Does a successful randomized experiment lead to successful policy? Project Challenge and what happened in Tennessee after Project STAR

Paul T. von Hippel
LBJ School of Public Affairs
University of Texas, Austin

Chandi Wagner
Office of the State Superintendent of Education, District of Columbia

Did a successful randomized experiment lead to successful policy? Project Challenge and what happened in Tennessee after Project STAR

Abstract

Evidence-based policy is the practice of basing policy decisions on rigorous research evidence, such as randomized experiments. But it is unclear how often evidence-based decisions produce more effective policy. We evaluate an evidence-based policy implemented in 1989-93, after the state of Tennessee completed the famous Project STAR randomized experiment, which showed that reducing average class sizes from 23 to 15 could raise test scores by nearly 0.2 standard deviations (SD). After Project STAR, the state launched Project Challenge, which tried to achieve similar score gains by earmarking $5 million to reduce class sizes in the state's 17 poorest districts.

We evaluate the effects of Project Challenge by applying regression discontinuity and difference in differences analysis to data from district report cards. The results offers no evidence that Project Challenge districts raised test scores, and raise questions about whether districts actually reduced class sizes. After Project Challenge, Tennessee's Basic Education Plan did reduce class sizes, but only by a token amount, from 26 to 25. In this example, it seems that a successful randomized experiment did not lead to successful policy.

Did a successful randomized experiment lead to successful policy? Project Challenge and what happened in Tennessee after Project STAR

Evidence-based policy is the practice of using rigorous empirical evidence to guide policy decisions and evaluate their effects (Campbell, 1969; Haskins & Margolis, 2014; Karlan & Appel, 2012). In its ideal form, evidence-based policy begins with a review of research to identify promising policy ideas. It continues with a limited but rigorous trial, often randomized, of a candidate policy. If the results of the trial are promising, then the policy is rolled out on a larger scale, and evaluation continues.

Although the idea of evidence-based policy holds promise, there are hazards at every step. Interest groups and fiscal constraints may limit the range of policies that are considered. The review of research may be selective. The researchers engaged may have ideological commitments which shape their interpretation of results. The evidence emerging from a trial may be mixed or confusing. If the evidence is clear, it may still have little influence on policy. The policy that is rolled out may differ from the intervention that was tested in the trial. The population and context in which the policy is applied may differ from those in which the trial was conducted.

For these reasons and others, it is not clear how often real evidence-based policy is practiced, or how often it leads to better policy outcomes. Does a successful randomized experiment lead to successful social policy? There are few examples to learn from. The evidence base for evidence-based policy is itself thin.

To shed light on the promise and hazards of evidence-based policy, we turn to an important chapter in its history. The chapter involves Tennessee's handling of school class

size policy between 1984 and 1993. We chose this chapter because the available history (reviewed in Kim, 2006) suggested that the state followed nearly all the principles of evidence-based policy. A key state legislator reviewed the research literature in 1984. From 1985 to 1989 the state funded one of the most famous randomized experiments in policy history—Project STAR—which reduced class sizes in randomly selected K-3 classrooms. Evaluation of STAR concluded that class size reduction had raised reading and math scores by 0.15-0.20 standard deviations (SD) on average, with slightly larger effects for poor children and African Americans. In response, in 1989-93 the state funded a program—Project Challenge—that reduced K-3 class sizes in the poorest school districts. After evaluation of Project Challenge also reached positive conclusions, Tennessee funded statewide class size reduction under its Basic Education Program, which started in 1993. Over the next two decades, at least 23 other states reduced class sizes, and advocates for these reductions often cited evidence from Project STAR.

Unfortunately, the two largest and most expensive state class size reduction policies, in California and Florida, did not produce test score gains comparable to those in Project STAR. In fact, neither produced any gains at all. In California, a 1996 law reduced K-3 class sizes from 30 to 20, but the law had several unintended consequences that squelched its potential to raise achievement or shrink achievement gaps. The surge in demand for teachers forced high-poverty schools to hire novice teachers with emergency credentials, while new vacancies in affluent neighborhoods permitted experienced teachers to transfer out of high-poverty schools (Jepsen & Rivkin, 2009). Many schools lacked enough classrooms for the new teachers, and had to use portable classrooms and staggered schedules under which different teachers and students were in the building during different weeks (Graves, McMullen, &

Rouse, 2013). Although class size reduction, in itself, raised test scores by 0.1 SD, this benefit was canceled by the influx of new and underqualified teachers—especially in the high-poverty schools that class size reduction was supposed to help the most (Jepsen & Rivkin, 2009).

In Florida, a 2002 constitutional amendment reduced class sizes to 18 in grades PK-3, 22 in grades 4-8, and 25 in grades 9-12—at an annual cost of a billion dollars in the early years, and three billion in recent years. These reductions did not increase test scores, either, and it does not appear that inexperienced teachers were the reason (Chingos, 2012; Dieterle, 2015).[1] Evidence from California and Florida is sometimes dismissed because it is not as rigorous as the randomized experiment conducted in Tennessee (Schanzenbach, 2011). However, it is difficult to keep basing policy on a small randomized experiment from thirty years ago when much larger and more recent interventions have not, according to available evidence, produced similar gains (Whitehurst & Chingos, 2011).

California and Florida's experiences may seem disappointing in light of Project STAR, but perhaps we should not be surprised. The external validity of Project STAR was limited. We cannot expect an experiment in Tennessee in the 1980s to predict the effects of interventions in Florida and California in the 1990s and 2000s. California and Florida have different demographics than Tennessee (more Hispanics, for example), and they use different tests. In addition, the class size reductions in Florida and California were not as deep as those tested in Project STAR, though the PK-3 reductions in Florida came close.

About all we can expect a randomized experiment to do is to predict the results of a similar intervention carried out on similar students in approximately the same time and place. On those grounds, we might expect Project STAR to predict the effects of Project Challenge. But did it? Did Project Challenge districts experience test score gains comparable to the gains

in Project STAR? We don't really know because the evaluation carried out at the time of Project Challenge was unconvincing. Indeed, the technique needed to rigorously evaluate Project Challenge—regression discontinuity with not one but two different forcing variables—has only recently been developed.

In this article, we use modern techniques to analyze old data from Project Challenge. Unfortunately, we find that it did not produce the benefits that Project STAR led the state government to expect. Not only did Project Challenge fail to increase test scores in affected districts, we find that it did not even reduce class sizes. While it appears the state did send the promised money to affected districts, it appears that they did not hire the new teachers needed to reduce class size. In addition, it seems that the Basic Education Program, a statewide school budget that followed Project Challenge, only reduced class size by a token amount, from 26 to 25. In short, it seems that, despite having rigorous evidence from Project STAR, later state policy did not reduce class size by the amount necessary to increase student achievement. As an exemplar of evidence-based policy, Tennessee's response to Project STAR was a disappointment.

History

Before presenting our results in detail, let's review the history of Tennessee's class size policy in the 1980s and early 1990s.

The Better Schools Project

The idea of reducing class sizes did not simply leap onto Tennessee's political agenda from the pages of the research literature. Class size reduction was a longstanding goal of

teachers' unions. In the late 1960s, union leader Albert Shanker had settled on class size reduction as a priority that "could be seen both as an education policy issue and as a working condition" (Kahlenberg, 2007). Shanker saw the interests of children and teachers as aligned: he believed smaller classes would benefit children, and he knew they would make teachers' jobs easier.

In Tennessee, the principal lobbyist for class size reduction was Professor Helen Pate Bain of Tennessee State University—a former president of the National Education Association, the nation's largest teacher's union, and an ally of that union's state affiliate, the Tennessee Education Association. Bain believed that the ideal class size was ten (Ritter & Boruch, 1999), less than half the average K-3 class size, which at the time was 22.5, in Tennessee.

Bain and the teachers' union had allies among liberal Democrats in the Tennessee legislature, but Republicans and fiscally conservative Democrats viewed class size reduction, which required hiring more teachers, as too expensive. The idea might have died in gridlock, except that Republican Governor Lamar Alexander (later US Secretary of Education, now a US Senator) needed union support for his Better Schools Program, an education reform package whose centerpiece was merit pay. Unions generally oppose merit pay, which they fear will increase managerial power reduce union solidarity by treating workers unequally. But the Tennessee Education Association was willing to accept Alexander's merit pay in exchange for an across-the-board 10 percent pay raise and class size reduction (Ritter & Boruch, 1999).

When a compromise version of the Better Schools Program passed in 1984, it included a statewide pay raise but did not include statewide class size reduction. The Plan did, however, include funding for a study of class size reduction at a single elementary school in a suburb of Nashville. The principal investigator for that study was Helen Bain.

A key legislator in the debate over the Better Schools bill was state Representative Steve Cobb, PhD, a liberal Democrat who served on the state House Ways and Means Committee. A former professor who had worked with Helen Bain at Tennessee State University, Cobb "did the unthinkable," according to evaluator John Foster: "he read the literature" on class size (interview with John Forster, quoted in Ritter & Boruch, 1999). This may be an overstatement, since the literature at the time included at least two reviews which might have fostered skepticism about class size. One was a narrative review by the Educational Research Service (ERS, 1978), which concluded that research on class size was "contradictory and inconclusive." Another was a vote count study by Hanushek (1981), titled "Throwing Money at Schools," which reviewed estimates from 29 studies on the effects of various school resources, including class size (proxied by the teacher-student ratio). Hanushek reported that the vast majority of estimates were statistically insignificant, and about as likely to be negative as positive. The histories of Tennessee's class size debate do not indicate whether Cobb and Bain read these skeptical reviews, or what they thought of them.

The histories are clear, however (Boyd-Zaharias, 1999; Ritter & Boruch, 1999), that both Cobb and Bain read a meta-analysis by Glass and Smith, which was one of the first meta-analyses ever performed (Glass, 1982; Glass & Smith, 1979). Glass and Smith acknowledged that past research summaries had produced mixed and confusing results, but they attributed the confusion to literature reviews that were incomplete or selective, as well as a failure to distinguish studies according to rigor and power. They also alleged that some review authors had conflicts of interest: for example, the Educational Research Service, despite its official sounding name, was actually an individual named P.J. Porwell, whose self-published research was financed by "the American Association of School Administrators, the Council of Chief State

School Officers, and several other professional administration groups" seeking to undermine teachers' unions and their demands for smaller classes (Glass and Smith, 1979). The ERS published a criticism of Glass and Smith's meta-analysis, which Glass rebutted in a terse reply subtitled "No Points Conceded" (ERS, 1980; Glass, 1980).

