Behavioural Public Policy (2021), 5: 1, 60–70
© The Author(s) 2020. Published by Cambridge University Press doi:10.1017/bpp.2020.19

Scaling up experiments to reduce educational inequality

STEPHEN W. RAUDENBUSH*

University of Chicago Department of Sociology, Harris School of Public Policy, and Committee on Education, University of Chicago, Chicago, IL, USA

Abstract: For educational researchers, Al-Ubaydli *et al.* raise a crucial question: How can the science of scaling experimental innovations contribute to school improvement? By assessing how particular innovative programs work, why, for whom and under what conditions, experimenters test theories: about how children and youth learn, about how adults can collaborate to create the required learning opportunities and about how policy can supply the required incentives and resources to support such effective collaboration. My focus in this response article thus shifts from the perspective of innovators who hope to scale their interventions to the perspective of practitioners who face the challenge of adopting, adapting or borrowing ideas from experimental studies and harmonizing those ideas with expert clinical judgment.

Submitted 24 February 2020; accepted 20 April 2020

Al-Ubaydli *et al.* have contributed a comprehensive discussion of the challenges facing those who aim to expand promising innovations to new settings, and they offer recommendations to overcome these challenges. This welcome effort to create a 'science of scaling' is tethered to the goal of supplying policy-makers with better information, enabling them to make better decisions.

Over the past half century, most of those engaged in research and development on schools, including myself, have sought to discover and test new strategies for making schools work better for the most disadvantaged children and youth. For those of us who pursue this goal, Al-Ubaydli *et al.* raise a crucial question: How can the science of scaling experimental innovations contribute to school improvement that targets the reduction of inequality?

* Correspondence to: University of Chicago Department of Sociology, Harris School of Public Policy, and Committee on Education, University of Chicago, Chicago, IL, USA. E-mail: sraudenb@uchicago.edu

https://doi.org/10.1017/bpp.2020.19 Published online by Cambridge University Press

An obvious answer would be that this science can help policymakers select particular new interventions that work reliably for the most disadvantaged students. My answer is a bit different. Supporting decisions about the adoption of standalone interventions, while useful, is of comparatively modest importance in this project of school improvement. Instead, the big payoff from experimentation and scale-up is the advancement of theory that enables schools to reliably supply effective teaching and learning. Such theory would include understandings about who is likely to benefit from a particular intervention within a particular setting, how a particular intervention operates to produce its effect, holding all else constant, and how much the adoption of such an intervention will cost. These are precisely the theoretical issues that Al-Ubaydli *et al.*'s science of scaling is tailored to address. However, as Hassrick *et al.* (2017) wrote: "Effectively implementing a series of interventions in the daily life of a school is not like plugging new appliances into an electrical outlet. To be effective, new interventions have to be integrated together and meshed with current practice to improve the overall life of a school."

Such integration is predicated upon assumptions about how diverse children learn, how the daily work of teaching supports such learning and how schools conceived as instructional systems can support the work of teaching. The science of scaling holds promise to support this project inasmuch as it can help us to probe, test, revise and ultimately strengthen the theoretical understanding that underlies these assumptions.

Many of the most promising randomized trials in education to date reveal the impact on students of attending highly effective schools. Such schools are not interventions or innovations in the usual sense. Rather, they mesh a range of ideas, beliefs and dispositions that support coherently organized practice. When we look at effective schools, we do see that particular interventions have contributed significantly to effective practice. Moreover, I will review evidence that research on effective schools generates many questions about new interventions. In turn, the study of new interventions can feed back to support powerful school practice. It is this virtuous cycle that can, in my view, strengthen the theoretical basis of practice and generate major progress in the face of strong evidence that inequality is remarkably resilient. Proposals for improving intervention science, like those made by Al-Ubaydli *et al.*, seem essential to this project.

My focus thus shifts from the perspective of innovators who hope to scale their interventions to the perspective of the school effectiveness researcher who wishes to exploit the science of scaling to make schools work better.

Let's first review why the reduction in inequality looms so large in school effectiveness research. I will then consider the theoretical questions posed by the science of scaling, how school improvement can leverage these theoretical

https://doi.org/10.1017/bpp.2020.19 Published online by Cambridge University Press

gains and how the study of particularly effective schools can generate new intervention studies. From this vantage point, I will comment on the suggestions of Al-Ubaydli *et al.* for improving intervention science, with a focus on applications in school improvement.

