Can Using Rigorous Evidence To Guide Federal Education Funds Improve Student Achievement?

Randomized Trials Show Encouraging Initial Results for DoED's Investing in Innovation Fund

An important recent development in evidence-based policy is the federal government's use of a "tiered evidence" approach to allocating funding in grant programs such as the U.S. Department of Education's Investing in Innovation Fund (i3). A central feature of this approach is that the largest grants are awarded to fund large-scale implementation ("scale up") of program models backed by strong scientific evidence of effectiveness. Among the federal tiered-evidence programs, i3's evidence standard for such scale-up grants is particularly rigorous, requiring a demonstration of policy-important effects in scientifically-credible evaluations, with a preference for well-conducted randomized controlled trials.

The key question is whether the government's use of such evidence to select program models for scale up will actually lead to the hoped-for gains in educational achievement. In other words, can the original findings of effectiveness for these models be reproduced as their implementation is expanded to new educational settings and conditions?

Encouraging initial answers to that question are emerging from new randomized trials evaluating three of the four program models awarded scale-up grants in i3's initial year: Reading Recovery, Success for All, and Teach for America.[1] The trials of Reading Recovery and Success for All evaluate these program models as they are being scaled up with i3 funding. The trial of Teach for America, which was commissioned by the Institute of Education Sciences, evaluates this program model as implemented with other (non-i3) funding. In all three cases, the randomized trials evaluate the program models as delivered on a relatively large scale, across multiple school districts and states.

The findings (summarized briefly below) are not yet definitive – e.g., because they are all relatively short-term – but they are consistently positive, and therefore raise confidence that i3's strategy is successfully identifying and funding program models that reliably increase student achievement when implemented on a large scale. The positive results are a notable departure from the usual findings of weak or no positive effects in large randomized trials in education.[2] If the findings hold up in longer-term study reports, they would constitute an important validation of i3's evidence-based approach to scale up.[3]

New Evaluation Findings for Programs Receiving i3 Scale-Up Awards:

Reading Recovery

- **Program description:** Reading Recovery is a program that provides struggling readers in 1st grade with one-on-one tutoring by highly-trained, certified teachers for 30 minutes daily, over a 3-5 month period.

- **Study design:** The program was evaluated in a randomized controlled trial (RCT) of 1,253 1st graders with low reading ability, who were randomly assigned to either a group that received Reading Recovery right away or a wait-list control group that received it one semester later.[4]

- **Key findings:** One semester after random assignment, Reading Recovery was found to have a sizable, statistically-significant effect on reading comprehension: on average, the Reading Recovery group scored at the 39th percentile nationally on the Iowa Test of Basic Skills (ITBS) versus the 19th percentile for the control group (standardized effect size of 0.54).

- **Study quality:** Our review found that the study meets widely-recognized criteria for a well-conducted RCT,[5] with the following limitations that hopefully can be addressed in future research:
  - At the one-semester follow-up, the study had moderately high sample attrition (test scores were collected for 69% of both the Reading Recovery and control groups).
  - The ITBS – a written test – was administered by Reading Recovery trainers and teachers, who then entered the results into Reading Recovery's central database. Preferably, the administration of the test, and handling of the data, would have been done by individuals who were unaffiliated with the program, so as to help ensure impartiality.

Success for All (SFA)

- **Program description:** SFA is a comprehensive school-wide reform program, primarily for high-poverty elementary schools, with a strong emphasis on early detection and prevention of reading problems before they become serious.

- **Study design:** The program was evaluated in an RCT that randomly assigned 37 elementary schools to either a group that implemented SFA or a control group that did not.[6] Outcomes are being tracked for the 2,956 students who were in kindergarten at the start of the study.