Glass and Smith summarized 725 effect estimates from 77 class size studies—14 of which were randomized—conducted between 1900 and 1979. Like previous authors, they found that estimates varied in size and significance—but by applying meta-analysis, they could explain much of the variation. In particular, studies tended to have larger, more significant effect estimates if they more rigorously controlled confounding variables through matching or random assignment, if students were assigned to large and small classes that differed substantially in size, and if students were exposed to their assigned class size for at least 120 days (more than a semester).

Glass (1982) summarized effect sizes with the log-linear model $E(Z_S - Z_L) = \beta \ln(n_L/n_S)$, where n_L/n_s was the ratio between the number of students in large and small classes in a particular study, and $E(Z_S - Z_L)$ is the expected difference between the standard scores of students in large and small classes. The coefficient β was estimated to be 0.23 for the 14 randomized studies, and 0.45 for the 5 randomized studies that lasted more than a semester. According to the latter estimates, a policy that reduced elementary class sizes from 22.5 (the Tennessee average) to 20 would be expected to raise test scores by only 0.05 SD. A more substantial reduction, from 22.5 to 15, would be expected to raise test scores by 0.18 SD.[2]

Glass and Smith's meta-regression convinced Cobb that class size reduction had potential benefits. But it also convinced him those benefits would be negligible unless class

sizes were reduced well below 20—and reductions that deep would be expensive. Cobb's ambivalence made him an unreliable advocate for the Better School Program's class size amendment, which would have reduced class sizes only slightly, leaving them over 20. That is one reason he let the amendment die.

Project STAR

In 1985, the Tennessee legislature engaged the question of class size anew. Again, the policy catalysts mixed politics and evidence. New evidence came from Bain's class size reduction study and the early results of Prime Time, a class size policy in Indiana. In retrospect, neither of these studies contributes much to the evidence on class size. Bain's study involved just a single school in suburban Nashville, and Prime Time was a nonrandomized policy whose evaluation, which was never peer reviewed, is difficult even to find today (Gilman, Swan, & Stone, 1988). Yet both studies carried some weight with Tennessee legislators, or at least served as fodder for those who already favored class size reduction (Ritter & Boruch, 1999).

Perhaps the more important catalyst for the 1985 class size debate was a political calculation. Bain convinced the Tennessee teachers' union to settle for reducing class sizes in grades K-3 only (Ritter & Boruch, 1999). The union disliked favoring teachers at certain grade levels, which was contrary to union solidarity, and would have preferred to reduce class sizes in all grades. Yet if class sizes needed to be cut to 15—as Bain, relying on Glass' reviews, believed (Boyd-Zaharias, 1999)—the only way to reduce costs to an acceptable level was to limit the reform to certain grades.

Bain's decision to focus on grades K-3 was not evidence-based. Glass' meta-analysis had concluded that the effects of class size reduction were greatest in high school, not elementary school (Glass & Smith, 1979, Figure 3). But high school teachers have higher salaries, and reducing their class sizes below 20 would have been more expensive. So although Bain argued that class size reductions would have their greatest benefit in the early grades (Ritter & Boruch, 1999), this argument was not based on the research evidence. Bain proceeded from either personal conviction or political and financial expediency.

In May 1985, Tennessee's legislature passed, and Governor Alexander signed, legislation authorizing the study that became Project STAR. The reason for the study was not simply the state's thirst for policy-relevant knowledge. Instead, the bill that passed was a compromise between proponents like Bain and Cobb, who wanted class sizes reduced statewide, and more fiscally conservative legislators who wanted to minimize costs (Ritter & Boruch, 1999).

The design and implementation of Project STAR were exemplary. It was a block-randomized experiment, designed and carried out by a consortium that included representatives from state government and five universities. Within each of 79 schools, the experiment randomly assigned kindergarten student and teachers to one of three experimental conditions: a regular-sized class with 22-25 students, a regular-sized class with an aide, and a small class with 13-17 students.[3] Fidelity and compliance were excellent; 99.7 percent of kindergarteners enrolled in the condition to which they were assigned, class sizes were usually close to the target ranges, and student characteristics were well-balanced across experimental conditions, as random assignment would predict (Krueger, 1999; Schanzenbach, 2006). Small class sizes were maintained for four years, from kindergarten in 1985-86

through third grade in 1988-89. Later, a followup called the Lasting Benefits Study continued into higher grades (Achilles & Others, 1993), and a more recent followup continued into young adulthood (Chetty et al., 2010).

Annual reports on Project STAR's results were provided to the Tennessee Department of Education starting in 1987, and in 1989 a special issue of the *Peabody Journal of Education* was devoted to it (Various, 1989). The original evaluators shared the data with other scholars from 1989 onward (e.g., Folger & Breda, 1989; Krueger, 1999), and in 2007 they released the data publicly with meticulously detailed documentation (Finn, Boyd-Zaharias, Fish, & Gerber, 2007).

The basic results of Project STAR are well-known. Teachers' aides had no effect, and small classes had the effect of increasing student math and reading scores by an average of 0.15-0.20 standard deviations (SD), with larger effects on poor children and African Americans.[4] The size of the effect was approximately what would have been predicted by Glass and Smith's summary of longer randomized trials. The full effect size was evident by the end of kindergarten, and persisted but did not grow in 1st-3rd grade. Later followups would find that three-quarters of the effect on test scores disappeared when small classes were discontinued in 4th grade (Schanzenbach, 2006). In adulthood, students assigned to small classes were a little more likely to attend college, but did not earn higher incomes (Chetty et al., 2010). However, evidence on long-term outcomes was not yet available when the state took its next step.

Project Challenge

Thanks to Project STAR, by 1989 Tennessee had rigorous evidence about the potential effect of small classes on young Tennessee schoolchildren. In addition, the state—now led by Democratic Governor Ned McWherter, who had been Speaker of the House under Governor Alexander—was under pressure to help rural districts, which in 1988 had filed a lawsuit, *Small School Systems v. McWherter*, claiming that their low levels of state funding violated the state constitution (Martin, 1993).

Out of this environment came a new class size policy targeting poor rural districts. Called Project Challenge by evaluators, the policy was one aspect of the statewide *21st Century Challenge*, whose goal was to increase Tennessee's economic competitiveness by improving children's academic performance. Other parts of the 21st Century Challenge included raising curricular standards, adopting a new state test—the Tennessee Comprehensive Assessment Program (TCAP)—that would reflect the new standards, and measuring teacher effectiveness with the nation's first statewide value-added system (Smith, 1990).

In 1989-90, Project Challenge provided the state's 17 poorest districts[5] with $2.8 million in state funds to reduce average K-3 class sizes to 15, which was the level that was tested in Project STAR. Districts were expected to supplement the state funds with $1.7 million from the federal "Chapter 1" program (today called Title I) (Smith, 1991). It is not clear from program descriptions whether districts would receive new Chapter 1 grants to pay for class size reduction, or were simply expected to repurpose Chapter 1 funds that they were already receiving. Naturally new funds would have make class size reduction more practical, while the suggestion to repurpose existing funds would have had the whiff of an unfunded

mandate. Project Challenge continued for at least 3 years, through 1991-92, after which the next policy phase began with Tennessee's Basic Education Program.

Eligibility for Project Challenge was determined by county per capita income and the percentage of district students applying for free and reduced price lunch. Every district poor enough to qualify was rural, although being rural was not an explicit criterion for eligibility. The focus on poor rural districts was politically expedient given the claims of the *Small School Systems* lawsuit, and could be considered evidence-based because Project STAR had found that small classes have more impact on poor students. On the other hand, Project STAR did not find that small classes had more impact in rural areas, and it found that they had more impact on black students, who were rare in the rural, mostly east Tennessee districts that participated in Project Challenge.[6]

Unlike Project STAR, Project Challenge has received next to no attention from scholars. Two evaluations were written for the state by a team that included three of the Project STAR evaluators (Nye et al., 1995; Nye, Achilles, Boyd-Zaharias, Fulton, & Wallenhorst, 1992), but there were no peer reviewed journal articles, and the evaluation of Project Challenge was not nearly as rigorous as the evaluations of Project STAR. The data from Project Challenge were not archived, and as far as we know, no outside scholars have ever requested the data or attempted to re-estimate the program's effects—until now.

Despite its neglect, Project Challenge is an important program because it speaks not just to the class size debate but to the question of how effectively state and local governments convert rigorous evidence into effective policy. Did the state and participating districts succeed in reducing K-3 class sizes to 15? Did those reductions have effects commensurate with those observed in Project STAR? We address those questions in the analyses below.

Data

Project Challenge began almost 30 years ago, and data from the original evaluation are no longer available. We contacted living members of the evaluation team, and one of them shared a few SPSS files, but the files were undocumented and represented only a small fragment of the data that would be needed to replicate the original evaluation or conduct a new one. We also contacted Tennessee State University, Vanderbilt University, the University of Tennessee in Knoxville, the Tennessee Department of Education, and the SAS Institute, which has conducted contract education research for Tennessee since the 1990s. None of these institutions had data from the Project Challenge evaluation, or any data from the Project Challenge era.

We did, however, find data in paper "report cards" that the Tennessee Department of Education published annually for each of the 134 school "systems," or districts, in Tennessee (Tennessee Department of Education, 1990, 1992a, 1992b, 1993). Although data at a lower level of aggregation would be valuable, district-level data is adequate to evaluate Project Challenge because Project Challenge was a district-level intervention.

Report cards were published for the first three years of Project Challenge (1989-90, 90-91, and 91-92) as well as the year before (1988-89). Some report cards included data from previous years as well as the current year, and this expanded the number of years with data available for analysis.

We scanned the paper report cards and transcribed variables which we imported into Stata for analysis. Selected variables and years were transcribed twice, independently, by

different transcribers. A comparison of the two transcriptions revealed (and corrected) less than one error per variable per year.

