Association for Public Policy Analysis and Management
www.appam.org

UNIVERSITY OF MARYLAND 1856
SCHOOL of PUBLIC POLICY

Peter H. Rossi Award Lecture
November 9, 2012

Thomas D. Cook
Awardee

Douglas J. Besharov
Moderator

Douglas Besharov: Welcome to the 2012 Peter H. Rossi Award Lecture.

Peter Rossi died in 2006. He was a dear friend of mine, and he was a great friend of the Association of Public Policy and Management (APPAM) and many of its members. He loved coming to these meetings. And so when we were thinking about having an award to commemorate his career and lifetime of contributions, it was natural that we would have it at APPAM.

Let me read from the New York Times obituary about Peter: "His research centered on the effectiveness of social programs in several areas: poverty, hunger, and prison reform, although he personally favored social policies that helped to disenfranchise." Actually, in college he was a Trotskyite, and he was in Alcove 2 at City College of New York, which you may know the PBS special and the book, *Arguing the World*. Those were his ethical and normative origins. "He sought to determine objectively whether programs designed for the disadvantaged actually worked. As a result, his work seemed to appeal equally to conservatives, who invoked it to highlight how programs never work and liberals, who invoked his work as showing the need to improve programs designed to help the disadvantaged."

The Peter H. Rossi Award honors a lifetime of achievement of Peter Rossi by recognizing the important contributions to theory or practice in program evaluations. The award can be for a specific project or product, or a lifetime.

The awardee receives a plaque—Pete used to say he saw one of these at a country club for winning a golf tournament, but that's what we have—plus a check for five thousand dollars. Most important, the winner is invited to give a lecture here at APPAM.

Past awardees have been Fred Mosteller of Harvard University, Rob Hollister of Swarthmore College, Russ Whitehurst now at the Brookings Institution, Judy Gueron of MDRC, Becka Maynard of the University of Pennsylvania, and Howard Bloom of MDRC.

The award committee was made up of a combination of past APPAM presidents and past awardees. This year's award committee was Judy Gueron, Rob Hollister, Becka Maynard, and David Weimer.

And as you know, this year's awardee is Thomas Cook of Northwestern.

Tom's primary field of work is the methodology of program evaluation and, in particular, field experimentation and quasi-experimentation. He's co-authored some of the leading text in the field, including: *Quasi-Experimentation: Design and Analysis Issues for Field Settings* with Donald Campbell*, Foundations of Program Evaluation: Theories of Practice* with William Shadish and Laura Leviton, and *Experimental and Quasi-Experimental Designs for Generalized Causal Inference* with Campbell and Shadish.

In selecting the Rossi Awardee, we solicit nominations from APPAM members and others. I would like to read from one of them.

> One of the most striking things about reading Tom's substantial evaluation work is that he rarely chooses questions that can be answered by a straight forward randomized trial or regression discontinuity design. Instead, he chooses difficult and important substantive questions that he examines with often imperfect research designs, motivated by his social conscience and a genuine desire to know what programs and policies can improve the lives of children and families.

If you hear an echo of Pete's life, it's deliberate, I think.

Tom, please accept this year’s Peter H. Rossi award.

* * *

The year before Pete died he was able to give a talk here with some of the same people as discussants. And so we’ve adopted that format for the Rossi Award. Tom will talk for about 30-35 minutes. And then we’ll do 10 minutes each from our discussants, Judy Gueron, Rob Hollister, and Becka Maynard, in that order.

Tom, welcome. Thank you and we look forward to your presentation.

Thomas Cook: What an honor to get a Pete Rossi Award.

I have had two-and-a-half mentors in my life, and Pete was one of them. When I was quite young, I met with him seven or eight times a year for six years. And we had the whole afternoon and evening to spend together. Sometimes Hank Levin was with us, but mostly it was just the two of us. That’s an amazing amount of time for a 30-year-old to be spending with a 55-year-old superstar, as he was at that time.

Pete taught me tremendous amounts about evaluation and about the politics of evaluation. He taught me tremendous amounts about ideas he was thinking of. He invented the vignette studies about that time. And he told me a lot about how the worlds of research worked at universities and in government in a realistic, hard-headed way that I could not have gotten anywhere else. And my life has been characterized by a bit of schizophrenia between Donald Campbell talking at a high level of how science should be done, and Peter initiating me into how science is done. And it’s a wonderful tension to live with.

I got to love the guy. I loved his scintillating wit. Who else could have written about the “metallic” laws of evaluation like he did?

I loved his trenchant intellect and creativity. The way he invented vignette studies so that he could have a randomized experiment within a survey with random selection, as he used to describe it. And how he did those wonderful surveys of when the president dies to show public reaction to the death of President Kennedy.

So he was a great intellect. He was also a great deflator of all those full of pedantry, full of posing, full of cant. I could tell you stories about being present when he stuck in his dagger and the deflation occurred to those pompous S.O.B.s that he was skewering. But my wife assured me that I should not tell such stories in public.

He was also a tremendous defender of the weak. When he saw people with power exercising that power in ways that demeaned or in any way diminished the weak, he was there standing up to the powerful, always in defense of the weak.

How loyal he was. Not just to Alice and to Pete Jr., but also to his students and friends and colleagues.

Tremendous person.

What a miracle worker he was to have created for himself a heart that was at once so stout and so soft. A very special person, and I am very honored to accept this award.

* * *

Some years ago (1958 to be precise) at Northwestern, Don Campbell invented the regression discontinuity design. Many of us wrote about it for thirty years or so. And all our writings fell into a big deep void. Nobody wanted to do anything about regression discontinuity, and it was a design sitting there with no takers.

Thank goodness that there are so many well-connected, young, applied micro-economists and that this wonderful coterie of people in about 1995 discovered it lying dust-covered on a shelf somewhere, dusted it off and applied it. And as they wrote about it people started discovering, “Oh boy, I can apply this here. I can apply this there. I can apply it somewhere else.” And soon regression discontinuity came to take an acknowledged, but maybe lesser role in the quiver of causal arrows at the disposition of the hard-headed among the social scientists.

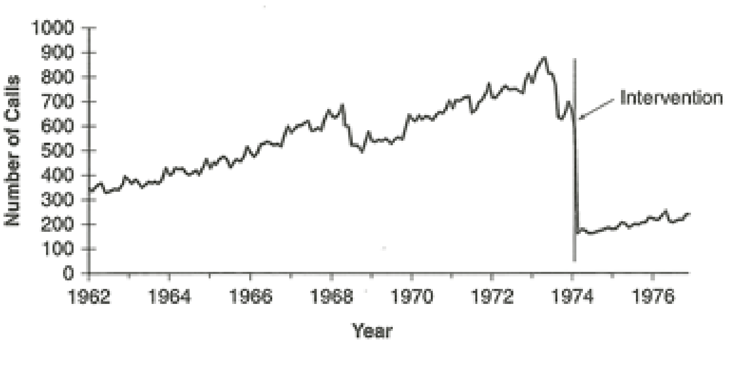
I want today, in a very minor way, to maybe begin the process of thinking whether it is time to reappraise the role of interrupted time series in the policy sciences. It has been around even longer than regression discontinuity. It is routinely used in many fields—engineering, finance, parts of medicine—where it’s accepted routinely as valid, i.e., capable of evidence-proof methodology for probing causal claims.

In doing this, I want to suggest that not all kinds of interrupted time series are worth worrying a lot about in the policy sciences. I'm going to therefore give you a sense of what I think are some of the better ways of doing interrupted time series.

I do not care if you carry away the thesis of this talk. What I care about is that you carry away an impression—a sense that this is something worth thinking about a bit more. Because if you're thinking about it in a cognitive science sort of way, you are primed. And once you are primed, you will find applications. Not applications for the sake of application, but applications because they're important.

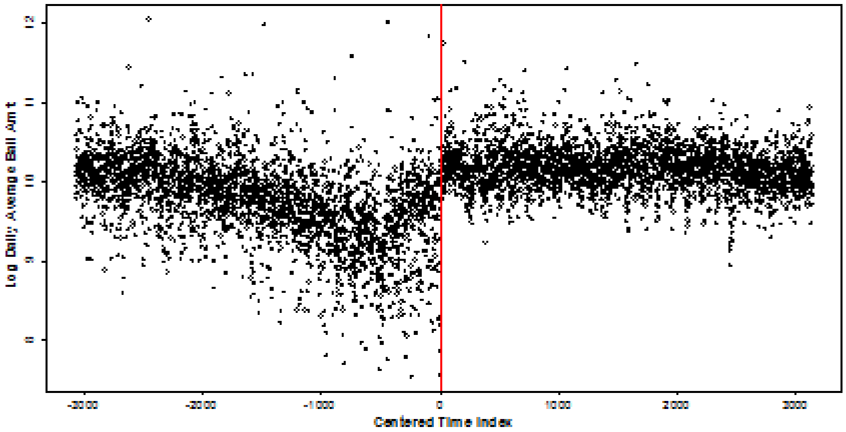
Here are some time series you may all know well. In Cincinnati, Bell began to charge for directory assistance calls. Some of you may not have much sense of this, but at one time people called to find out the telephone number of friends and acquaintances, and you got the telephone number for free. Well, Cincinnati Bell decided it was going to start charging for these calls. And here is a time series with weekly data. You can see that the calls for directory assistance are going up and up and up. But at the time of the intervention, in the very next week after it was implemented, there was a big drop off the cliff. And this drop off the cliff could perhaps be a causal effect of charging for directory assistance.