- **Key findings:** After one school year (i.e., end of kindergarten), SFA was found to have a positive effect on students' word attack skills – e.g., their ability to sound out words – which equated to moving the average student from the 50th percentile to the 57th percentile in these skills (standardized effect size of 0.18). This effect was statistically significant or very close to it (p=0.03 under the study's main analysis, p=0.06 after adjusting for multiple comparisons). SFA had no significant effect on students' letter-word identification skills, but this is consistent with findings of an earlier RCT of SFA, in which an effect on this outcome only emerged in 1st and 2nd grade.

- **Study quality:** Our review found that the study meets widely-recognized criteria for a well-conducted RCT.[5]

Teach for America (TFA)

- **Program description:** TFA recruits college seniors and college graduates with strong academic records to teach in low-income schools for a minimum of two years, and provides them with brief, intensive training to prepare them for the classroom (e.g., in classroom management, learning theory).

- **Study design:** The program was evaluated in an RCT, which randomly assigned 5,790 middle and high school students in 45 schools to math classes taught by TFA teachers or to similar math classes taught by non-TFA (control group) teachers.[7]

- **Key findings:** After one school year, TFA was found to have a statistically-significant positive effect on students' math scores on state tests, roughly equivalent to an extra 2.6 months of learning during the school year (standardized effect size of 0.07).

- **Study quality:** Our review found that the study meets widely-recognized criteria for a well-conducted RCT.[5]

Conclusion: New randomized trials have produced encouraging initial evidence that i3's scale-up strategy is successfully identifying and funding program models that reliably increase student achievement when implemented on a large scale.

References

[1] The fourth scale-up awardee – the Knowledge is Power Program (KIPP) Leadership Design Fellowship – is also being rigorously evaluated, but findings from that study have not yet been reported.

[2] Coalition for Evidence-Based Policy, *Randomized Controlled Trials Commissioned by the Institute of Education Sciences Since 2002: How Many Found Positive Versus Weak or No Effects*, July 2013, linked here.

[3] We note that i3 has other goals in addition to scaling up program models backed by strong evidence of effectiveness. While such scale up is the focus of i3's highest tier, the other tiers of i3 make smaller grants to fund and rigorously evaluate innovative program models that have moderate evidence of effectiveness (in i3's "validation" tier) or "evidence of promise" (in i3's "developmental" tier). The goal is to strengthen the evidence base for these models and, over time, to build the number that are backed by strong evidence and are therefore worthy candidates for scale up. Because innovation, by its nature, involves trial-and-error, it can be expected that only a subset of the evaluations in these tiers will show positive effects, and that many will not.

[4] May, H., Gray, A., Gillespie, J.N., Sirinides, P., Sam, C., Goldsworthy, H., Armijo, M., and Tognatta, N., *Evaluation of the i3 Scale-up of Reading Recovery: Year One Report, 2011-2012*, Consortium for Policy Research in Education, August 2013, linked here.

[5] The criteria we used to assess whether the trial was well-conducted are summarized in the Top Tier Evidence initiative's RCT Checklist, and include such items as: (i) the program and control groups were similar in their pre-program characteristics; (ii) the study had low sample attrition, and similar attrition rates for the program versus control group; (iii) the study measured outcomes for all individuals assigned to the program group, regardless of whether or how long they participated in the program; (iv) study outcomes were assessed with valid measures; and (v) where appropriate, research staff collecting outcome data were kept unaware of which sample members were in the program versus control group.

[6] Quint, J.C., Balu, R., DeLaurentis, M., Rappaport, S., Smith, T.J., & Zhu, P., *The Success for All Model of School Reform: Early Findings from the Investing in Innovation (i3) Scale-Up*, MDRC, October 2013, linked here.

[7] Clark, M.A., Chiang, H.S., Silva, T., McConnell, S., Sonnenfeld, K., Erbe, A., & Puma, M., *The Effectiveness of Secondary Math Teachers from Teach For America and the Teaching Fellows Programs (NCEE 2013-4015)*, Washington, DC: National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, U.S. Department of Education, September 2013, linked here.