Program participation and eligibility

We flagged each district according to whether it participated in Project Challenge. A list of Project Challenge districts was given in the original evaluations (Achilles, Zaharias, Nye, & Fulton, 1995; Nye et al., 1992). According to the evaluations, 17 districts started Project Challenge 1989-89, and one, Van Buren County, later dropped out. We also dropped two nonparticipating districts that had missing values in one or more years; neither of these counties was close to the eligibility threshold, so dropping them presumably had no practical effect on the results.

For each district and year, the report cards give the two variables whose values in 1989-90 were used to determine Project Challenge eligibility:

- the districts' percentage of students qualifying for free or reduced price lunch, and
- the per capita income of the county in which the district was located.

In Tennessee, most districts and counties are geographically identical, so county income is the same as district income. A few counties contain more than one district—one district for a city and one for the surrounding area—but these are in metropolitan areas, while all Project Challenge districts were rural.

Program implementation

A crucial question is whether participating districts actually reduced class sizes, and whether they received adequate funding to do so. District report cards contain variables that shed light on these questions:

District report cards contain total spending per student in every year of Project Challenge, as well as the year before. In 1991-92, district report cards also contain the percentage of spending that came from state, federal, and local sources. From these we can calculate state spending per student and zoom in to the question of whether state or federal spending increased in Project Challenge Districts.

District report cards give the number of teachers and students in each district. Report cards do not give counts for individual grades, but they do give counts for grades K-3 combined, which are the grade levels targeted by Project Challenge (and earlier targeted by Project STAR). Dividing the number of students by the number of teachers gives the K-3 student/teacher ratio.

Student/teacher ratio is not always the same as class size, but the terms were often used interchangeably in Tennessee. For example, the acronym STAR stands for Student Teacher Achievement Ratio. Most of the ways that student/teacher ratio can differ from class size are not relevant to Project Challenge. In some data, the student/teacher ratio can underestimate class size by counting teachers of art, music, physical education, and vocational skills—but the Tennessee report cards count those teachers separately, and our student/teacher ratio is limited to "regular" classroom teachers. In middle and high school, the student/teacher ratio can differ from class size because teachers specialize in different

subjects and see several classes per day. But in elementary school, students spend the bulk of the day in a single classroom, and the number of classes taught per "regular" teacher is one. The exception to this occurs in kindergarten, where some teachers may have two half-day classes—one that meets in the morning and one that meets in the afternoon. In schools with half-day kindergarten, the K-3 student/teacher ratio exceeds average class size by a factor of 5/4.[7] Even then, a school could not reduce its student/teacher ratio without reducing class size, unless it switched from half- to full-day kindergarten at the same time.

Test scores

The next question is whether participation in Project Challenge improved academic performance. We can answer this using test scores in the report cards.

At the time of Project Challenge, Tennessee tested two of the four grades whose class sizes were reduced by Project Challenge. The state did not test grades K and 1, but it did test grades 2 and 3 (as well as 4 through 10). As the original evaluators pointed out, students' exposure to small classes increased by grade and year. After year 1 of Project Challenge, 2nd and 3rd graders had one year of small-class exposure; after year 2, they had two years' exposure, and after year 3, they had three years exposure. After year 4, 2nd graders had three years exposure (K-2), and 3rd graders had four (K-3)—but by then Tennessee was changing class sizes again with the Basic Education Program.

Unfortunately, the state changed tests in the year that Project Challenge began. This makes impact evaluation more challenging, though not impossible. In the years before Project Challenge, Tennessee students took the Stanford Achievement Test, version 7 (SAT-7) in grade 2 and the Basic Skills First Achievement Tests (BSF) in reading and math in grade 3. But

during Project Challenge, students took the Tennessee Comprehensive Assessment Program (TCAP) tests, which they still take today. There are TCAP tests in reading, language arts, math, science, and social studies, as well as a TCAP battery score that combines all 5 subjects.

Normalization of TCAP scores

Report cards summarized TCAP scores in a variety of formats, which we converted to the mean or median of a student-level *Z* score.

District TCAP battery scores were summarized by the national percentile *P* that corresponded to the districts' median score. We converted the percentile to a median *Z* score m_Z by the inverse standard normal transformation: $m_Z = \Phi^{-1}(P/100)$.

Subject TCAP scores were summarized by the percentages of students who scored in three different ranges on a nationally normalized stanine scale *S* with a mean of 5 and an SD of 2. We converted these stanines to a standard normal *Z* score with a mean of 0 and an SD of 1. The report card gave the percentage P_1 who scored "below average" ($S<3.5$, $Z<-.75$), the percentage P_2 who scored near average ($3.5 \leq S \leq 6.5$, $-.75 \leq Z \leq .75$), and the percentage P_3 who scored "above average" ($S>6.5$, $Z>.75$). Then by properties of the normal distribution,[8] the mean of student *Z* scores in the district is

$$\bar{Z} = .75 - s_Z \Phi^{-1}\left(1 - \frac{P_3}{100}\right)$$

where

$$s_Z = \frac{1.5}{\Phi^{-1}\left(1 - \frac{P_3}{100}\right) - \Phi^{-1}\left(\frac{P_1}{100}\right)}$$

is the SD for student *Z* scores within the district.

BSF and SAT-7 scores

Before Project Challenge, BSF and SAT-7 scores were reported in a less useful format.

For the SAT-7, report cards gave the mean stanine score $\bar{S}$ but rounded it to one digit, making it so coarse as to be practically useless. When rounded to one digit, nearly all districts have a mean stanine score of 5 or 6.

For the BSF, report cards gave the percentage of questions answered correctly. This cannot be converted to the mean of the student level Z scores, but it has two digits so it can be used to rank districts. District rankings will be useful in some analyses.

Other potential outcomes

We considered as possible outcomes some other variables that appear in district report cards for some year, including student attendance and promotion rates. We decided against including these outcomes because they were aggregated across grades K-12 and it would be impossible to isolate the effect of Project Challenge on the grades that it targeted (K-3).

Years of data

Ideally, the report cards would provide every variable for every year of Project Challenge and at least one year before. Instead, some variables are missing from the report cards for some years. The availability of data by year is summarized in Table 1.

Variables that we have both before and during Project Challenge can be analyzed longitudinally in a panel regression with district fixed effects. Variables that we only have

during Project Challenge can be analyzed cross-sectionally in a regression discontinuity design.

The Basic Education Program

Project Challenge lasted (at least) four years, but we restrict our evaluation to years 1-3, since in year 4 (1992-93) the state launched the Basic Education Program, which among other things reduced class sizes in every district and not just districts that participated in Project Challenge. District report cards from 1991-92 indicate how much new state money each district would receive under the Basic Education Program and what it had budgeted that money for, including new teachers. We summarize the new state funding received and the new teachers hired statewide.

Methods

The original evaluators of Project Challenge were understandably tentative in their choice of an evaluation design. They recognized that Project Challenge was harder to evaluate than Project STAR. Unlike Project STAR, Project Challenge was not randomized, and there was no obvious comparison group since the districts that participated in Project Challenge were by design poorer than the districts that did not. Data on class size was lacking, and the state had changed tests in the year that Project Challenge began, making it difficult to compare test scores before and after intervention began (Achilles et al., 1995; Nye et al., 1992, 1995).

Regression with district fixed effects

Although the original evaluators recognized these challenges, they were nonetheless charged with estimating the impact of Project Challenge. To do so, they converted average district test scores to ranks and asked whether Project Challenge districts ranked higher after intervention than before. The evaluators reported that the average ranks of Project Challenge districts climbed from below-average before intervention to average or-above-average afterward., and concluded that the intervention had raised test scores (Achilles et al., 1995; Nye et al., 1992, 1995).

There are some problems with using ranks in this way. Scores on different tests are not interchangeable, and even without intervention, we would not expect districts to have the same ranks on one test as on another. In addition, Project Challenge was applied selectively to poor districts, which tend to be low-scoring, so there is a danger that participating districts' scores would improve, relative to other districts, simply because of regression toward the mean. Regression toward the mean is a particular danger for ranks, since ranks have a "floor," or minimum possible value, of 132. For districts ranked near the floor, there is hardly anywhere to go but up, and there is a danger of spuriously attributing such improvements to Project Challenge.

Despite our reservations, we analyzed change in district ranks simply to see whether we could replicate the original investigators' results and understand their optimistic conclusions. While the evaluators approached their rank analysis descriptively, without statistical inferences, we recast their analysis more formally as a difference-in-difference

analysis applied to ranks or, equivalently, as a rank regression with district fixed effects. Our model was

$$R_{dt} = \alpha_d + \beta\ Challenge_{dt} + e_{dt}$$

where R_{dt} was the rank of district *d* in year *t*, and $Challenge_{dt}$ was a dummy variable indicating whether district *d* participated in Project Challenge in year *t*. Setting reservations about the model aside, the coefficient β represented the average effect of participation in Project Challenge on a district's test score rank. α_d was a district fixed effect, and e_{dt} was a random residual. There is no need for a year fixed effect because by definition the average rank must remain constant across years.[9] We clustered standard errors at the district level (Cameron & Miller, 2015).

We fit the model to 3rd grade reading and math scores, which were the only scores that were available to us for at least one year before and after the start of Project Challenge. We used BSF scores before Project Challenge and TCAP scores afterward. The original evaluators' analysis was similar, except that they used 2nd grade scores and used the SAT-7 before project Challenge. In our data, the SAT-7 scores were not usable because district means were rounded to a single digit.

We used a similar model for spending per student. We did not convert spending to ranks and we did include year fixed effects since average spending levels rise over time. Regression toward the mean remained a concern since the districts that were eligible for Project Challenge had lower-than-average funding levels before intervention.

We could not apply a similar longitudinal model to student/teacher ratios, since we did not have data on student/teacher ratios until after the start of Project Challenge.

Regression discontinuity

Today evaluators have a tool—regression discontinuity (RD) analysis—that is suited to evaluating programs like Project Challenge, which assign treatment to districts that are just above some threshold for poverty. RD fits a smooth curve to the relationship between poverty and some dependent variable (test scores, funding, student/teacher ratio), and looks for a discontinuity or jump in the curve at the point where district poverty levels cross the threshold between eligibility and ineligibility. The size of the discontinuity estimates the effect on Project Challenge on districts that were near the eligibility threshold.