I. Cincinnati Bell and Charging for Directory Assistance



Here is another more recent one that comes from Cook County, which decided after a while it was going to institute a new practice of bail hearings. These bail hearings were not going to be done face-to-face with a real life judge, but instead were going to be done by video. The presumption being that this would make it faster and more inexpensive. What you see are daily data for eighteen years. (Not a very frequent occurrence, so don't expect to find these around the corner.) It's daily data for eighteen years on the amount of bail that a person had to pay in order to go home in the interim.

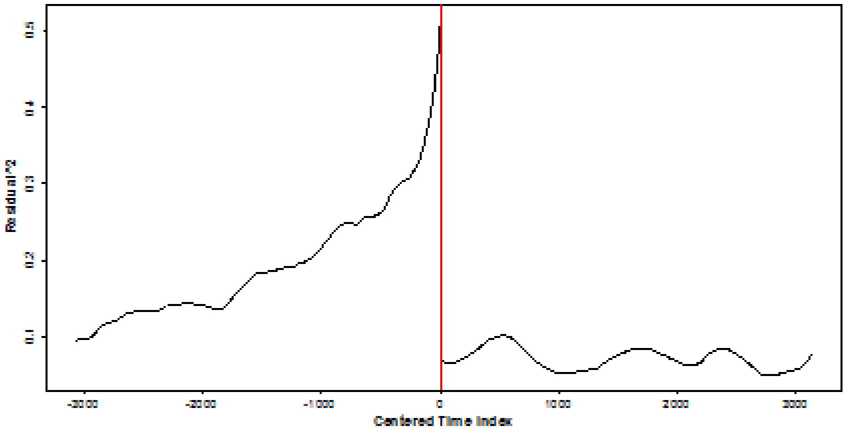
I. How Bail Hearings by Video affect Bail Amounts Set



You can see here that there are three things happening which statistical analysis corroborates. There is a downward trend initially with a little kick-up. There is a big mean difference at the intercept—the day after the video system is instituted. You can see that there is a change in slope and you can probably see that there is change in variance so that the variation around the trend line is less than it was beforehand.

A version of this time series with monthly data was used to halt the video proceedings for bail hearings, because there was no earthly reason why bail costs should go up just because they were using a video system. And with these bail amounts that means that people can't make the bail are more likely to spend time in prison, in jail, etc. So this has been used.

I. How Video Bail Hearings affect Variation in Bail Amounts



These examples seem to work because they have a long and very stable pre-test time series, and this promotes more confident extrapolation—this extrapolation beyond the intervention point that is supposed to be the counterfactual. As we all know, we are probably all somewhat leery about extrapolations. The further out, the more leery. So, initially around the cutoff—see I'm slipping into regression discontinuity language because they have a lot in common—the extrapolation is stronger. In this particular example, these are daily or weekly data. And there are very few alternative interpretations that are going to be related to the outcome and that occur in this one day after the video system is instituted or one week after Cincinnati Bell charges people.

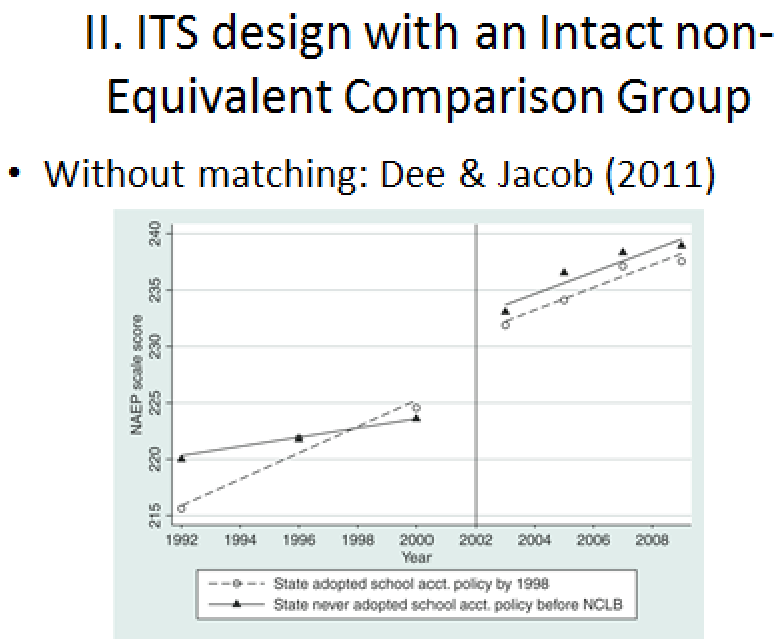
That short time interval is crucial for ruling out history alternative interpretations. It has immediate treatment onset—there's no diffusion of the treatment. In many cases in the real world, you gather treatment and it slowly diffuses through the population. It's not instantaneously available. Also in these cases there is an immediate response function: It dropped off the cliff straight away.

In many cases, there is delay in causation and we do not always know the theory of the length of the delay. We have a classical example where we do know. It is the time series studies of the effects of blackouts in New York City, when nine months later there was an increase in births. Right? Nine months? We know about it.

Also here there are very large effect sizes. The Cincinnati Bell is about twelve standard deviations relative to the intertemporal variation that preceded it. We should all power our studies to detect twelve standard deviation effects.

These conditions are very rare. We have not long intervals between observations, and more history alternative interpretations can come in then. When you have a single interrupted time series, the treatment diffusion is often a temporal process, many effects have a delay, and we power our studies to detect effects of one fifth of the standard deviation, and not twelve standard deviations.

How to improve the single interrupted time series? One way is to add a comparison time series. Here is some data by Dee and Jacob as part of the evaluation of a mechanism in No Child Left Behind, the federal initiative instituted under the second President Bush. What you see here are two sets of states before the intervention. You see states that already had an accountability system before No Child Left Behind occurred; and the other line, the less steep line in the pre-treatment period is of states that had no consequential accountability system before, but those states would have to institute one in 2001 when No Child Left Behind passed. If you were to extrapolate those two trend lines the extrapolation would be such that in the states that already had a system of accountability in place, the slope did not change and the intercept may have changed a little. But the important thing is what happens to the two slopes relative to each other. You can see in the states that did not have a consequential accountability system. When they got them, they really jumped up in slope and really jumped up in mean.

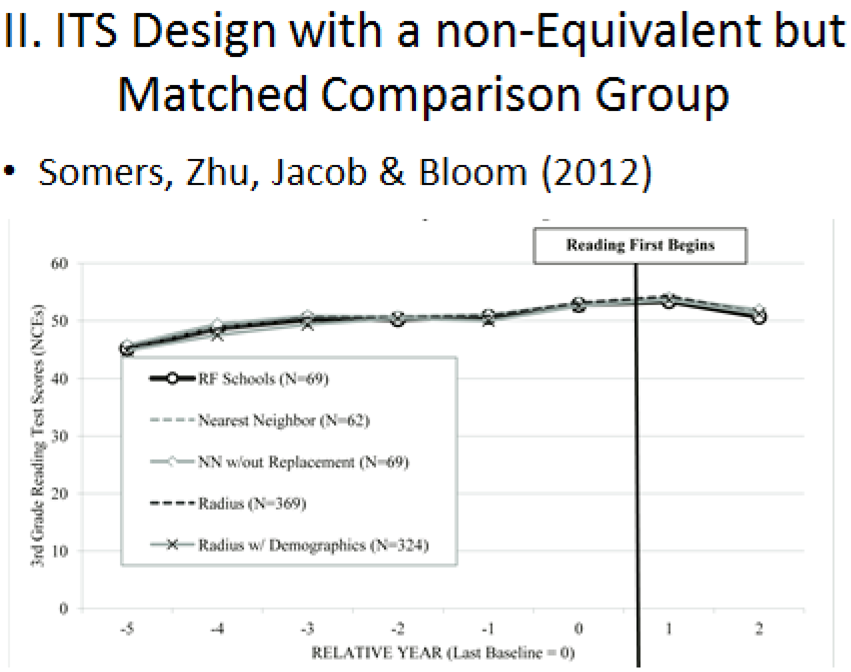


II. ITS design with an Intact non-Equivalent Comparison Group

- Without matching: Dee & Jacob (2011)

This was interpreted as an effect of the consequential accountability mechanism within No Child Left Behind increasing achievement. Now the logic of the analysis and the problem we have to wrestle with is: Do the obtained post-intervention differences in means or slopes differ from the predicted values for them after extrapolating from the pre-intervention slope difference? The logic here, of course, is akin to a difference-of-differences logic. And the main problem has to do with interaction of selections of history, that is to say, outcome-correlated events that co-occur immediately after the treatment and do so more with one population than another. For example, the time series I just showed you is for fourth-grade mathematics, using NAEP as the outcome variable. We know that about this time period, in 2000, the National Association of Teachers of Mathematics set out new standards. Could those new standards be an alternative interpretation, a delayed effect of new standards? Well if, and only if, there was some reason to suppose that those standards were implemented differently in that kind of state that already had a consequential accountability system versus that state which only got one because of No Child Left Behind. That's the logic here. It's a difference-of-differences kind of logic. And the main problem is: Are there differential effects occurring after the intervention that affected one group more than another?

