 7th international conference on learning sciences* (pp. 196–202). International Society of the Learning Sciences.

Harris, S. (2016). Retrieved from homepage: https://www.samharris.org

Hedges, L. V. (1981). Distribution theory for Glass's estimator of effect size and related estimators. *Journal of Educational and Behavioral Statistics, 6*(2), 107–128.

Hedges, L. V. (2013). Recommendations for practice: Justifying claims of generalizability. *Educational Psychology Review, 25*(3), 331–337. doi:1040-726X.

Holland, P. W. (1986). Statistics and causal inference. *Journal of the American Statistical Association, 81*(396), 945–960.

Hoxby, C. M. (Ed.). (2007). *College choices: The economics of where to go, when to go, and how to pay for it*. Chicago: University of Chicago Press.

Imbens, G. W., & Rubin, D. B. (2010). Rubin causal model. In S. N. Durlauf & L. E. Blume (Eds.), *Microeconometrics* (pp. 229–241). New York: Macmillan.

Ingersoll, R. M. (1998). The problem of out-of-field teaching. *The Phi Delta Kappan, 79*(10), 773–776.

Ingersoll, R. M. (1999). The problem of underqualified teachers in American secondary schools. *Educational Researcher, 28*(2), 26–37.

Ingersoll, R. M. (2004). *Why do high-poverty schools have difficulty staffing their classrooms with qualified teachers?* Center for American Progress, Institute for America's Future.

Ingersoll, R. M., Han, M., & Bobbitt, S. (1995). *Teacher supply, teacher qualifications, and teacher turnover: 1990–1991* (pp. 95–744). Washington, DC: U.S. Department of Education, Office of Educational Research and Improvement, National Center for Education Statistics, NCES.

Kaput Center for Research and Innovation in STEM Education. (2016). University of Massachusetts, Dartmouth. Retrieved from: http://www.kaputcenter.umassd.edu/projects/simcalc

KIPP: About KIPP. (2016). Retrieved from homepage: http://www.kipp.org

KnowHow2Go. (2013). Retrieved from homepage: http://www.knowhow2go.org

Milesi, C., Brown, K., Hawkley, L., Dropkin, E., & Schneider, B. (2014). Charting the impact of federal spending for education research: A bibliometric approach. *Educational Researcher, 43*(7), 361–370. doi:10.3102/0013189X14554002.

National Center for Education Statistics. (2011a). *College Navigator.* Retrieved from http://nces.ed.gov/collegenavigator

National Center for Education Statistics. (2011b). *The nation's report card: Reading 2011 (NCES 2012–457)*. Washington, DC: Institute of Education Sciences, U.S. Department of Education.

National Center for Education Statistics. (2016). *Schools and Staffing Survey (SASS)*. Washington, DC: Institute of Education Sciences, U.S. Department of Education. Retrieved from: http://nces.ed.gov/surveys/sass/index.asp.

National Research Council. (2002). Scientific research in education. Committee on Scientific Principles for Education Research. In R. J. Shavelson & L. Towne (Eds.), *Center for education, division of behavioral and social sciences and education*. Washington, DC: National Academy Press.

National Science Foundation. (2010a). *Preparing the next generation of stem innovators: Identifying and developing our nation's human capital*. National Science Foundation. Retrieved from: https://www.nsf.gov/nsb/publications/2010/nsb1033.pdf

National Science Foundation. (2010b). *Research and Evaluation on Education in Science and Engineering (REESE). Program Solicitation.* Retrieved from: http://www.nsf.gov/pubs/2010/nsf10586/nsf10586.pdf

No Child Left Behind Act of 2002, S. 1115, 107th Cong. (2002). Washington, DC: US Department of Education. Public Law 107–110.

O'Muircheartaigh, C., & Hedges, L. V. (2014). Generalizing from unrepresentative experiments: A stratified propensity score approach. *Journal of the Royal Statistical Society, Series C, 63*(2), 195–210. doi:10.1111/rssc.12037.

Organization for Economic Co-operation and Development, OECD/EU. (2015). *Indicators of Immigrant Integration 2015: Settling In*, OECD Publishing, Paris. 1–348. doi:http://dx.doi.org/10.1787/9789264234024-en

Pellegrino, J. W., Chudowsky, N., & Glaser, R. (Eds.). (2001). *Knowing what students know: The science and design of educational assessment*. Washington, DC, National Academies Press.

Pellegrino, J. W., Wilson, M. R., Koenig, J. A., & Beatty, A. S. (Eds.). (2014). *Developing assessments for the next generation science standards*. Washington, DC: National Academies Press.

Principled Assessment Designs for Inquiry. (2003). Retrieved from homepage: http://padi.sri.com

Quint, J., 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) evaluation*. New York: MDRC.

Reardon, S. F. (2011). The widening academic achievement gap between the rich and the poor: New evidence and possible explanations. In R. Murnane, & G. Duncan (Eds.), *Whither opportunity? Rising inequality and the uncertain life chances of low-income children*. New York: Russell Sage Foundation.

Ritter, S., Anderson, J. R., Koedinger, K. R., & Corbett, A. (2007a). Cognitive Tutor: Applied research in mathematics education. *Psychonomic Bulletin & Review, 14*(2), 249–255.

Ritter, S., Kulikowich, J., Lei, P. W., McGuire, C. L., & Morgan, P. (2007b). What evidence matters? A randomized field trial of Cognitive Tutor Algebra I. *Frontiers in Artificial Intelligence and Applications, 162*, 13.

Roschelle, J., Tatar, D., Shechtman, N., Hegedus, S., Hopkins, B., Knudsen, J., & Stroter, A. (2007). Can a technology-enhanced curriculum improve student learning of important mathematics. *Results from 7th grade, year*, *1*.