Unlike a longitudinal analysis, RD does not require variables to be available before intervention, though it can use those variables as covariates if they are available. This is a major advantage in our Project Challenge data, where one variable (student/teacher ratio) is not available before intervention, and another variable (test scores) is not measured the same way before the intervention as during.

The original evaluation of Project Challenge did not use RD because in the early 1990s the technique was not in evaluators' toolbox. Although RD was first proposed by psychologists in 1960 (Thistlethwaite & Campbell, 1960), after 1972 the technique went into a dark age of disuse before being revived, refined, and extended by economists starting around 1995 (Cook, 2008). Project Challenge and its evaluation took place during the dark age of RD.

RD requires a *forcing variable* that determines program eligibility, so that districts are eligible only if the forcing variable is above (or below) some threshold. In the case of Project Challenge there were two forcing variables. The first forcing variable was county per capita income PCI_d, and the second forcing variable was the district's percentage of students FRL_d

who were eligible for free or reduced price lunch. Districts were eligible if PCI_d was below some threshold τ_{1d} and FRL_d was above some threshold τ_{2d}. Eligibility was determined by the value of the forcing variables in 1989-90.[10]

There is a small literature on how to conduct an RD when there are two forcing variables (Papay, Willett, & Murnane, 2011; Porter, Reardon, Unlu, Bloom, & Robinson-Cimpian, 2014; Wong, Steiner, & Cook, 2012). This literature is recent and did not exist at the time of the original Project Challenge evaluation. So even if the evaluators had been aware of RD, they would not have known how to apply it to Project Challenge.

Some techniques for RD with two forcing variables require estimating the discontinuity around the threshold of each variable separately.[11] These techniques are not viable in Project Challenge since they require a large number of eligible and ineligible districts with a high density of points around each threshold. Our Project Challenge data, by contrast, have only 16 or 17 eligible districts out of a total of 132. The data are somewhat sparse near the eligibility thresholds, and the forcing variables are strongly correlated, so that a district which is eligible on one forcing variable will typically be eligible on the other as well.

Under these circumstances, the only viable approach is to combine the two forcing variables into a single *binding score* (Porter et al., 2014),[12] which we call district *poverty*:

$$Poverty_d = \min(-std(PCI_d), std(FRL_d))$$

where each forcing variable has been centered around its eligibility threshold and then standardized with respect to its SD:[13]

$$std(PCI_d) = \frac{(PCI_d - \tau_{1d})}{SD_{PCI}}$$

$$std(FRL_d) = \frac{(FRL_d - \tau_{1d})}{SD_{FRL}}$$

Then we can analyze the effect of Project Challenge as an ordinary RD with $Poverty_d$ as the only forcing variable. Districts were eligible for Project Challenge if $Poverty_d > 0$ and ineligible otherwise.

For a single year of data a sharp RD model is

$$Y_d = f(Poverty_d) + \beta\ Challenge_d + e_d \qquad (1)$$

Here $f(Poverty_d)$ is a smooth continuous function of the forcing poverty variable measured in 1988-89. $Challenge_d$ is a dummy indicating whether the district participated in Project Challenge, and β estimates the effect of participation. e_d is a random residual.

The outcome Y_d can be a district's average test scores, or one of the resources that was supposed to improve under Project Challenge: total spending per student, state spending per student, or the student/teacher ratio. We can include pre-program spending and test scores as covariates. (The fact that a different test was used before the program is not necessarily a problem.) We cannot use pre-program student/teacher ratio as a covariate, because student/teacher ratio was not reported before Project Challenge began.

The sharp RD model in (1) assumes that all eligible districts participated in Project Challenge. This is not quite true, so instead we fit a fuzzy RD model that uses eligibility as an instrument for participation. The model has two stages or equations:

$$P(Participate_d) = g(Poverty_d) + \gamma\ Eligible_d + u_d \qquad (2a)$$

$$Y_d = f(Poverty_d) + \beta\ \hat{P}(Participate_d) + e_d \qquad (2b)$$

Equation (2a) estimates the probability of participation as a smooth function g() of district poverty, with a discontinuity γ at the eligibility threshold. Equation (2b) estimates the effect β

of participation on the outcome Y_d, modeled in terms of a smooth function *f()* of poverty and the probability of participation, as estimated from equation (2a).

The smooth functions *f()* and *g()* are estimated by local linear regression. In large samples, the optimal bandwidth for the local regression is the *IK bandwidth* (Imbens & Kalyanaraman, 2012). In small samples like ours, though, the optimal bandwidth is unknown and the IK bandwidth may be too narrow and risk overfitting. We experimented and obtained a relatively smooth fitted curve with a small standard error by using twice the IK bandwidth and omitting one outlier. Using the IK bandwidth and including the outlier yielded a more jagged curve and a larger standard error, but did not change our substantive conclusions.

One assumption of RD is that districts cannot deliberately change which side of the threshold they are on. This assumption is plausible here because school districts cannot change their county incomes or the percentage of students qualifying for free and reduced price lunch. Even if districts could misreport how many students qualified for free and reduced price lunch, they couldn't engineer eligibility for Project Challenge since the lunch numbers were already reported before the eligibility criteria for Project Challenge were decided.

Another assumption of RD is that nothing besides program eligibility changes at the threshold. That assumption too is plausible here. It is true that during Project Challenge, small rural districts were suing the state for more funding in the *Small School Systems* lawsuit filed in 1988. But that lawsuit did not change school funding until a state Supreme Court decision in 1992 triggered the state to launch the Basic Education Program of 1992-93. Even if the state had tried to preemptively increase funding to small districts before the lawsuit was resolved, there were 66 to 78 districts in the *Small Systems* suit, vs. just 17 that Participated in

Project Challenge. So the poverty threshold to benefit from the *Small Systems* lawsuit was clearly different than the threshold to benefit from Project Challenge.

Results

Description of eligible districts

Figure 1 describes the 17 Tennessee school districts that participated in Project Challenge. All were rural, largely white, and poor. All but two were in east Tennessee, and all were small, with average K-3 enrollments of just 790, or a total of 13,432 K-3 students across the 17 eligible districts in 1991-92.

Eligibility was determined by thresholds for county per capita income and the percentage of students qualifying for federal lunch subsidies when Project Challenge launched in 1989-90. Although the eligibility thresholds are not specified in extant policy or evaluation documents, they are clear from Figure 1. All but one of the participating districts (and just one nonparticipating district) had per capita incomes under $9,700 and at least 47 percent[14] of students qualifying for federal lunch subsidies. The only participating district that did not meet both eligibility criteria was Van Buren County, which met one criterion and only participated in some years. We coded Van Buren's participation as 0.5, and coded other districts' participation as 0 or 1.

We combined district income and lunch subsidy into a binding score called *poverty*. (See the Methods section for details.) Figure 2 shows how the poverty score discontinuously predicts district participation in Project Challenge. At the poverty threshold of zero, participation in Project Challenge jumped from practically zero to nearly 90 percent.

Spending trends

The first thing to verify is that money was spent on Project Challenge. According to a contemporary account by Tennessee's Commissioner of Education (Smith, 1990), in the first year of Project Challenge (1989-90) participating districts received $4.7 million—$2.8 million in state funds and $1.9 million in federal Chapter 1 funds—to reduce class sizes in grades K-3. If these funds were new, and not simply reallocated from another purpose, they would amount to an increase of $361 per K-3 student in participating districts. Although the money was intended for grades K-3, district reports of per-student spending are aggregated across all students in grades K-12, and on that basis, we would expect Project Challenge to increase reported spending by $111 per student in participating districts.

The spending data in district report cards match this account. According to Table 1, in the first year of Project Challenge (1989-90) participating districts increased their per-student spending by $103 more than other districts. During Project Challenge, participating districts continued to spend less per student than other districts, but the average difference was $103 smaller in the first year of Project Challenge than it was the previous year. Across the four years from 1989-90 to 1991-92, participation in Project Challenge was associated with a statistically significant spending increase of $96 per student ($p$=.003), according to a panel regression with year and district fixed effects.

Student/teacher ratio

The next question is whether districts used the money, as intended, to reduce K-3 class sizes as the program intended. According to a contemporary account by Tennessee's

Commissioner of Education (Smith, 1990), Project Challenge districts reduced K-3 class sizes to 15, on average, from a previous class size of 25 to 28.

Our results contradict this account. According to Table 3, the average student/teacher ratio in Project Challenge districts was 24 in year 2 of the program (1990-91) and 26 in year 3 (1991-92)—higher than the average ratio of 22 in nonparticipating districts.

These ratios contradict the claim that participating districts had K-3 class sizes of 15. Although class size is not always the same as the student/teacher ratio, these is no way to make a K-3 student/teacher ratio of 24 to 26 equivalent to a class size of 15. The teacher counts in Table 3 are limited to "regular" classroom teachers and exclude teachers of special education, art, music, physical education, and vocational skills. In grades 1-3, regular teachers typically have one class each. In Kindergarten, teachers can have two classes, one in the morning and one in the afternoon, but even if every Kindergarten teacher in Project Challenge districts taught two classes (which is unlikely), a K-3 student/teacher ratio of 24 to 26 would equate to an average class size of 20—not 15.

Even if Project Challenge districts did not reach their target class size, it could be that their class sizes were smaller during Project Challenge than they were before. We cannot assess this possibility directly, because district report cards did not report student/teacher ratios until after Project Challenge began. However, if Project Challenge did reduce student/teacher ratios, we would see evidence of that in a regression discontinuity plot. And we do not.

Figure 3 displays discontinuity plots of the student/teacher ratio as a function of district poverty, during years 2 and 3 of Project Challenge. If Project Challenge districts reduced their class sizes, we would expect to see a downward discontinuity at the poverty

threshold of zero. But we do not. Instead, we see an *upward* discontinuity, indicating that districts that were barely eligible for Project Challenge had *higher* student/teacher ratios than districts that were barely ineligible. Near the eligibility threshold, the average student/teacher ratio in participating districts exceeded the ratio in ineligible districts by 1.6 in 1990-91 and by 2.0 in 1991-92. The latter difference is significant at $p<.10$, though the former is not.