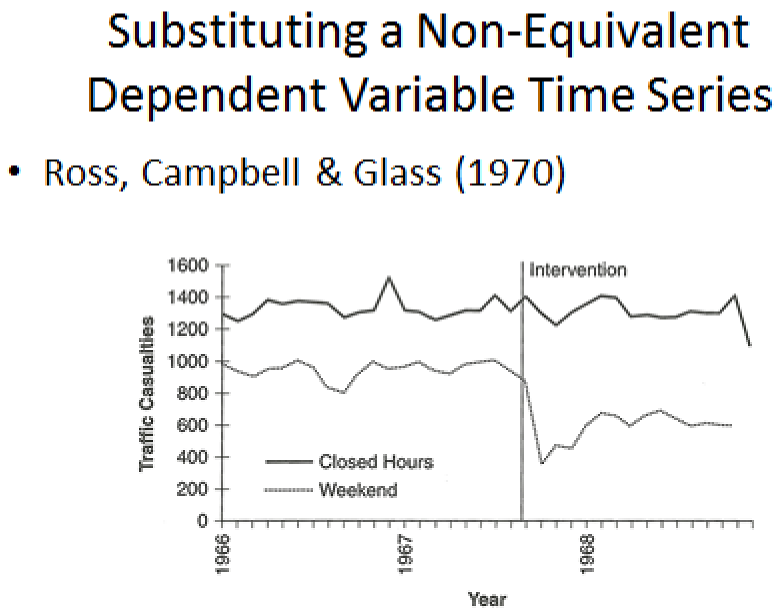
You can do this with matching the time series. Here is a plot of some Reading First data, with four different ways of matching the schools that got Reading First with other schools that were around. The idea is to match them so that we somewhat recreate the logic of the randomized experiment. We can look here and say, “Oh boy, these groups do not differ. And they do not differ however you choose to match them. And if they do not differ over such a long period of time, what reason do we have to think they may differ afterwards?” Well, they might differ afterwards, because it is still the case that some schools that get Reading First or the other control schools have things happening commonly in one group or another immediately after the treatment. So matching helps, but it does not get around this problem completely. But if you are embedded in the anthropology of the setting, it will often be the case that no one can come up with alternatives.



II. ITS Design with a non-Equivalent but Matched Comparison Group

- Somers, Zhu, Jacob & Bloom (2012)

Here is another way of having just two groups in which it is done. This is substituting a non-equivalent control group series, or match control group series, with a non-equivalent dependent variable time series. It is a classical example here from Britain of what happened when they changed their drunk driving laws and the outcome variable here is traffic casualties. If you look at the bottom of these series you will see that when the drunk driving was changed there was a drop in traffic fatalities and serious accidents. So what is the top darker line? In Britain, there’s much more social drinking in pubs than there is in other countries where there is more solitary drinking alone. And there are certain hours in which pubs are open and closed. What you have here is two time series, one of which is hours when pubs are open and the two hours afterwards versus hours when pubs are closed: mornings, through the night, etc.



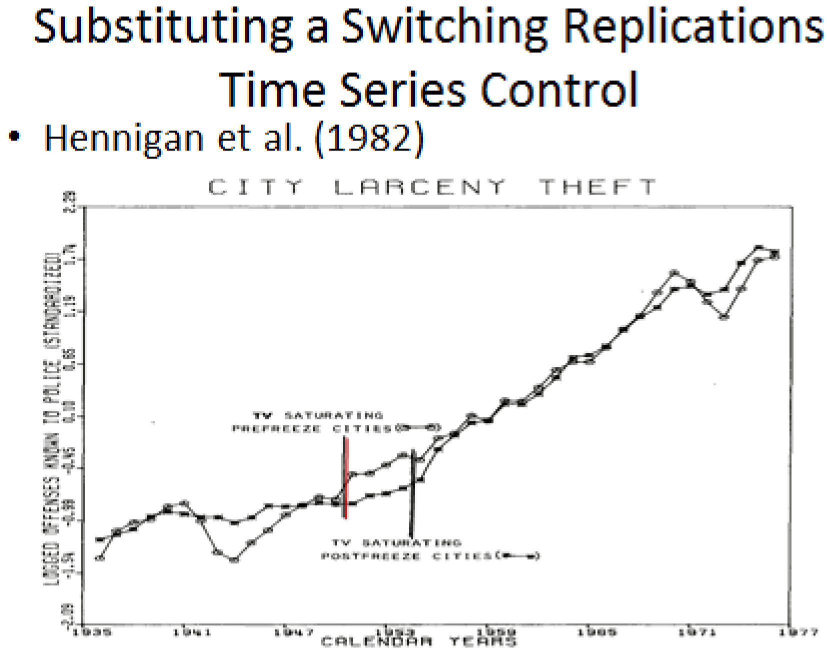
Substituting a Non-Equivalent Dependent Variable Time Series

- Ross, Campbell & Glass (1970)

Because if you look at just the bottom time series and you see that drop you can say, well, that may just be due to safer cars. It may be due to some other campaign that was taking place in Britain at the time about safer driving. It could be due to weather, because weather affects traffic accidents and the like. But it's pretty hard to say that the safer cars are on the roads when pubs are closed and not on the roads as much when pubs are open, or that the weather is different when pubs are open or closed, etc.

So what you have here is a time series control which is theoretically subject to the same alternative interpretations that happen around the time of the implementation of the treatment, and many of those history alternative interpretations go away. The alternative interpretation has to be something that operates at exactly that time and differently when pubs are open than when pubs are closed.

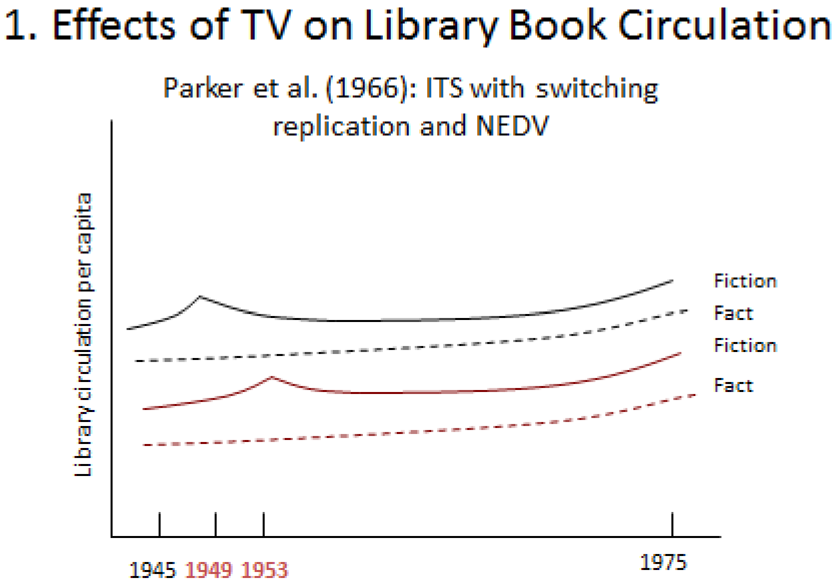
Sometimes you can substitute, still on two group designs, a switching replications design. This is a study of the effects of the introduction of television on larceny in the United States. The two series go from 1935 to 1977. Here you have one set of communities in the United States which introduces television early. They tend to be richer, more densely populated places. We look at when saturation reaches 90 percent plus of homes. We have the diffusion curves, and it takes about three years to get there. We can see where the larceny increases, which it seems to, but it doesn't increase in the communities that still have no television. Television licenses in the other communities got to be reissued in 1953. Because there had been a lot of corruption around the issuance of television relicenses, it was a license to print money, in essence. So, the Federal Communications Commission had stopped the diffusion of licenses and opened it again in 1953. Then everybody crowded into the market, and what you see here is that the original control series that did not have television early but got it later now goes up and the treatment series stays flat. And that period in the middle, which is highlighted, is potentially the effects of television on larceny. The earlier one you see is during the war. For some reason, police reported larcenies differently in larger cities than in smaller cities from 1942 to 1946.