Roschelle, J., Shechtman, N., Tatar, D., Hegedus, S., Hopkins, B., Empson, S., Knudsen, J., & Gallagher, L. P. (2010a). Integration of technology, curriculum, and professional development for advancing middle school mathematics three large-scale studies. *American Educational Research Journal, 47*(4), 833–878.

Roschelle, J., Tatar, D., Hedges, L., & Shechtman, N. (2010b). Two perspectives on the generalizability of lessons from scaling up SimCalc. *Society for Research on Educational Effectiveness*.

Rowan, B., Correnti, R., Miller, R., & Camburn, E. (2009). School improvement by design: Lessons from a study of comprehensive school reform programs. *Consortium for Policy Research in Education*, 1–62. doi:10.12698/cpre.2009.sii.

Rubin, B. (2005). Bayesian inference for causal effects. In C. R. Rao & D. K. Dey (Eds.), *Handbook of statistics, volume 25: Bayesian thinking: Modeling and computation* (pp. 1–16). Amsterdam: Elsevier.

Schanzenbach, D. W. (2006). What have researchers learned from Project STAR? *Brookings Papers on Education Policy, 9*, 205–228.

Schneider, B. (2015). 2014 AERA Presidential Address, The College Ambition Program: A realistic transition strategy for traditionally disadvantaged students. *Educational Researcher, 44*(7), 394–403.

Schneider, B., & McDonald, S. K. (2007). Scale-up in practice: An introduction. In B. Schneider & S. K. McDonald (Eds.), *Scale-up in education: Vol. 2: Issues in practice* (pp. 1–12). Lanham: Rowman & Littlefield.

Schneider, B., & Stevenson, D. (1999). *The ambitious generation: America's teenagers, motivated but directionless*. New Haven: Yale University Press.

SimCalc, the mathematics of change. (2011). Retrieved from: http://math.sri.com/index.html

Society for Research on Educational Effectiveness (SREE): Mission. (2010). Retrieved from https://www.sree.org/pages/mission.php

Songer, N. B., Myers, P., & Gotwals, A. W. (2007). *DeepThink: Fostering and measuring learning progressions focused on deep thinking about biodiversity*. Poster presented at the Principal Investigators Meeting of the National Science Foundation, Washington, DC.

Songer, N. B., Kelcey, B., & Gotwals, A. W. (2009). How and when does complex reasoning occur? Empirically driven development of a learning progression focused on complex reasoning about biodiversity. *Journal of Research in Science Teaching, 46*(6), 610–631.

Stevenson, D. L. (2000). The fit and misfit of sociological research and educational policy. In M. T. Hallinin (Ed.), *Handbook of the sociology of education* (pp. 547–563). Springer US.

Success for All Foundation: Our Story. (2005). Retrieved from http://www.successforall.org/who-we-are

The Center for Health and the Social Sciences. (2016). *High school students: Training early achievers for careers in heatlh (TEACH).* The University of Chicago. Retrieved from: http://chess.uchicago.edu/TEACH

Tipton, E. (2011). *Improving the external validity of randomized experiments using propensity score subclassification*. Working Paper.

Tipton, E. (2014). How generalizable is your experiment? An index for comparing experimental samples and populations. *Journal of Educational and Behavioral Statistics, 39*(6), 478–501.

Tipton, E., Hedges, L. V., Borman, G., Vaden-Kiernan, M., Caverly, S., & Sullivan, K. (2014). Sample selection in randomized experiments: A new method using propensity score stratified sampling. *Journal of Research on Educational Effectiveness, 7*(1), 114–135. doi:10.1080/19345747.2013.831154.

U.K., House of Commons. (2006). *Science and technology committee: scientific advice, risk and evidence based policy making* (Vol. 1). House of Commons.

U.S. Department of Labor. (2011). *A profile of the working poor, 2009.* U.S. Department of Labor, U.W. Bureau of Labor Statistics. March 2011. Retrieved from: http://www.bls.gov/opub/reports/working-poor/archive/workingpoor_2009.pdf

Urban Prep Academies: History. (2012). Retrieved from: http://www.urbanprep.org/about/history-creed

Walters, P. B., Lareau, A., & Ranis, S. (Eds.). (2008). *Education research on trial*. Taylor & Francis.

Weiss, C. H. (1982). Policy research in the context of diffuse decision making. *The Journal of Higher Education, 53*, 619–639.

Weiss, C. H. (1989). Congressional committees as users of analysis. *Journal of Policy Analysis and Management, 8*(3), 411–431.

What Works Clearinghouse. (2009). Intervention report: Cognitive tutor algebra I. Retrieved from https://www.mbaea.org/documents/filelibrary/pdf/cognitive_tutor/WWC_CogTutor_Report_July2009_B2A3C279D0481.pdf

What Works Clearinghouse. (2010). *What works clearinghouse: Quick review of the report "Student Characteristics and Achievement in 22 KIPP Middle Schools.* U.S. Department of Education, Institute of Education Sciences. Retrieved from: http://ies.ed.gov/ncee/wwc/Docs/QuickReview/kipp_092110.pdf

What Works Clearinghouse. (2016). What we do. Retrieved from: http://ies.ed.gov/ncee/wwc/WhatWeDo

Word, E., Johnston, J., Bain, H., Fulton, B. D., Zaharias, J. B., Achilles, C. M., Lintz, M. N. Folger, J. & Breda, C. (1990). The State of Tennessee's student/teacher achievement ratio (STAR) Project. *Tennessee Board of Education*.

World Education Research Association (WERA). (2016). Retrieved from homepage: https://www.wera.org

Yamada, H., & Bryk, A. S. (2016). Assessing the first two years' effectiveness of statway® a multilevel model with propensity score matching. *Community College Review*. 0091552116643162.