Was the budget of Project Challenge enough to fund the intended reductions in class size? By our calculations, it could have come close. According to the 1992-93 district report cards, the average cost of a new teacher in Project Challenge districts was about $21,800, so that the 1989-90 Project Challenge budget of $4.7 million would have been enough to hire 216 new teachers. And adding 216 to the teacher counts for Project Challenge districts in Table 3 would reduce the K-3 student/teacher ratio from 24 to 17 in 1990-91 and from 26 to 18 in 1991-92.[15] If some kindergarten teachers teaching two classes, a student/teacher ratio of 17 or 18 could correspond to an average class size of 15.

Effect on test scores

If Project Challenge districts did not in fact reduce class sizes, we would not necessarily expect them to improve test scores. As we will show, they did not.

The effect of Project Challenge on test scores is somewhat difficult to estimate, because the state gave a different test during Project Challenge (when it gave the TCAP) than it did in the year before (when it gave the BSF and the SAT-7). But there are a couple of potential approaches.

Analysis of ranks

The approach taken by the original evaluators was to convert district average test scores, before and during Project Challenge, to a rank. They reported that after Project Challenge began, the ranks of participating districts improved in both math and reading. The validity of this result is questionable, though, because different tests were used before and after Project Challenge began.

Setting the issue of validity aside, our data cannot reproduce the finding that the rankings of participating districts rose during Project Challenge. To the contrary, in 3rd grade reading, we find that the average rank of Project Challenge districts actually fell from 72.5 (out of 132) in the year before Project Challenge (1989-90) to 86.2 by year 3 of the program (1991-92)—a statistically significant drop (p<.03). In 3rd grade math, the average rank rose from 65.6 to 61.8, but the rise was not statistically significant (p=.7).[16]

Regression discontinuity

A different and perhaps more defensible way to estimate Project Challenge's impact on test scores is to apply RD to scores on the same test in the same year. Figure 4 and Figure 5 present RD graphs that compare Challenge-eligible and ineligible districts near the eligibility threshold (i.e., near poverty=0). Figure 4 presents results for 2nd grade, and Figure 5 presents results for 3rd. Both figures give scores for 5 different subjects in 1991-92 (year 3 of Project Challenge), as well as an all-subject "battery" score in 1990-91 (year 2).

If Project Challenge had improved 2nd and 3rd grade test scores, we would expect to see an upward discontinuity at the eligibility threshold. But we do not. Across grades 2 and 3, 11

or the 12 discontinuity graphs show either no discontinuity, or a discontinuity that is negative, nonsignificant (p>.10) and very small. The only graph to show a positive discontinuity is the one for 3rd grade language arts, and in that graph the discontinuity is nonsignificant (p>.10) and the estimated effect size is negligible (.06 SD).

The power of our RD analysis is limited because there were only 17 participating districts, but our lack of significant effects is not due to a lack of power. Even directionally, only one of the 12 discontinuities is in a direction suggesting that Project Challenge raised test scores. And our earlier analysis showed that none of the participating districts had class sizes of 15 as they were supposed to. The fact that participating districts had an average class size of 25 instead of 15 is not subject to sampling error and not an artifact of lower power.

In general, the results suggest that Project Challenge had practically no impact on test scores, perhaps because it had practically no effect on class sizes.

Class size reduction under the Basic Education Program

In 1992, as the Small School Systems lawsuit neared a judgment in favor of the plaintiff districts, the state passed the Basic Education Program (BEP), which increased total school funding by $111 million, or 4%, and reduced funding disparities between districts.

Districts' 1992 report cards included a detailed budget describing how they would spend the new state funds that it would receive under the Basic Education Program. They planned to spend 39% of the money on new teachers (34% in Project Challenge districts), which implied some reduction in the student/teacher ratio.

The reduction, however, was quite small. Districts would increase the total number of K-12 teachers by just 4% (8% in Project Challenge districts). Before BEP, Tennessee had a

ratio of 21 K-12 students for every teacher, or 26.1 students for every "regular" teacher (excluding teachers of special education, art, music, physical education, and vocational skills). After BEP the ratio would fall to 20 students for every teacher, or 25 for every regular teacher. (Our calculation may be optimistic since it assumes no increase in enrollment, and no change in the proportion of "regular" vs. other teachers.)

Clearly a reduction of the student/teacher ratio from 26 to 25 is nowhere close to the reform tested in Project STAR, where class sizes were reduced from approximately 24 to 15. Although BEP reduced class sizes, the size of the reduction was not informed by evidence from Project STAR.

Conclusion

The history of class size in Tennessee is a cautionary tale about evidence-based policy. The state commissioned a rigorous randomized trial, and the trial produced clear evidence that reducing K-3 class sizes to 15 could substantially raise test scores, at least in the short term. Yet the class size policies that followed were token or ineffective. Although Project Challenge was supposed to raise test scores by reducing K-3 class sizes to 15 in poor rural districts, our analysis of the districts' own report cards finds no evidence of reduced class sizes or increased scores. While it is true that the Basic Education Program reduced class sizes statewide, the reduction was token—from 26 to 25 students per regular teacher—nowhere near the reductions that were tested in Project STAR.

Why were the policies that followed Project STAR so anemic? Although poor implementation played a role, there is also a fundamental question regarding what the

purpose of Project STAR really was. A previous history has suggested that the purpose of Project STAR was not entirely to inform future policy decisions. The study also served a political purpose; it mollifed advocates for class size reduction while allowing skeptics to limit the time and money that would be spent on for the next four years (Ritter & Boruch, 1999). It seems that Project STAR, like some other government studies, served not just to inform action, but to delay it.

Several aspects of Tennessee's policy history make more sense if Project STAR was in part a delaying tactic. For example, why did the state wait four years to take action when by the end of Project STAR's first year, it was clear that class size reductions in kindergarten could raise test scores? Why did Project STAR test class size reductions far deeper than anything Tennessee or other states would prove willing to fund statewide? Why did the state spend more on the Project STAR than it would spend on the Project Challenge policy that came afterward? Project STAR cost $28 million over 4 years, or $7 million per year, while Project Challenge got just $2.8 million in state funds per year, which districts were expected to supplement with $1.7 million in federal funds. None of this makes sense if the sole purpose of Project STAR was to inform policy decisions.

If the state wanted evidence to be aligned with realistic policy options, it could have tested a variety of class size reductions in a wider range of grades. It could have taken preliminary policy steps after just the first year or two of results. And it would have spent more on the policy than on the trial.

Project STAR may have had more policy influence in other states than it did in Tennessee (Kim, 2006; Ritter & Boruch, 1999). Evaluations of Project STAR published between 1989 and 1995 played a role in California's 1996 decision to cap class sizes in grades

K-3. That decision did not raise achievement, as we discussed in the introduction, yet Project STAR was cited again as recently as the 2019 Los Angeles teacher's strike, when the union demanded caps on class sizes in grades 4 and up.[17]

Why does Project STAR still hold sway 30 years after it ended, in settings where its findings have already been shown not to generalize? Our continued reliance on a 30-year-old study testifies to how seldom we conduct randomized experiments and how naïve we can be about their generalizability. In a more robust environment for evidence-based policy, every state would conduct randomized experiments on a regular basis, and there would be less need to hearken back to a single experiment—one that did not yield effective policy in its home state.

References

Achilles, C. M., & Others. (1993). The Lasting Benefits Study (LBS) in Grades 4 and 5 (1990-1991): A Legacy from Tennessee's Four-Year (K-3) Class-Size Study (1985-1989), Project STAR. Paper #7. Retrieved from https://eric.ed.gov/?id=ED356559

Achilles, C. M., Zaharias, J. B., Nye, B. A., & Fulton, D. (1995). Analysis of Policy Application of Experimental Results: Project Challenge. Retrieved from Tennessee State University, Center of Excellence for Research in Basic Skills website: http://eric.ed.gov/?id=ED393151

Boyd-Zaharias, J. (1999). Project STAR: The Story of the Tennessee Class-Size Study. American Educator, 23(2), 30–36.

Cameron, A. C., & Miller, D. L. (2015). A Practitioner's Guide to Cluster-Robust Inference. Journal of Human Resources, 50(2), forthcoming.

Campbell, D. T. (1969). Reforms as experiments. American Psychologist, 24(4), 409–429. https://doi.org/10.1037/h0027982

Chetty, R., Friedman, J. N., Hilger, N., Saez, E., Schanzenbach, D. W., & Yagan, D. (2010). How Does Your Kindergarten Classroom Affect Your Earnings? Evidence From Project STAR (Working Paper No. 16381). Retrieved from National Bureau of Economic Research website: http://www.nber.org/papers/w16381

Chingos, M. M. (2012). The impact of a universal class-size reduction policy: Evidence from Florida's statewide mandate. Economics of Education Review, 31(5), 543–562. https://doi.org/10.1016/j.econedurev.2012.03.002

Cook, T. D. (2008). "Waiting for Life to Arrive": A history of the regression-discontinuity design in Psychology, Statistics and Economics. Journal of Econometrics, 142(2), 636–654. https://doi.org/10.1016/j.jeconom.2007.05.002

Dieterle, S. G. (2015). Class-size reduction policies and the quality of entering teachers. Labour Economics, 36, 35–47. https://doi.org/10.1016/j.labeco.2015.07.005

Educational Research Service. (1978). Class size : A summary of research. Arlington, VA: Educational Research Service.

Finn, J. D., & Achilles, C. M. (1990). Answers and Questions About Class Size: A Statewide Experiment. American Educational Research Journal, 27(3), 557–577. https://doi.org/10.3102/00028312027003557

Finn, J. D., Boyd-Zaharias, J., Fish, R. M., & Gerber, S. B. (2007). Project STAR and Beyond: Database User's Guide. Lebanon, TN: HEROS, Incorporated.

Folger, J., & Breda, C. (1989). Evidence from project STAR about class size and student achievement. Peabody Journal of Education, 67(1), 17–33. https://doi.org/10.1080/01619569209538668

Gilman, D. A., Swan, E., & Stone, W. (1988). The Educational Effects of a State Supported Reduced Class Size Program. Contemporary Education; Terre Haute, Ind., 59(2), 112–116.

Glass, G. V. (1982). School Class Size: Research and Policy (1 edition). Beverly Hills, Calif: SAGE Publications, Inc.