Substituting a Switching Replications Time Series Control

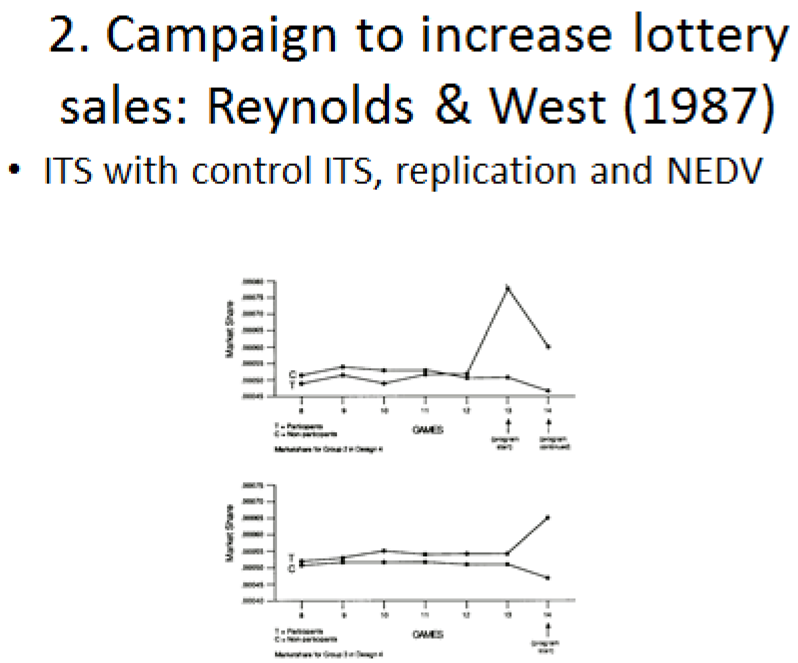
- Hennigan et al. (1982)

It is much better to add multiple interrupted time series. Going back to this television example: What we are doing here is we are looking at the effects of time interaction of television on library book circulation. Because watching television takes time, mostly it is about leisure activities and the question is: Does it displace reading of library books? You can see here that in communities that got television earlier there was a decrease in the circulation of fiction books, where there was no such change in the circulation of fiction books in those that got television later. Those that got television later had a change in fiction books at the right time. But neither of these had a change in the circulation of fact books. Why? You don't go to television to learn about the connective tissue in helix nemoralis. You’re not going to get that. So there was no reason to suspect that the introduction of television should have affected the circulation of fact books. What you have here is a time series, supplemented by a control time series, supplemented by a replicated later time series, supplemented by non-equivalent dependent variable series.


1. Effects of TV on Library Book Circulation

Parker et al. (1966): ITS with switching replication and NEDV

Here is another example of the same notion. This is a study in Arizona of the effects of having to advertise that you could get tickets for the state lottery. The state lottery was really worried that people were not buying tickets so they had a campaign. They would not let it be done by random assignment. The campaign was to hang in some stores a notice at the checkout counter that says: “If we have not asked you, ‘Do you want to buy a lottery ticket?’ then you can have one for free.” That’s all the intervention was. And what they did here was match stores. They matched them by chain. In the same chain, there was a treatment control. They matched them by zip code. They had many prior time points, as you see at the top one here, and that matching was pretty good and you can see that not much is happening. In the top series the treatment is instituted between 12 and 13, the penultimate points, and you see a big uptick. In other stores, it was implemented at a different time later and now you see the same uptick, but at a different time. So what you have here is matched control interrupted time series (ITS). You have a replication. And they also had non-equivalent dependence variables because you might say that this buying of tickets is due to the fact that they have just become suddenly more affluent around there. There is also a time series, which I can’t get into one graph here, showing that there was no uptick in the sales of gasoline and no uptick in the sales of convenience goods. Just an uptick in the sale of lottery tickets. So we have multiple controls here.



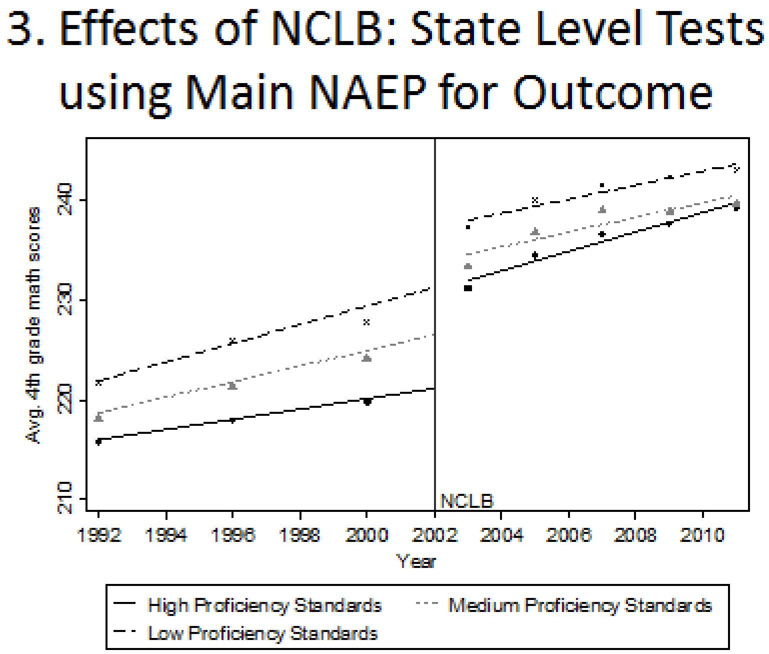
2. Campaign to increase lottery sales: Reynolds & West (1987)

- ITS with control ITS, replication and NEDV

I now want to move to another example quickly that does not lend itself to one figure. But the several figures, if you can put them into one figure in your mind, exemplify multiple time series controls being used to facilitate causal inference by making the alternatives less and less and less plausible. We saw before this work of Dee and Jacob on No Child Left Behind, where they tested this consequential accountability mechanism and used the difference between these pre-intervention time series to extrapolate the lines and claim that there was an effect of suddenly getting accountability systems in 2002.

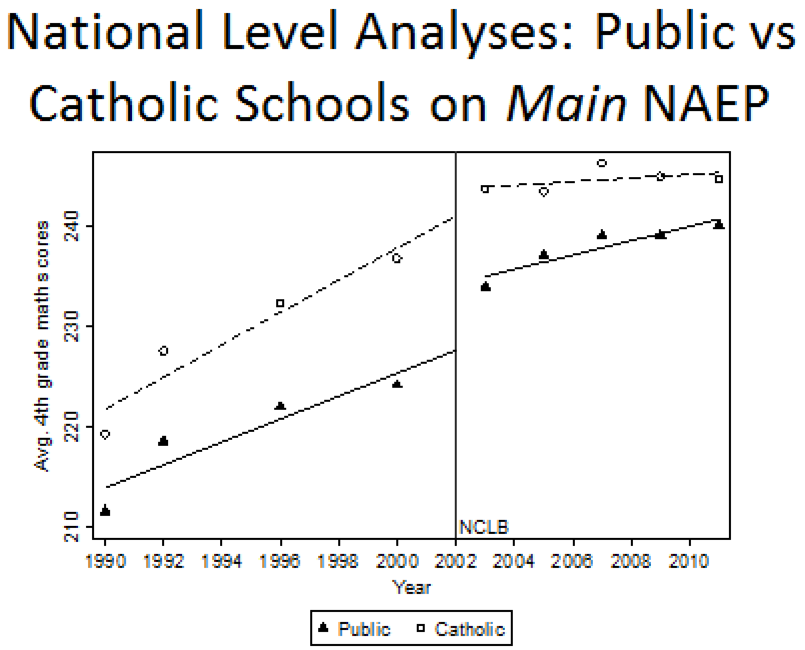
Here is another quite different mechanism involved in No Child Left Behind because you can suddenly get a consequential accountability system, but states were free to do many things in the initial regulations about No Child Left Behind—things that could strengthen or weaken the accountability system. For example, if you had an easy state test or if you had a very low threshold for children passing, then not many children were going to fail and not many schools were going to fail. Then you would not have to implement many of these consequential mechanisms. So a state could implement consequential accountability with high proficiency standards, which means a lot of kids would fail, a lot of schools would have to change, and a lot of change would have to take place in the education reform efforts. Or you could implement with low proficiency standards which meant little change would take place. Here are three levels of proficiency standards in the year 2003 by state.

The bottom line here is states that have hard tests and have higher proficiency standards, and you can show whether they were indeed called upon to make more change. You can see here a bigger change in intercept and a bigger change in slope such that fiscal analyses show a change in mean, slope, and final time point here. And it has three levels. So here is a different mechanism also intrinsic to No Child Left Behind, just like consequential accountability is, and having the very same implications about the effects of it.



3. Effects of NCLB: State Level Tests using Main NAEP for Outcome

These are state level studies of mechanisms, though; they are not a national study of a national program. There is no national estimate you can get out of this. In order to get national estimates of the effect of No Child Left Behind, since it's a public school intervention, you have to compare it to private schools in some way. Here, public schools are compared to Catholic schools on the main NAEP test. You can see that the public schools initially underperform the Catholic schools. But they have a bigger change in intercept, a bigger change in slope, and with very few degrees in freedom the fine endpoints here are reliably different from each other.