Why focus on educational inequality?

Over the past half-century, most of those engaged in research and development in education have sought to reduce educational inequality. In part, this reflects the goals of federal educational policy since the passage of the Elementary and Secondary Education Act, better known as Title I, in 1965. In his 'Great Society' speech of 1964, President Lyndon Johnson said, "Poverty must not be a bar to learning, and learning must offer an escape from poverty."

Since the passage of Title I, the federal government has required the evaluation of new initiatives that aim to achieve these goals. In 2001, Congress passed the 'No Child Left Behind' law, a new incarnation of Title I, which aimed to hold school leaders accountable not only for their students' average achievement, but also for the achievement of subgroups defined by socioeconomic status and racial/ethnic background. The law also mandated the formation of the Institute of Education Sciences (IES), with the purpose of creating a new scientific basis for educational research. In 2002, Russell Whitehurst became the founding director of IES. With Whitehurst's commitment to experimental evaluation, and supported by a large increase in the budget for educational research, the IES funded more than 175 large-scale randomized controlled trials (RCTs) during the first decade of its existence (Spybrook, 2014; Spybrook *et al.*, 2016).

Following federal money has not been the only motive for testing innovations that aim to reduce inequality. Reducing inequality has also been key for private funders of education research, including the W.T. Grant Foundation, the Russell Sage Foundation and the Gates Foundation. The goal of expanding opportunity and reducing inequality has captured the imagination of many researchers and educators. It is no surprise, then, that prominent lists of educational innovations with strong evidentiary support, including the federally funded 'What Works Clearing House' (https://ies.ed.gov/ncee/wwc), the Arnold Foundation's 'Social Programs That Work' (https://evidencebasedprograms.org/policy_area/k-12-education) and Robert Slavin's Best Evidence Encyclopedia (http://www.bestevidence.org), mostly cite programs that were designed to reduce inequality.

Even if the sole aim of policy were strictly to improve average achievement, reducing inequality would loom large. The surprisingly poor performance of US students on international assessments of mathematics achievement is

https://doi.org/10.1017/bpp.2020.19 Published online by Cambridge University Press

arguably attributable primarily to the vast inequality of achievement in the USA. Park (2013), for example, showed that on the 2009 test administered by the Programme for International Student Assessment, students at the 90th percentile of math achievement in the USA scored about 25% of a standard deviation below those at the 90th percentile in South Korea. However, students at the 10th percentile in the USA were 75% of a standard deviation behind students at the 10th percentile in Korea. This is such a massive difference that it is hard to see how average achievement in the USA can approximate that of the highest-achieving countries without dramatically increasing the skills of the lowest-achieving students, many of whom are the low-income students that Title 1 was designed to help and that most experimental programs in the USA have targeted. How to increase these skills on a mass scale requires theory and evidence on how those skills develop over the life course and how interventions at each stage can jointly promote them.

Effective scaling requires and generates theoretical advance

The science of scaling asks how particular innovative programs work, why, for whom and under what conditions. Applied to school improvement, the answers to these questions entail propositions about how children and youth learn, how adults learn to collaborate in order to create the required learning opportunities and how policy can supply the required incentives and resources to support such effective teaching. Confidence that an intervention will work in a new setting depends on knowing whether the children and youth in that setting stand to benefit from the intervention and knowing whether local resources, including money, but also the skills, beliefs and commitments of local practitioners, are sufficient to put the intervention into practice. One must presumably have ways of assessing early on whether the implementation of the innovation is on track for success and strategies for correcting shortcomings. This process of short-term evaluation and correction requires theory about changes that are crucial to the intervention and measures of those changes. These are all theoretical questions that a science of scaling would help us answer – and questions that are central to improving schooling generally.

The task of producing powerful schooling is fraught with complexity and uncertainty, which is why learning is so central. Administrators and incumbent teachers need to learn about the qualities of applicants for teaching jobs and how to evaluate teaching practice; teachers have to learn about their children's skills and how to manage the heterogeneity of skills and dispositions; and children have to 'learn how to learn', which requires self-control, focused attention, persistent effort, prosocial behavior and the capacity to cope with

https://doi.org/10.1017/bpp.2020.19 Published online by Cambridge University Press

adverse feedback. Particular interventions can assist in particular aspects of these challenges. But how the knowledge from such particular interventions is mobilized to shape powerful practice is more complicated than simply adopting the intervention.