Glass, G. V., & Smith, M. L. (1979). Meta-Analysis of Research on Class Size and Achievement. Educational Evaluation and Policy Analysis, 1(1), 2–16. https://doi.org/10.2307/1164099

Graves, J., McMullen, S., & Rouse, K. (2013). Multi-Track Year-Round Schooling as Cost Saving Reform: Not Just a Matter of Time. Education Finance and Policy, 8(3), 300–315. https://doi.org/10.1162/EDFP_a_00097

Hanushek, E. A. (1981). Throwing money at schools. Journal of Policy Analysis and Management, 1(1), 19–41. https://doi.org/10.2307/3324107

Haskins, R., & Margolis, G. (2014). Show Me the Evidence: Obama's Fight for Rigor and Results in Social Policy. Washington, D.C: Brookings Institution Press.

Imbens, G., & Kalyanaraman, K. (2012). Optimal Bandwidth Choice for the Regression Discontinuity Estimator. The Review of Economic Studies, 79(3), 933–959. https://doi.org/10.1093/restud/rdr043

Jepsen, C., & Rivkin, S. (2009). Class Size Reduction and Student Achievement: The Potential Tradeoff between Teacher Quality and Class Size. Journal of Human Resources, 44(1), 223–250.

Kahlenberg, R. D. (2007). Tough Liberal: Albert Shanker and the Battles Over Schools, Unions, Race, and Democracy. Columbia University Press.

Karlan, D., & Appel, J. (2012). More Than Good Intentions: Improving the Ways the World's Poor Borrow, Save, Farm, Learn, and Stay Healthy (Reprint edition). New York: Plume.

Kim, J. S. (2006). The Relative Influence of Research on Class Size Policy. Brookings Papers on Education Policy, 2006(1), 273–295. https://doi.org/10.1353/pep.2007.0004

Krueger, A. B. (1999). Experimental Estimates of Education Production Functions. The Quarterly Journal of Economics, 114(2), 497–532. https://doi.org/10.1162/003355399556052

Martin, K. V. (1993). Constitutional Law - Tennessee Small School Systems v. McWherter: Opening the Door for Education Reform Symposium - The Tennessee Supreme Court: Judicial Activists: Comment. Memphis State University Law Review, 24, 393–412.

Nye, B. A., Achilles, C. M., Boyd-Zaharias, J., Cain, V. A., Fulton, B. D., & Tollett, D. A. (1995). Project Challenge Fifth-Year Summary Report: An Initial Evaluation of the Tennessee Department of Education "At-Risk" Student/Teacher Ratio Reduction Project In Sixteen Counties, 1990 Through 1994. Nashville, TN: Tennessee State University, Center for Excellence for Research and Policy on Basic Skills.

Nye, B. A., Achilles, C. M., Boyd-Zaharias, J., Fulton, B. D., & Wallenhorst, M. P. (1992). Project Challenge Preliminary Report: An Initial Evaluation of the Tennessee Department of Education "At Risk" Student/Teacher Ratio Reduction Project in Seventeen Counties.

Retrieved from Tennessee State University, Center of Excellence for Research in Basic Skills website: http://eric.ed.gov/?id=ED352180

Papay, J. P., Willett, J. B., & Murnane, R. (2011). Extending the regression-discontinuity approach to multiple assignment variables. Journal of Econometrics, 161(2), 203–207.

Porter, K., Reardon, S. F., Unlu, F., Bloom, H., & Robinson-Cimpian, J. P. (2014). Estimating Causal Effects of Education Interventions Using a Two-Rating Regression Discontinuity Design: Lessons from a Simulation Study. Retrieved from MDRC website: http://www.mdrc.org/publication/estimating-causal-effects-education-interventions-using-two-rating-regression

Ritter, G. W., & Boruch, R. F. (1999). The Political and Institutional Origins of a Randomized Controlled Trial on Elementary School Class Size: Tennessee's Project STAR. Educational Evaluation and Policy Analysis, 21(2), 111–125. https://doi.org/10.3102/01623737021002111

Schanzenbach, D. W. (2006). What Have Researchers Learned from Project STAR? (Harris School Working Paper No. 06.06). Retrieved from http://harrisschool.uchicago.edu/About/publications/working-papers/abstract.asp?paper_no=06%2E06+++

Schanzenbach, D. W. (2011). Review of Class Size: What Research Says and What It Means for State Policy. Retrieved from National Education Policy Center, University of Colorado website: http://nepc.colorado.edu/thinktank/review-class-size-what-research-says-and-what-it-means

Smith, C. C. E. (1990, November). Goals and Objectives of the 21st Century Challenge Plan. Tennessee Department of Education.

Smith, C. C. E. (1991, November). Goals and Objectives of the 21st Century Challenge Plan. Tennessee Department of Education.

Tennessee Department of Education. (1990). Commissioner's report card, 1988-89. Nashville, TN: Tennessee Department of Education.

Tennessee Department of Education. (1992a). 21st century schools report card. Nashville, TN: Tennessee Department of Education.

Tennessee Department of Education. (1992b). 1991 Commissioner's report. Nashville, TN: Tennessee Department of Education.

Tennessee Department of Education. (1993). 21st century schools value added assessment. Nashville, TN: Tennessee Department of Education.

Thistlethwaite, D. L., & Campbell, D. T. (1960). Regression-discontinuity analysis: An alternative to the ex post facto experiment. Journal of Educational Psychology, 51(6), 309–317. https://doi.org/10.1037/h0044319

Various. (1989). Project STAR and class size policy. Peabody Journal of Education, 67(1), 17–33. https://doi.org/10.1080/01619569209538668

Whitehurst, G. J. "Russ," & Chingos, M. M. (2011). Class Size: What Research Says and What it Means for State Policy. Retrieved from Brookings Institution website: http://www.brookings.edu/research/papers/2011/05/11-class-size-whitehurst-chingos

Wong, V. C., Steiner, P. M., & Cook, T. D. (2012). Analyzing Regression-Discontinuity Designs With Multiple Assignment Variables: A Comparative Study of Four Estimation Methods. Journal of Educational and Behavioral Statistics, 1076998611432172. https://doi.org/10.3102/1076998611432172

Tables

Table 1. Data availability by year and grade

a. District resources

School year	Student/teacher ratio (grades K-3)	Spending per student			
		Total	Federal	State	Local
1988-89 (before Project Challenge)		√			
1989-90		√			
1990-91	√	√			
1991-92	√	√	√	√	√

b. District average test scores

School year	Grade	Test	Subject					
			Language arts	Math	Reading	Science	Social studies	Battery†
1988-89 (before Project Challenge)	2	SAT-7*		√	√			
	3	BSF		√	√			
1989-90	2	TCAP						√
	3	TCAP						√
1990-91	2	TCAP						√
	3	TCAP						√
1991-92	2	TCAP	√	√	√	√	√	
	3	TCAP	√	√	√	√	√	

*The SAT-7 scores are unusable because the district average stanine scores are rounded to one digit, which for nearly all districts was 5 or 6.

†The battery score summarizes all subjects.

Table 2. Spending per student

Year	Project Challenge districts	Other districts	Difference
1988-89 (before Project Challenge)	$ 2,865	$ 3,053	$ 188
1989-90	$ 3,189	$ 3,274	$ 85
1990-91	$ 3,372	$ 3,444	$ 71
1991-92	$ 3,355	$ 3,474	$ 119

Note. When Project Challenge began, participating districts increased their spending by more than other districts. This table excludes Van Buren County, which participated in some years but not others.

Table 3. Student/teacher ratios, grades K-3

	Project Challenge districts			Other districts		
	Students	Teachers	Ratio	Students	Teachers	Ratio
1990-91	13,006	533	24	243,194	10,647	22
1991-92	13,054	502	26	242,598	10,725	22

Note. This table is limited to "regular" classroom teachers, excluding teachers of special education, art, music, physical education, and vocational skills. The table also excludes Van Buren County, which participated in Product Challenge during some years but not others.