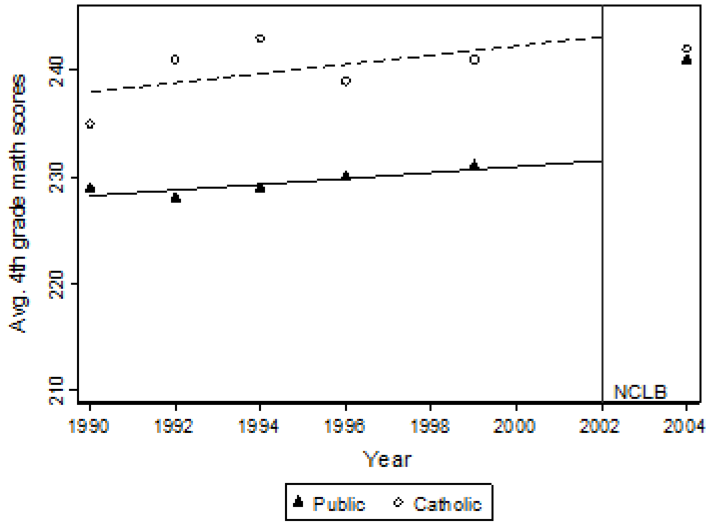


National Level Analyses: Public vs Catholic Schools on *Main* NAEP

But maybe NAEP has a test that changes its items every so often in order to reflect presumptive changes in the national curriculum. We also have a NAEP test called Trend NAEP which does not change its items. Trend NAEP, however, did change its sampling design so there is only one time point that's comfortable. And for Trend NAEP where the items have not changed, looking at Catholic versus public schools, you get the same difference this time, although it can only be tested at change of intercept.

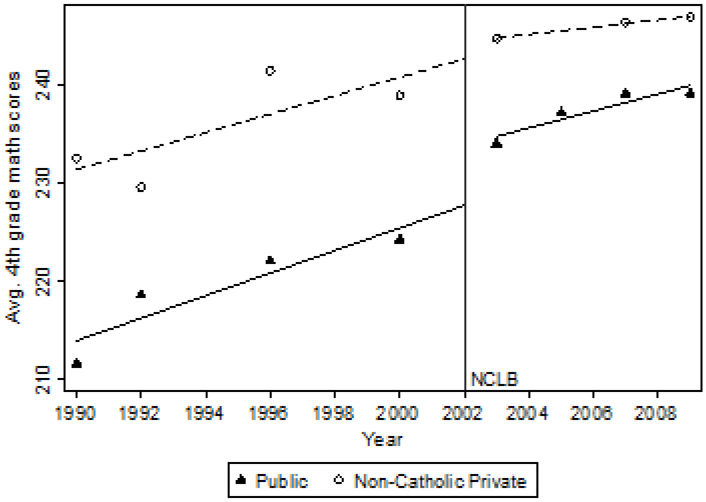
What about other kinds of private schools? We have non-Catholic private schools. If you do it for non-Catholic private schools, you get basically the same pattern result. You get almost the same causal estimates as with Catholic schools, but there is more variability because of the way NAEP is conducted with Catholic schools and it's not quite statistically significant in that one case.

National Level Analyses: Public vs Catholic Schools on *Trend* NAEP



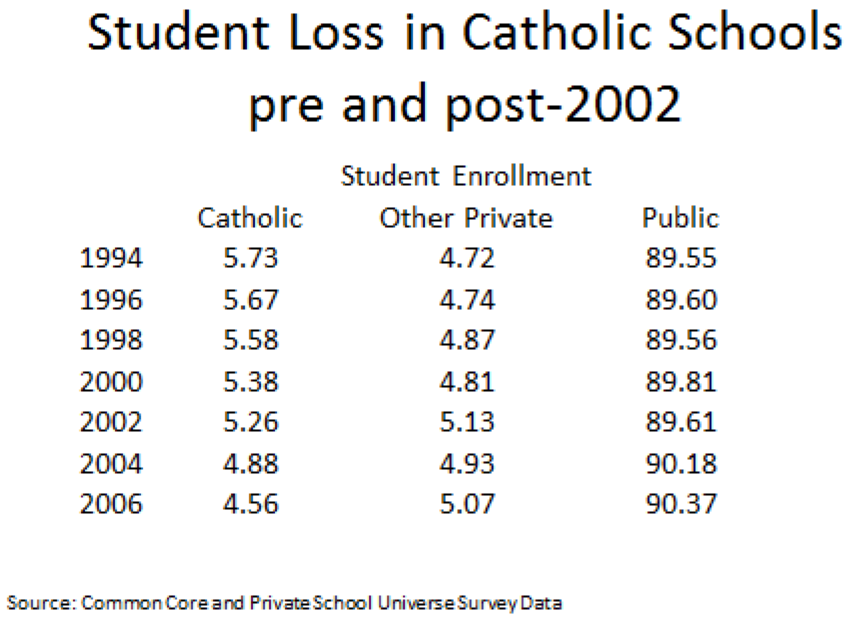
Thus, if you can imagine all of that as one big time series, we see that in two un-correlated state partitions, there is no correlation between the states that institute higher performance standards and the states that have a consequential accountability system. We see basically the same results. In national tests using public versus Catholic, public versus non-Catholic, and main and Trend NAEP tests, we see basically the same patterned results. There is a total consistency across national and state level, across types of comparison group, and across achievement test measures in these interrupted time series.

Public vs *Non-Catholic* Private Schools on Main NAEP



But consistency is not causation. What about student loss from Catholic schools? In 2002, in Boston, we began to hear about the sexual abuse concerns. Did it somehow lead to students leaving Catholic schools, which reduced the mean for Catholic schools? Did they move into public schools and increase the mean in public schools? And others would have moved into non-Catholic private schools and lowered the means there. With the ITS design approach which gave you the state tests with two different mechanisms, for the loss at Catholic schools to work, you would have to assume that the exodus from Catholic schools into public schools was differential in such a way that there was more of it in consequential accountability states than non accountability states, and there was more of it in states with high proficiency standards than not. This is not very causal.

Anyway, you can squint at numbers of enrollment in Catholic schools, in other private schools and in public schools. You can see there is a little change of one-fifth of 1 percent in Catholic schools after 2002. There’s an increase of two-fifths of 1 percent in the public schools.



Student Loss in Catholic Schools pre and post-2002

		Student Enrollment	
	Catholic	Other Private	Public
1994	5.73	4.72	89.55
1996	5.67	4.74	89.60
1998	5.58	4.87	89.56
2000	5.38	4.81	89.81
2002	5.26	5.13	89.61
2004	4.88	4.93	90.18
2006	4.56	5.07	90.37

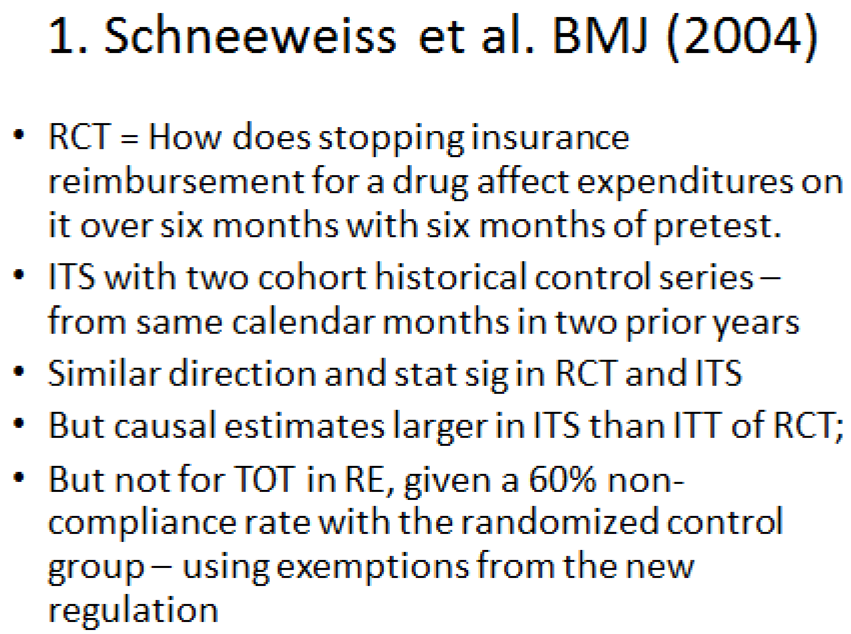
Source: Common Core and Private School Universe Survey Data

If you want those tiny changes to account for all of the indifferences that we have seen, the children who move would have to be scoring way off the scale for this to happen, so far off the scale that it is impossible. There were also new math standards instituted in 2000. That would be a viable alternative interpretation if (1) we were willing to assume that math standards affect achievement with a two year delay, (2) if math standards are differently implemented in states with consequential accountability and in states with high proficiency standards given that those two kinds of states are not correlated with each other, (3) if we also believe that the diffusion of new math standards affected public and private schools differently, and (4) if you believed that just issuing new standards could have those kinds of effects on national estimates.