Learning to read

Perhaps the best example of how multiple interventions can shape practice is research on reading. No other aspect of schooling has been more studied, and many conclusive findings have emerged. A stream of evaluations of reading programs has strongly shaped practice, but the efforts to use these findings to improve practice has unearthed major gaps in knowledge. To fill these gaps requires theory that can enable us to combine insights across interventions over the early life course of the child.

Over several decades of the last century, researchers conducted a series of experimental and quasi-experimental studies, often supported by the National Institute of Mental Health (NIMH). Their aim was to reduce inequality in reading skill, focusing primarily on early reading, which typically occurs around ages 5–6. The aim at this stage is not to learn new ideas through reading, but rather to decode the text; that is, to recognize in printed text the words and phrases that the children already know. The National Reading Panel *et al.* (2000), organized by the NIMH, asserted that a scientific consensus had emerged from their review of thousands of studies. The central proposition was that the vast majority of children would learn to read if exposed to systematic instruction in word decoding, including the cultivation of phonemic awareness (knowing how sounds blend to make words and how sounds are related to letters within words), guided oral reading in small groups and discussions about the meaning of texts. Moreover, exposure of all children to such effective instruction would reveal that a small number of children required much more intensive intervention. A series of careful clinical studies showed that most members of this smaller group of children, who failed to benefit from regular instruction, could nonetheless become fluent readers if exposed to intensive, expert instruction over many weeks provided in very small classes.

The National Reading Panel framed the emerging theory in terms of public health. Effective classroom instruction was conceived as 'primary prevention', which minimizes the number of children who would need intensive treatment. A failure of primary prevention, resulting from inattention to explicit instruction in decoding in regular classroom instruction, would swamp a school with 'curriculum casualties' – typically developing children who would have learned to read had they received good early instruction. These non-readers would be

https://doi.org/10.1017/bpp.2020.19 Published online by Cambridge University Press

mixed with a much smaller but hidden group of children who would have required intensive intervention even if they had received good early instruction. Without primary prevention, it would be extremely expensive to mobilize the resources needed help children catch up.

Reading to learn

These findings were exciting. But even if perfectly implemented, successful word decoding, supported by extensive intervention research, would not be sufficient to produce high levels of reading comprehension. Learning to read – that is, learning to decode familiar text – is not the same as using reading skills to acquire new knowledge, often called 'reading to learn'. When one sees text that conveys new ideas, one encounters new vocabulary conveyed in academic English syntax. Assuming they are proficient decoders, children with large vocabularies and those who use academic English at home will have comparatively little difficulty comprehending novel text. Other children will find comprehension more challenging. We can regard this as a 'missing data' problem. If I encounter new text containing only a small fraction of unfamiliar words, I can readily impute the meanings of the novel text. However, if the fraction of missing words in the text is large, my capacity to impute is much more limited.

'Reading to learn' is essential to all subsequent academic pursuits. Promoting such reading poses a major challenge for elementary schools because social inequality in this skill forms early in life. Parental education and investments in children's academic learning early in life underlie the emergence of children's conceptual knowledge, vocabulary and familiarity with the syntax of academic English. Highly educated parents possess the kind of language that will underlie the later emergence of vocabulary and are especially likely to buy books for their children, to read to their children and to expose their children to various other learning opportunities (for a review, see Hassrick *et al.*, 2017). For these reasons, children come to school with vastly varied skills in academic English. Such differences may be muted when the aim is to teach the decoding of simple text. However, such differences in linguistic background will become readily apparent when the challenge is to comprehend new ideas in text, which happens in the USA in the second and third grades.