Figures

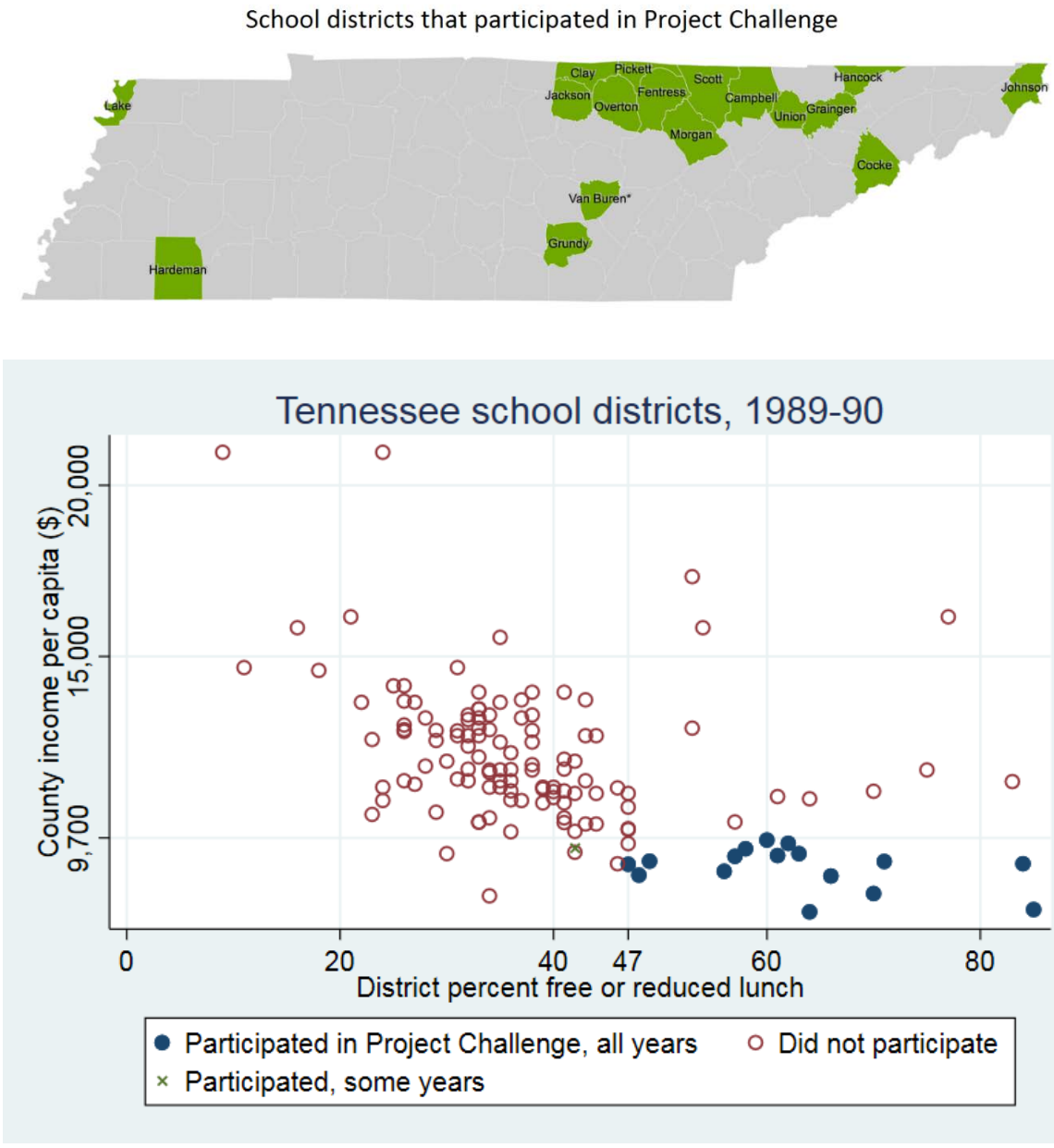

Figure 1. Project Challenge districts were in poor rural counties. All but one of the participating districts had county per capita incomes of less than $9,700 and at least 47 percent of student qualifying for federal lunch subsidies. (The one exception, Van Buren County, participated only in some years.)

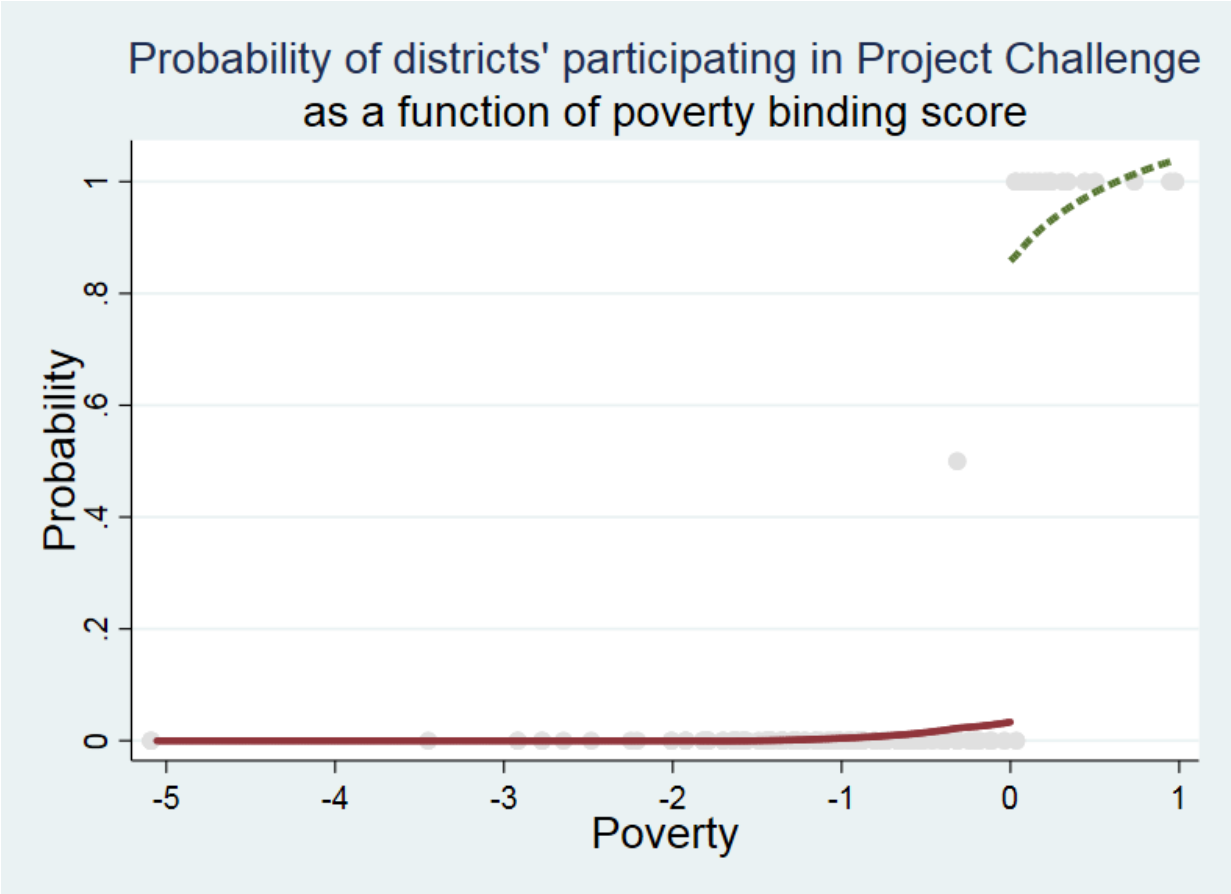

Figure 2. A binding score called poverty, constructed from per capita income and percent lunch subsidy, predicts a discontinuity in districts' Project Challenge participation. At the poverty threshold of zero, participation jumps from practically 0 to nearly 90 percent.

Note. All districts' participation values are coded as 0 or 1, except for Van Buren County, which is coded as 0.5 because it participated in some years but not others.

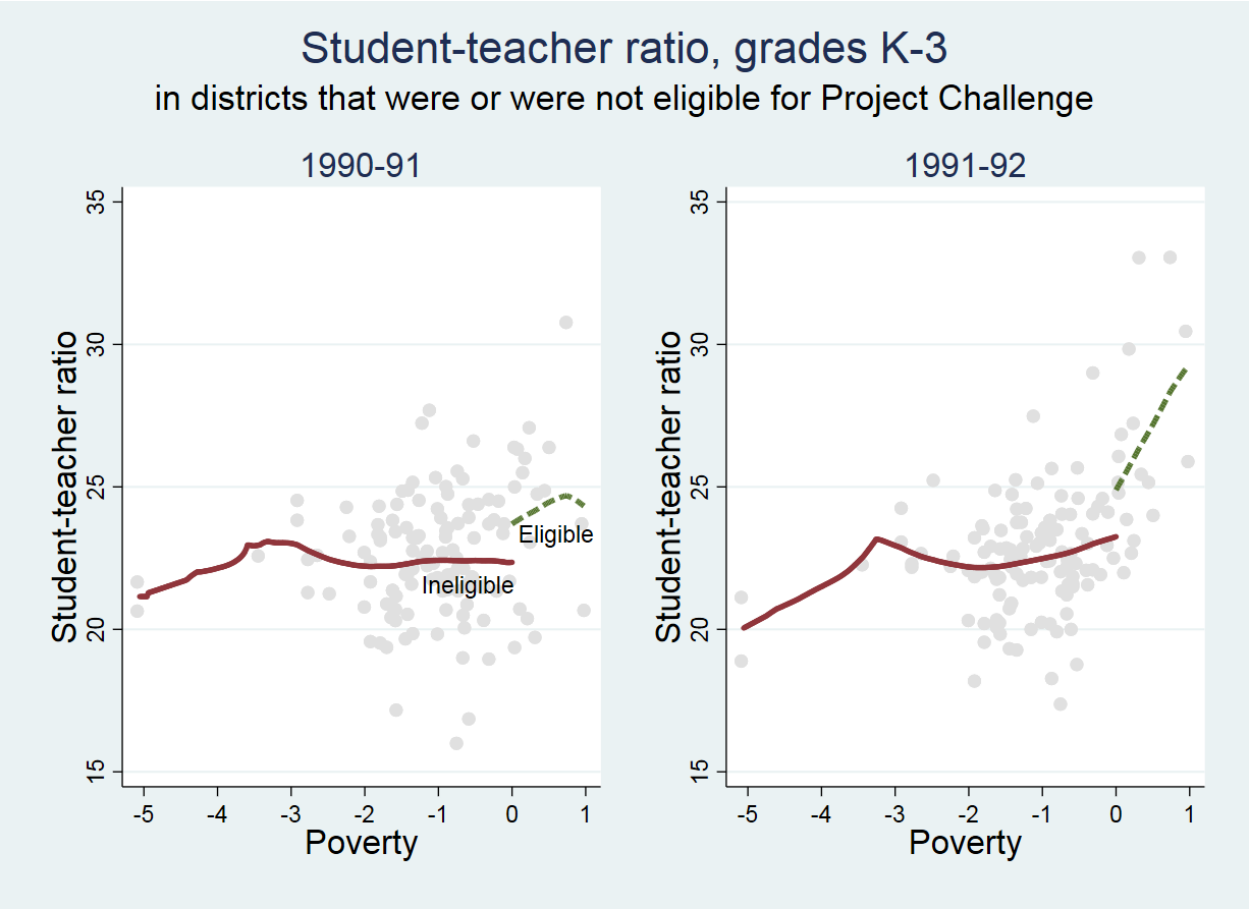

Figure 3. K-3 student/teacher ratios for Tennessee school districts during years 2 and 3 of Project Challenge. The districts right of the poverty cutoff of zero were supposed to be eligible for class size reduction, but the discontinuity plot shows that their student/teacher ratios were higher than those of districts that were barely ineligible.