What have we done here? We have used multiple control, interrupted time series in a pattern matching way. (I am glad to see Paul Rosenbaum now moving to use this kind of pattern matching language.) The principle is that differences between pretest time series can help observe a large part of the selection process, but the comparability is stable over time. Now, a causal hypothesis is created so as to have multiple implications in the data. Implications about when there should be a change, whether it should be a change in one series at one time or a different series another time, whether there should be a change when the treatment is operative and theoretically it should affect the outcome, or when you have a non-equivalent dependent variable that should not.

You want to assume that no alternative interpretation can be fit to this complex pattern of data that is being predicted. This is very often likely to be true, but not willy-nilly certain to be true. So we need a critical pattern-matching perspective that adds design elements to the interrupted time series to rule out the alternative threats to internal validity.

There are four studies in which people have deliberately designed a study to see if a randomized clinical trial and an interrupted time series with controls give the same results or give comparable results within a reasonable sampling error framework. I am not going to walk you through all four of them. Three are from medicine; one is in the social policy world. All four of them give comparable results, or certainly not statistically significant different results. They give the same results, basically. It is sobering that an interrupted time series and a randomized experiment deliberately designed to test the same hypothesis come up with the same results.



1. Schneeweiss et al. BMJ (2004)

- RCT = How does stopping insurance reimbursement for a drug affect expenditures on it over six months with six months of pretest.
- ITS with two cohort historical control series – from same calendar months in two prior years
- Similar direction and stat sig in RCT and ITS
- But causal estimates larger in ITS than ITT of RCT;
- But not for TOT in RE, given a 60% non-compliance rate with the randomized control group – using exemptions from the new regulation

2. Fretheim et al. (2012)

- Quality improvement intervention on physicians' prescription behavior
- Single ITS design, prior observations on treatment group
- IN RCT and ITS, similar direction of results, stat sig patterns and comparable estimates
- This is not so strong – only a single ITS and they were lucky that nothing related to outcome co-occurred with treatment

3. Roifman et al. (1987), as reanalyzed by Shadish, Hedges & Rindskopf (2011)

- RCT is cross-over design – for first 6 time points group A is treatment and B controls; for next 6 group B is the treatment and A the controls.
- ITS is within individual subjects – hierarchical linear model - time nested within patients nested within study group A or B.
- Results:

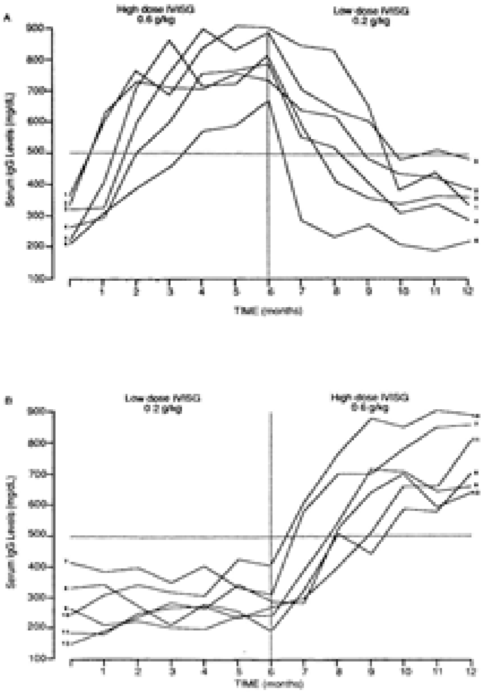
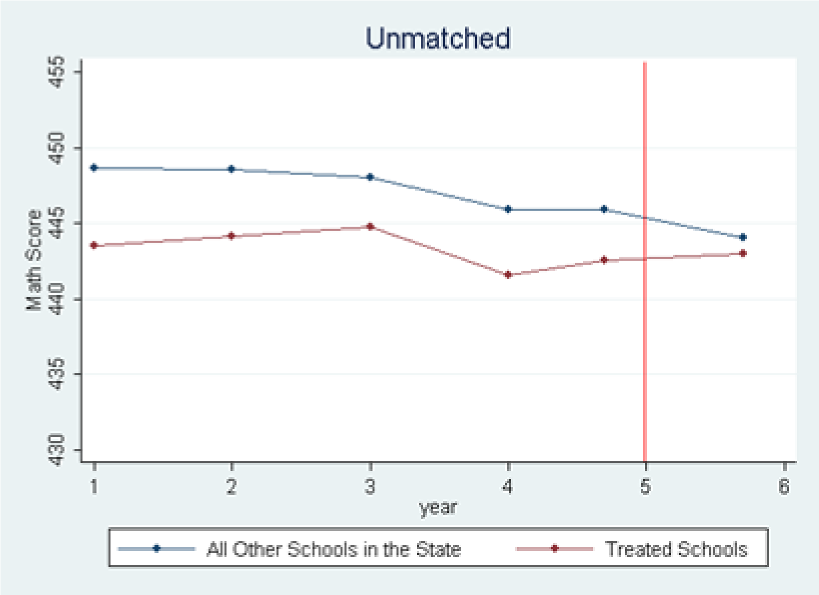


Fig 1—Serum IgG concentrations.

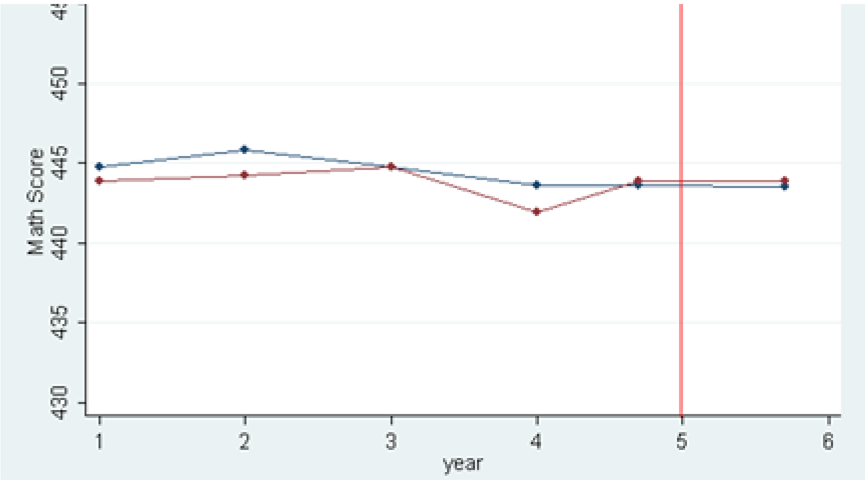
A, initial high dose; B, initial low dose.

4. Cook, St Clair & Hallberg (2012)

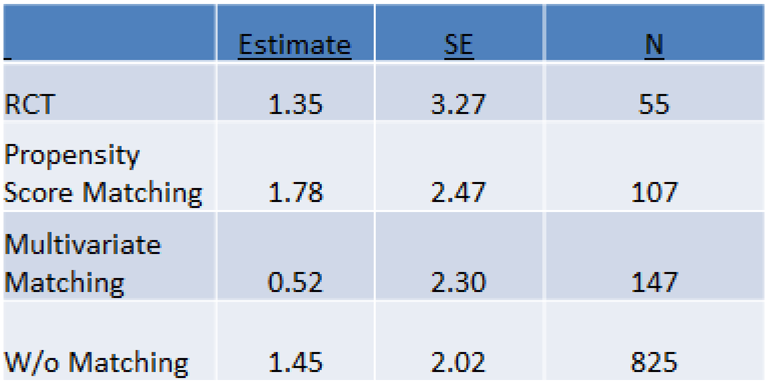
- RCT = does feedback to teachers about individual students increase achievement: 34 treatment/21 control schools with grades 3-5
- ITS = 5 prior school-level achievement means. Control schools are (a) all others in state with grades 3-5 and (b) matched schools with matching done several ways



4. Data with Matching

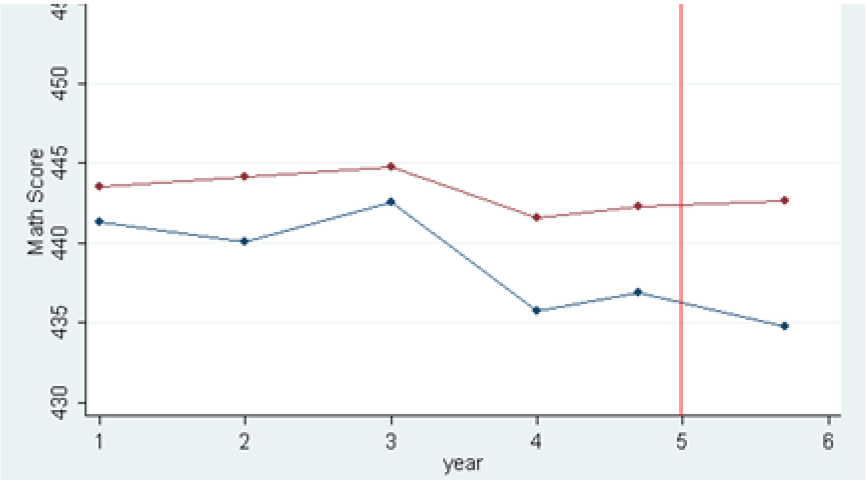


4. Causal Estimates Compared



	Estimate	SE	N
RCT	1.35	3.27	55
Propensity Score Matching	1.78	2.47	107
Multivariate Matching	0.52	2.30	147
W/o Matching	1.45	2.02	825

The Imperfect RCT



Allow me to give you my conclusions as questions. If I were George Box, a statistician of Wisconsin, I would be quaking. He has developed an aesthetically wonderful set of studies, most of which are predicated on at least fifteen, if not more, time points. You can collect brief interrupted time series like I showed you—often with only four, six, or eight prior time points. You can do this type of study because the data are in archives, or you can yourself collect the data. (There's a whole literature on single subject designs that use interrupted time series logic of the kind that I have talked about in order to test causal hypotheses. In fact, the Department of Education's What Works Clearinghouse has just added these within subject designs to its alimentarium of accepted practices.)