A dynamic treatment regime to reduce inequality

This reasoning suggests that a kind of dynamic treatment regime will be required to overcome social inequality in reading comprehension. Such a regime would begin at birth and continue into the early years of schooling.

https://doi.org/10.1017/bpp.2020.19 Published online by Cambridge University Press

This regime would rely for empirical support on a wide variety of lab and field studies conducted at each age. But its ultimate success would rely on theoretical understandings about how the stages of language and literacy are connected and how interventions at each stage amplify the effect of earlier interventions and create the foundation for later ones, similar in spirit to Cunha and Heckman's (2007) notion of dynamic complementarity. And it would require collaboration across institutions: the hospital, the preschool and the elementary school.

Following the work of Graf *et al.* (2017) on the 'Thirty Million Words' project, one can envision engaging all parents as language teachers as part of the regular care that we expect pediatrics to supply to young children. These authors report experimental evidence suggesting that, with some guidance, nearly all mothers, including those with limited educational attainment and low incomes, can enhance their very young children's language development. However, as the child's language skills develop, intervention outside the home will become essential to promote high levels of language development. Now the institutional framework is the highly diverse set of childcare providers, including many centers that might explicitly focus on academic instruction, and especially language development, as one aspect of their work. An emerging hypothesis is that reading aloud to children, discussing the text and engaging children in games and play that promote oral language is critical at this stage to support the later emergence of reading comprehension (Duursma *et al.*, 2008; Kendeou *et al.*, 2009), particularly for low-income children.

At age 5, most children enter the K–12 schooling system. The research just reviewed suggests the importance of explicit instruction in word decoding. But we cannot ignore language development. This can be an explicit goal in the teaching of science, social studies and literature (Gelman & Brenneman, 2004). Reading aloud to children and fostering discussion about the ideas in books would then supplement intensive instruction in early decoding, implying that expanded instructional time will be required to overcome inequalities that have their source in the early environment (Swanson *et al.*, 2011).

No single study will be sufficient to provide empirical evidence to support this regime. Its success will depend instead on a theoretical framing that collects and interprets the empirical evidence from many interventions and integrates this evidence with the insights of expert practitioners. This theoretical frame includes propositions about how learning is connected across stages of development, propositions about how teachers can manage the inevitable heterogeneity of child skills that teachers encounter in their classrooms and propositions about how policy can mobilize the incentives and resources that enable medical personnel, parents, preschool teachers and K–12 teachers to harmonize their efforts.

https://doi.org/10.1017/bpp.2020.19 Published online by Cambridge University Press

The power of an effective school

My thesis is that what is important for school improvement is not primarily the adoption of specific intact interventions, but rather combining the insights and practices of multiple interventions and integrating these with the understanding of expert practitioners. If this is correct, we would expect to see the most powerful experimental results at large scale when we study the effects of entire schools. And that is what we see. In these studies, more parents apply to have their children attend an innovative school than there are available places in the school. A randomized lottery is then a fair way to make admissions decisions, and this lays the basis for a randomized experimental trial of the effectiveness of the whole school. Neal (2018) reviews these studies. To date, most of the schools so evaluated have been publicly funded charter schools. The research has not shown that charter schools are, on average, more effective than regular public schools, but it has revealed significant heterogeneity in the effects of schools. A number of schools have been found to be particularly effective and capable of producing powerful results under replication, and most of these schools serve urban populations composed of large fractions of low-income and minority children. The KIPP Charter School Management Network has recently demonstrated the capacity to expand its model of schooling successfully (Cohodes *et al.*, 2019). But not all schools found to be effective using randomized methods have been charter schools. Among the most remarkable studies of effective schooling is the large-scale evaluation of Small Schools of Choice (SSC) in New York City. Involving over 100 newly designed regular public high schools, Bloom and Unterman (2014) found that SSC produced significantly and substantially higher graduation rates and higher examination scores on the New York State Regents exam. Taken together, these findings show that powerful schooling has the potential to dramatically reduce social class and racial inequality if replicated on a large scale.

These findings contradict the results of the massively influential 'Coleman Report', which suggested that school differences are not strongly predictive of student outcomes (Coleman *et al.*, 1966). This finding, replicated in every major survey of US schooling in the past half-century, cast a long shadow regarding the hypothesis that school improvement can reduce inequality in educational achievement. The only plausible explanation for the discrepancy between its findings and the results of recent randomized trials of particular schools is that the schools recently studied operate very differently from the schools studied in past surveys. It follows that understanding how these schools operate becomes a critical question for future inquiry.