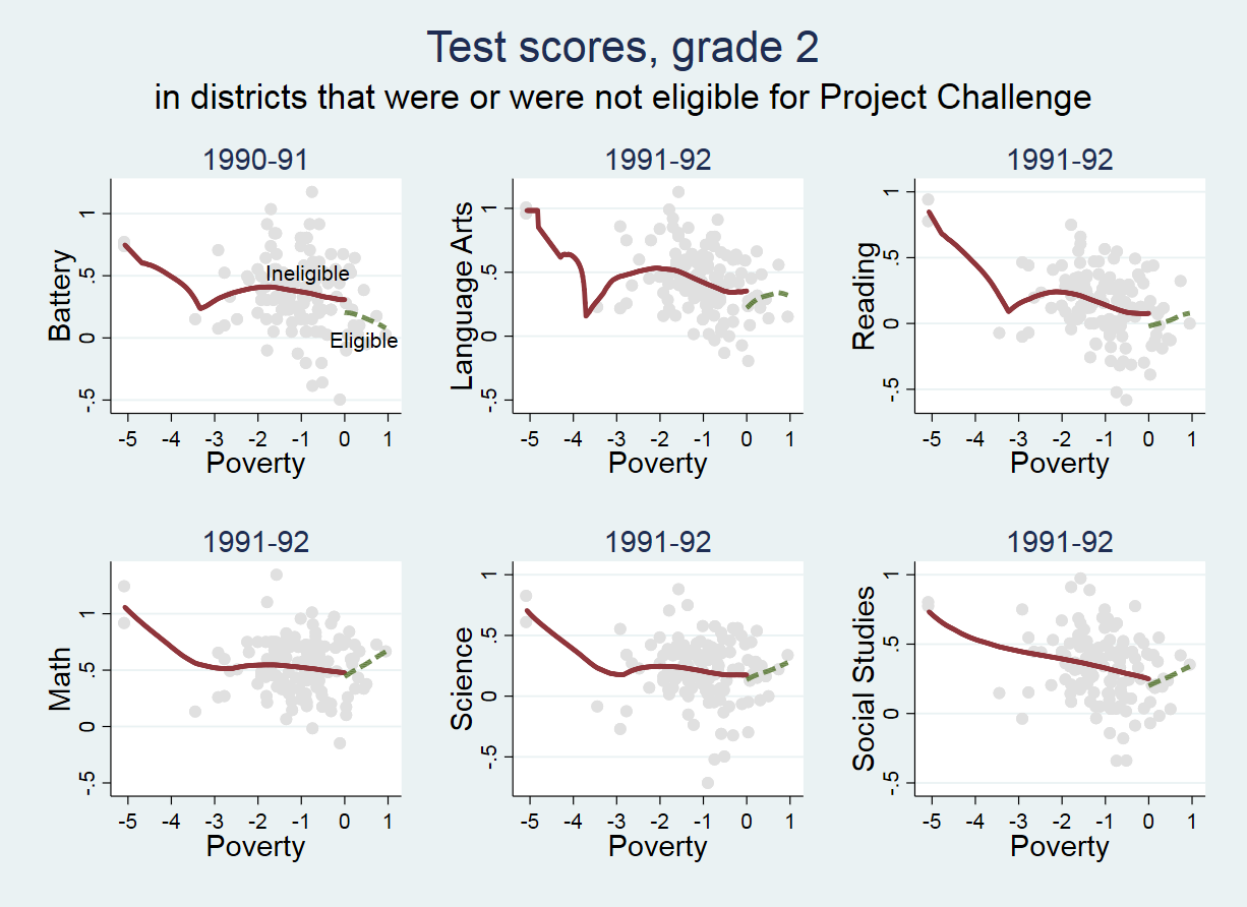

Figure 4. Average second grade test scores in all Tennessee school districts, standardized to a national reference. Districts just right of the poverty cutoff of zero had been eligible for Project Challenge for 2-3 years, yet their scores were no higher than those of nonparticipating districts just below the cutoff.

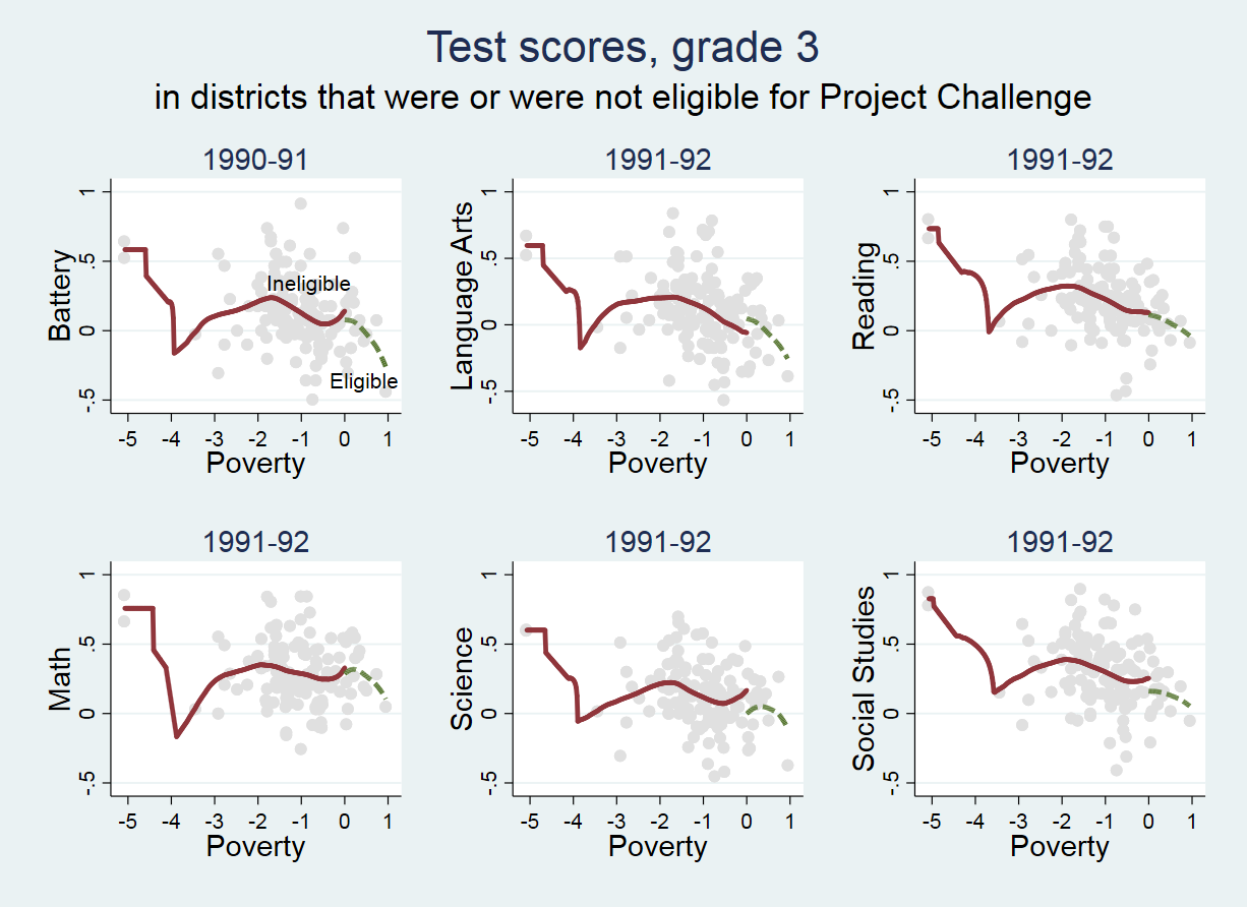

Figure 5. Average third grade test scores in all Tennessee school districts. As in second grade, there is no evidence of scores rising at the poverty threshold for Project Challenge.

Endnotes

[1] Dieterle's article describes an anonymous "State X," but the state is obviously Florida. Like Florida, State X passed class size reduction through a constitutional amendment in 2002, which capped class sizes at 18 in grades PK-3, 22 in grades 4-8, and 25 in grades 9-12.

[2] An alternative specification, using a quadratic model (Glass and Smith 1979), was less interpretable but agreed that test scores would improve little unless class size were reduced well below 20.

[3] We do not know why the aide condition was included. We have never found a motivation in the literature on Project STAR, or any debate of teachers' aides in policy histories from the era.

[4] The effect size has been reported in different ways, which can be confusing. As a fraction of the total SD (the usual standard), the effect size was 0.15-0.20, but as a fraction of the SD within treatment groups it was closer to 0.25, and as a fraction of the SD between classrooms it was 0.5-0.6 (Finn & Achilles, 1990).

[5] The Commissioner of Education's first description of the program stated that 15 districts were selected (Smith, 1990), but the evaluations consistently say the 17 districts participated (Achilles, Zaharias, Nye, & Fulton, 1995; Nye, Achilles, Boyd-Zaharias, Fulton, & Wallenhorst, 1992), and later statements by the Secretary of Education agree.

[6] East Tennessee has so few African Americans that it considered following West Virginia's example and seceding from the Confederacy during the Civil War.

[7] If there is one teacher in each grade from K through 3, then there are 4 classrooms of students if the district offers full-day kindergarten, and 5 classrooms of students if the district offers two half-day kindergartens. That is how we calculated the ratio $5/4$.

[8] In the reference population where the TCAP was normed, Z has a standard normal distribution with a mean of 0 and an SD of 1. Our formulas assume that Z scores are also normal within each district, with a different mean $\overline{Z}$ and SD s_Z for each district.

[9] Regardless of the year, the average rank among D districts ranked 1 to D is $(D + 1)/2$).

[10] Values of the forcing variables were highly correlated from year to year, but 1988-89 was the year in which districts were chosen for participation and the year in which the forcing variables most cleanly separate eligible from ineligible districts.

[11] There are two basic approaches. One is to restrict the data to districts that are eligible according to one forcing variable and then estimate the discontinuity around the threshold for the other forcing variable—and vice versa. The other is to fit a regression surface that estimates the relationship between the outcome and both forcing variables simultaneously. Neither of these approaches is viable in our small, sparse dataset.

[12] In previous literature, it has been assumed that eligibility requires being below the threshold on both forcing variable. In Project Challenge, eligibility required being below the threshold on one variable and above the threshold on the other. Our expression for the binding score is therefore a little more complicated.

[13] Division by the SD is reasonable but arbitrary. Different results could be obtained with a different divisor.

[14] The percentage of student with lunch subsidies was rounded in district report cards. We code the eligibility threshold as 46.5, so that districts with 47 percent of students subsidies could be eligible but districts with 46 percent could not.

[15] These calculations use teacher costs from 1992-93; if costs were lower in 1989-90, as they must have been, then reductions in the student/teacher ratio could have been deeper.

[16] These results exclude Van Buren County, which participated in some years but not in others.

[17] The settlement of the 2019 strike capped Los Angeles class sizes at 35 in grade 4-6, and 39 and middle and high school math and English. These caps were well above the class size of15 that was tested in Project STAR, and well above Los Angeles' average class size of 25.