So, the big question is: If you can perform an interrupted time series will it satis*fice* about causation? (Notice I'm using Herb Simon's word.) I'm not saying, "Is it perfect?" I'm not saying, "Is it better than X, Y, or Z?" I'm saying that it satisfices about causation.

A smaller question is: Do you agree with me? Obviously, my answer to the first question is that I think it satisfies—if carefully done and carefully interpreted.

A not so little question is: Is it dangerous to use a single interrupted time series design in most of the applications that we will get to see? I would say yes. And it needs therefore to be a supplemented interrupted time series. I have shown you eight different ways by which you can supplement the single interrupted time series.

Another question is: Will the collection of data amenable to interrupted time series increase in the future? You bet. As we lay down more and more and more data and data sets and make them more readily available.

Another not so little question is: When you are anticipating modest effects (which should be base-rate expectation) are RCT's still preferred for internal validity due to their fewer and more transparent assumptions? I think so.

Big question: If we were to use internal and external validity together to appraise the quality of designs (like Chuck Manski, my colleague, is always pushing) might interrupted time series often be the equal of randomized control trials when there is a choice among them? Of course, there often is no choice among them—each is different for different kinds of questions. But if there is a choice and you weigh internal and external validity, might you be more inclined to favor interrupted time series?

I want us to stand here and imagine a randomized control trial of No Child Left Behind. And quickly you would see it does not lend itself to any kind of randomized clinical trial, even with a state or local waiver option, which has a great deal of external validity.

Thank you for your kind attention. I hope that I have not bored you. And I am now curious to get my friends' and critics' appraisals and answer questions with them from those of you in the audience. Thank you.

Judy Gueron: This talk is a great example of why Tom won this award. I want to thank him.

Now forgive me, Tom, but as I was both listening to you and looking at these slides quickly before coming here I was reminded of Humphrey Bogart's description of Johnny Rocko in the movie, *Key Largo*. Has anyone here ever heard of that movie? If you remember, Bogart describes the essence of Rocko, who is this notorious gangster, as someone who always wants more. (I want to apologize to Tom, but he's Rocko in this example.)

That is the essence of Johnny Rocko. That he's never satisfied. And Tom is like that. Wanting more and continuing to look for it, but in a good sense. He is always looking to solve the next problem and always in a brilliant, encyclopedic, and provocative way.

I want to use my brief time today to do something very atypical. Not really to respond to the specifics of the paper, although I think the issue that Tom is dealing with is a very critical one to all of us. But, rather, by citing Tom's answer to a question that I asked him in an interview earlier this year. I asked Tom, "Over the last forty years have your views changed on the value of random assignment? And if so, why?" And this is a somewhat edited version of his answer:

> My views have changed. I've always been a cautious advocate. I've been an advocate because it works and it gives you a clear causal inference. I've been cautious for two reasons. It doesn't speak to the theory of causation that's most prized in science, which is the identification of causal explanatory mechanisms. Most of the variables that have been manipulated have been practical, solution-driven interventions and they haven't been manipulations of central variables that explain human behavior in general. So that's one reason for caution.
>
> The second reason for caution, and this relates to this paper, is that there are a lot of questions for which you cannot use random assignment because it's totally unethical. You couldn't have randomly assigned people to No Child

> Left Behind, or states to getting Race to the Top funds in the many millions of dollars.
>
> So I have always wanted to keep boiling on the back pot of my stove better quasi-experimental methods for use when you could not do random assignment experiments. And I think that the sub-agenda has gotten lost in the keenness to do random assignment studies.
>
> So I've been a cautious advocate. Advocate because it is the best single method for probing one theory of causation. Cautious because that theory of causation is not the most prized one in science and cautious because one needs to have the quasi-experimental agenda bubbling for cases when random assignment was not possible.
>
> Have my views changed? They've changed in so far as I think there is a growing realization that some things can't be manipulated. Even among those in the public policy arena where micro-economists operate. Even among people like myself, who is a sympathetic advocate of random assignment.
>
> I think there's a sense that the alternatives need to be examined in greater depth now. The alternatives for the most part should be probed because we need to discover those that are most likely to reproduce the same results as randomized experiments. So, oddly enough, the development of the best quasi-experiments is going to reify and deify random assignment studies because the best quasi-experiments are those that reproduce the same results as random assignment. So one has to pay very special attention only to those quasi-experimental methods that meet this goal of reproducing experimental results.

End of quote.

Over the years I think all three of us commenters have been among the initially small, but more recently exploding, group of people who have worked hard to show the feasibility and the potential of random assignment studies. In the process we've answered many questions—even some important questions. I am sure, however, that all of us would say that we have hit up against many other questions that we could not answer.

Back in the early days of MDRC, our first board chairman, Eli Ginsberg, warned me against having discovered the hammer of random assignment and only looking for nails, with the implication that I would be looking for small bore issues on which it could be used. What I value deeply about Tom is that he's always looking at the big ones. And not just with dismay, but seeking creative solutions. I want to thank him for this and for the wise assistance that he has given us over the years on how to make our studies better. If you ever want a reviewer when you are launching some kind of a study (or midway through it), you want Tom Cook as that reviewer because he will (1) do the work, and (2) always bring something new to the discussion.

Johnny Rocko, lest we not loop back, may have wanted more and ultimately, you may recall, he ended up dead on the deck in the ship. Tom Cook continues to point us in the right direction, and I know that he will continue to do it and to triumph.

Thank you.

Rob Hollister: As you know I am a random assignment fanatic, so you cannot count on any balanced views from me. But Tom is a long time friend and this is a really highly deserved reward. He has been considering and contributing to a broad array of evaluation methods.

Tom has been an excellent advisory committee member. I want to second Judy on that. I have been on several technical advisory groups with him, and he always comes thoroughly prepared (unlike the rest of us), and he always has something good to contribute—and he is very balanced in his discussions, even to a radical like me. I remember him saying in one MDRC project when they were going to do some other thing, and he said, "Why do you want to throw in the toilet your real advantage of randomized trials?"

So he has had great work: the Comer Evaluation, he had a key role in getting Head Start RCT going, and certainly his resurrection of regression discontinuity designs is a signal accomplishment.

Now I have two quibbles that I thought I should state. First of all, randomized control trials are often a hard sell to sponsors and the program implementers. My worry is that, when Tom demonstrates that a quasi-experimental method is a reasonable substitute for an RCT, then program operators and sponsors who, I think erroneously, regard RCTs as terribly hard to do, expensive, and so forth, will grab onto those reasons immediately. Hence, I am cautious about presenting these alternatives and presenting them in a good light even though it is perfectly honest to do so.

Second, after having done careful studies like this one, he has concluded that this particular quasi-experimental method should be used when you meet certain conditions. If you look at those conditions, however, in most cases they are ones in which you could do an RCT. If you could meet those conditions for the quasi-experimental study you could do an RCT. I would rather see people do the RCT, for, as Tom himself says, it is a better method in many cases.

I admit that there are many situations in which you just simply cannot do RCTs. I have done a lot of research that way myself on major things. I've tried to stress that I think it's really important to inform the client of what you are not going to be able to deliver to them, that you're not going to be able to deliver to them a definitive causality relationship.

You still do the best you can, of course, and Tom's work is very helpful in saying what might be the second best in these kinds of situations.

Tom mentioned external validity in passing. My own view on external validity is that in general, it's not an attainable goal. I know two or three studies that I think could make that claim: the Job Corps study where you could really say it was a random draw from a known population and maybe the Head Start study as well. (Tom was very important in getting the Head Start study off the ground. A randomized trial after all those decades of which we had nothing really good in that line.)