The most effective new schools have been found in various studies to foster among teachers high expectations for the growth of each student; to administer

https://doi.org/10.1017/bpp.2020.19 Published online by Cambridge University Press

frequent, timely assessments of student skills; to use those assessments to guide instruction with the aim of keeping every student on track for success; and to provide substantial instructional time focused on basic literacy and math skills and extra tutoring for students who have fallen behind (Fryer, 2014). In many of these schools, the aim is to foster a culture where 'no excuses' are tolerated for failure. Hassrick *et al.* (2017, pp. 3–4) wrote that, in these schools, "neither the staff nor the students tolerate the argument that family difficulties or neighborhood disadvantage prohibit success in school." Effective Charter School Management Organizations provide what Cohen and Moffitt (2009) call 'infrastructure for school improvement'. Such infrastructure supports school improvement on a large scale by setting standards for the hiring, evaluation and promotion of teachers, promoting incentives and learning opportunities to school leaders and teachers and guiding the start-up of new schools. These organizational conditions are similar to those found by Bryk *et al.* (2010) to support large-scale school improvement in a large urban district. Hassrick *et al.* (2017) conducted an intensive study of two such schools attended by African-American students on the South Side of Chicago. They showed how the schools found to be remarkably effective in promoting math and literacy in a randomized trial made use of a variety of findings regarding the teaching of reading and math and social services within a culture supported by a system of accountability and incentives that supported a relentless effort to promote the learning of children who came to school with starkly heterogeneous skills.

While research on particularly effective schools is exciting, it poses more questions than it answers. The set of policies and practices that we can observe in such schools are sufficient to produce good outcomes, but we can neither claim that each component of practice is necessary nor that the components taken together are efficient in producing good outcomes. Studying the particular components is essential. Fryer (2014) attempted to isolate particular practices observable in effective schools and sought to import those into existing schools. The results were encouraging, and they illustrate the potential for a virtuous cycle of research in which particular interventions contribute to whole-school practice and studies of whole schools generate hypotheses for the study of particular interventions or combinations of these. This is the spirit of a robust science of school improvement aimed at reducing inequality.

The science of scaling meets the science of school improvement

When I first read the recommendations of Al-Ubaydli *et al.* to improve the science of scaling, I found myself impressed by their comprehensiveness and mostly in agreement, but skeptical about Proposal #1, which says, in

https://doi.org/10.1017/bpp.2020.19 Published online by Cambridge University Press

essence, that we should recommend that a new intervention be adopted only if the data provide strong statistical support against the null hypothesis, particularly in cases where prior support for the policy is weak. This didn't seem quite right to me. Suppose, for example, we have no prior information about the efficacy of a new intervention, but that staying with the status quo is just as costly as enacting the new intervention. Then, given a study of the efficacy of the new intervention, we might decide what to do solely based on the sign of the effect, equivalent to using $\alpha = 0.50$. While obviously contrived, this scenario illustrates the potential value of using decision theory to evaluate policy options. In Bayesian terms, prior belief and new data are not sufficient for a policy decision; one must also impose a loss function (or utility function).

However, if the aim of scale-up is to support a robust science in support of practice, Proposal #1 makes perfect sense We might say, for example, that a proposed causal proposition in support of the science of school improvement should have to meet the same stringent requirements that would be required, say, in the science that underlies medical research.

Indeed, I react enthusiastically to Proposal #1 and all remaining proposals thus reframed. Promoting the pre-registration of experiments, rewarding replication studies, publishing null results, mining multi-site trials for evidence about where, how and for whom an intervention works and designing effectiveness studies with generalization to target populations in mind are crucial to a real science of education. I thank the authors for highlighting the question of general equilibrium and suspect this may be central to understanding the contribution of school improvement to social improvement. Some have argued that increasing educational achievement, while helping participants in a small study, will, if taken to scale, simply intensify competition for scarce jobs as part of a zero-sum game that reshuffles the winners and losers in the labor market. The alternative view, with which I agree, is that expanding human capital will increase the productivity of the entire society while also generating a more thoughtful and enriched civic life. This is a major question in social science that requires interdisciplinary attention.

My essay may seem to have downplayed the perspective of the innovator who seeks to scale up an intervention. However, my aim is not to diminish the importance of the perspective of the innovator, but rather to supplement that perspective with the perspective of users who face the challenge of adopting, adapting or borrowing ideas and practices from experimental studies of interventions and harmonizing those ideas and practices with expert practitioner judgment. As Dewey (1929) suggested, scientific findings can inform but never replace the intelligence of the effective practitioner. In this way, those who wish to advance intervention science in education to support

https://doi.org/10.1017/bpp.2020.19 Published online by Cambridge University Press

better and fairer schooling will benefit from the work of Al-Ubaydli *et al.* and from the current wave of social scientific reform that informs their essay.

References

Bryk, A. S., P. Sebring, E. Allensworth, S. Luppescu and J. Easton (2010), *Organizing schools for improvement: Lessons from Chicago*, Chicago: University of Chicago Press.

Cohen, D. K. and S. L. Moffitt (2009), *The Ordeal of Inequality: Did Federal Regulation Fix the Schools?* Cambridige, MA: Harvard University Press.

Cohodes, S., E. Setren and C. R. Walters (2019), *Can Successful Schools Replicate? Scaling up Boston's Charter School Sector* (No. w25796), National Bureau of Economic Research.

Coleman, J. S., E. Q. Campbell, C. J. Hobson, J. McPartland, A. M. Mood, F. D. Weinfeld and L. Robert (1966), 'York.', *Equality of Educational Opportunity*, (1966).

Cunha, F. and J. Heckman (2007), 'The technology of skill formation', *American Economic Review*, **97**(2): 31–47.

Dewey, J. (1929), *The Sources of a Science of Education*, New York: Horace Liveright.

Duursma, E., M. Augustyn and B. Zuckerman (2008), 'Reading aloud to children: the evidence', *Archives of Disease in Childhood*, **93**(7): 554–557.

Fryer, R. G. Jr (2014), 'Injecting charter school best practices into traditional public schools: Evidence from field experiments', *The Quarterly Journal of Economics*, **129**(3): 1355–1407.

Gelman, R. and K. Brenneman (2004), 'Science learning pathways for young children', *Early Childhood Research Quarterly*, **19**(1): 150–158.

Graf, E., L. Garofalo, A. C. Hundertmark, G. L. Montague, N. M. Polash, E. Suskind and D. L. Suskind (2017), 'Using formative research to develop a hospital-based perinatal public health intervention in the US: The Thirty Million Words Initiative Newborn Parent Education Curriculum', *Journal of Early Hearing Detection and Intervention*, **2**(1): 2–11.

Hassrick, E. M., S. W. Raudenbush and L. Rosen (2017), *The Ambitious Elementary School: Its Conception, Design, and Implications for Educational Equality*, University of Chicago Press.

Kendeou, P., P. Van den Broek, M. J. White and J. S. Lynch (2009), 'Predicting reading comprehension in early elementary school: The independent contributions of oral language and decoding skills', *Journal of Educational Psychology*, **101**(4): 765.

National Reading Panel, National Institute of Child Health, Human Development, National Reading Excellence Initiative, National Institute for Literacy, United States Public Health Service, & United States Department of Health (2000), *Report of the National Reading Panel: Teaching children to read: An evidence-based assessment of the scientific research literature on reading and its implications for reading instruction: Reports of the subgroups*, National Institute of Child Health and Human Development, National Institutes of Health.

Neal, D. A. (2018), *Information, Incentives, and Education Policy*, Harvard University Press.

Spybrook J. (2014), 'Detecting Intervention Effects Across Context: An Examination of the Precision of Cluster Randomized Trials', *The Journal of Experimental Education*, **82**(3): 334–357

Spybrook J., R. Shi and B. Kelcey (2016), 'Progress in the past decade: an examination of the precision of cluster randomized trials funded by the U.S. Institute of Education Sciences', *International Journal of Research & Method in Education*, **39**(3): 255–267

Swanson, E., S. Vaughn, J. Wanzek, Y. Petscher, J. Heckert, C. Cavanaugh, … K. Tackett (2011), 'A synthesis of read-aloud interventions on early reading outcomes among preschool through third graders at risk for reading difficulties', *Journal of learning disabilities*, **44**(3): 258–275.

https://doi.org/10.1017/bpp.2020.19 Published online by Cambridge University Press