I want to end my inadequate comments with two little stories. One was a joke that David Kershaw told me. It's one of the only jokes I've ever remembered. The Basque separatists were coming to Madrid to negotiate with the government about some kind of settlement. And they were mostly these rural people. They came to Madrid and they got put up in a hotel that had a revolving door entrance. And they'd never seen one like that before and they thought it was great fun and so they were running around and around and they all got crowded in and suddenly- Pff! - They all spilled out onto the street and the bus ran them over. So the moral of the story is: Never put your Basques in one exit.

The other one is: I recently went to Istanbul for the first time in my life. What you may not know, which I happen to know for a strange reason, is that Tom is a fanatic about rugs. So I was walking through the bazaar and they're all saying, “Want a rug here? I can give you a wonderful rug. Many centuries back here, and it’s in great shape,” and so forth. And so I said, “No, no I don’t.” And he said, “You’re American. Do you know this Cook … Cookie … Cook guy? He was just here a couple of weeks ago. And he came across this rug that I had. And the rug had a thread that ran in a straight line like this and then it jumped up. And it went in another line and he got tremendously excited. He said, ‘I must have this.’ And he says, ‘You see that, you know where that point where it jumps up is? That’s 1922, end of the Ottoman Empire and Ataturk comes in and the measure here is education—it goes steadily up like this.’” That’s my friend, Tom.

Becka Maynard: Wow, tough acts to follow. I have no movies, I have no jokes. So let me begin by congratulating Tom on this very well deserved award. Pete would be very pleased by the committee’s selection. And I want to echo Judy and Rob’s comments about Tom’s major contributions to this field.

This is, to me, a fascinating presentation and is bringing me back—I was going to say twenty-plus years—but Tom has said I probably need to be honest and go back more than that. Anyhow, back then there was a time when time series analysis was much more common than it is today, in part because of the data that were available for analysis, and in part because of the issues that were being addressed, and in part because we had not advanced into other methods.

What I want to do is offer three types of comments on your reappraisal and optimism about a larger role for interrupted time series in the evaluation methods arsenal.

The first set of comments relates to the fact that the emergence of more and better and better articulated databases greatly expands the opportunities to conduct time series analysis—and, if the data are there, they are going to be used. So it’s really timely to have this serious discussion of circumstances under which and how one should consider using these data. With a state longitudinal data system (for example, in education and the state welfare data systems out there), it is important that we think seriously about how to use these data.

You have identified opportunities for learning and circumstances where otherwise it would be necessary to continue acting on faith. I think it’s really important to start plugging that hole. My worry is that interrupted time series also may become a method in search of a use. I think that would be too bad. Like all statistical methods, the usefulness of interrupted time series depends on the research question and the extent to which the underlying assumptions are met. Tom's done, I think, a great job of making sure we are reminded of those restrictions.

I think it's easy to accept that interrupted time series analysis can provide reliable causal estimates when certain conditions prevail: for example, when you have long stable pretests, when you have immediate compliance, and when you have an immediate response function. Tom notes all of these.

The reality, however, is that these conditions often don't prevail, as Tom noted. And the workarounds for these condition failures of adding multiple comparisons and adding supplements to the comparisons, I think, are creative and are likely to work in some cases. If we are doing this analysis ex-post it seems reasonable and even clever and desirable to examine aggressively how well these various comparison groups are likely to work and under what conditions.

In this case, if we are talking about ex-post analysis, presumably the alternative to using imperfect information is having no information at all. So, I have a pretty easy time saying let us think hard about what we can learn from these data that are sitting there. We have no other way to address this question as opposed to when we are thinking about designing prospective studies.

For prospective studies, it seems much more likely that interrupted time series would be the method of choice only in those cases where it was in agreement that you were going to have a very large effect and only very obvious impacts would matter. That if you get little effects, we do not really care. I am thinking that this is something that is actually used a lot in management. Program operators use this kind of information. You make the change, results happen. If not, then so be it.

Having said that, I think it would be a great disservice to the field if this work leads to having the next cohort of Ph.D. dissertations demonstrating the causal prowess of interrupted time series. We have had prior waves (they come in about five year intervals) where everybody does the lambda coefficients; we do propensity score matching; we do fixed effects; we do RD designs. These are all methods in search of a use. Mind you, I believe firmly that each of these methods has merit. What I hope is that we can avoid the confusion over whether this is a method that should replace causal inference research that otherwise could be and would be better done using a randomized control trial. Or even some other non-experimental method.

The final thing I want to mention is that in much of the social sciences, the interventions we are studying have quite modest average effects or no effects at all.

When we have effects, they tend to be highly differentiated by subgroups or across settings. This is important. We focus on the average effects but often times we also find subgroup effects. Estimates that are based on administrative data are going to have less measurement error because of the large sample size than those from studies where they are prospective, smaller samples, etc. Hence, the estimates may be reliable measures of the outcomes for the generalizable population, but they may not have the same causal validity.

As a result, we may be estimating impacts of the intervention plus some other things that we don't know what they are. I think this goes back to what Rob was saying. I am not sure that internal validity versus external validity is really an even tradeoff there. Having an externally valid estimate of something that is not what you want, which is the estimate of the impact, does not really help you.

If the stakes are high, if the potential for confounds is high, and if implementation is prospective, it generally seems wise to opt for the experiment. Only if the experiment is not an option, do I think it would be profitable to explore the merits of interrupted time series or propensity score or RD or anything else—because there is a time and a place for all.

Regardless of the estimation method, I want to underscore something that I think Tom himself made very clear: It is that claims of causal inference should be carefully qualified regarding the strength of the warrant for the cause of validity and the context to which the estimates pertain and the context. I think we get sloppy about defining the context to which these causally valid estimates pertain.

Tom, this is really a great honor to be able to congratulate you in public on this very nice award.

Douglas Besharov: We have just under 15 minutes. Anyone who would like to ask a question, please come to one of the microphones. Please keep your questions relatively short so that we can have relatively short answers.

If you are willing, please tell us who you are, and direct your question to one member of the panel or the entire panel.

Steve Bell: I am Steve Bell from Abt Associates. My question is for Tom, but others may know of these. Are there examples that you have come upon where the case could be made that the conditions were right to use interrupted time series to discover an effect, somebody looked at the data and it flattened, that is, there was not what there appeared to be? That there were reasons in theory why some important effect may have really been hidden? Is it adequate to stop, at that point? Or is interrupted time series really a quick check for the big win? Then you may have to move on to something else.

Thomas Cook: That is a hard question. It is not something special about interrupted time series more than any other method you might use. If you do not get an effect at first, and you have reason to assume that there is some subgroup, some conditions under which you would get it and if hopefully you had specified that in advance, then you should go into the data and look for it.

We have been talking a lot here, because this is a public policy group, about using administrative data. If we were a group of people in psychology or in special education, we would have talked about interrupted time series where you collect your own data and you do it all the time and it would be normal to collect your own data and you do it over smaller time periods. Under that condition, where it does not take very long to do a study, people do sort of take null results, think it through, then they can quickly run another study to vary the conditions they think led to the no difference. In the administrative context, however, you are stuck by what's in the data set.

Eric Hanushek: I am Rick Hanushek of Stanford University. My question's for Tom. I'd like you to address your slight bastardization of the Hollister-Reeder hypothesis that cheap studies will drive out RCTs. My question is: How should we pay attention to the costs of the different studies when we're designing a set of evaluations?

Thomas Cook: It has been traditional in this organization, APPAM, to prioritize internal validity over any of the kinds of validity there are. If I were Chuck Menski, I think I would be pulling out my hair a bit over that because he has prepared to make tradeoffs between internal validity and other kinds of validity all the time—because there are different actors who have different interests. Getting the right population would be more important than getting an unbiased estimate on a population that I am not interested in.

So the key issue is if you can do a randomized clinical trial with (1) a population you are interested in, (2) at a time you want it generalized to, (3) with a treatment variant that is really the one you are going to implement in the real world later, (4) with outcome measures that correspond to the way you will do the monitoring later, (5) and if the randomized clinical trial is done in situations to which you want to generalize then, by all means, do a randomized clinical trial.

If, however, your inclination is: Let me do an RCT because of its primacy for internal validity then you've got to show to me and to anybody else that all those other conditions you want to generalize to are in fact the conditions met in your RCT. If not, you're in a tradeoff situation, rather than an inflexible, inviability of RCT as the only way to go.

Eric Hanushek: I actually want to push you in a slightly different direction. We are trained in this room to moan about the fact that the Institute of Education Sciences doesn't have enough money and the National Science Foundation's budget is not large enough to do all of our work. How should we make the larger tradeoffs that are involved in designing our research agenda?

Thomas Cook: Oh my goodness, now I understand your question. It is going to take me a week, Rick. Let's have a drink afterwards and talk about it. Does anybody else want to answer that?

Becka Maynard: Can I get it?

Thomas Cook: Okay. Becka is wiser than I.

Becka Maynard: I am not wiser, but I am maybe a little bolder on this point. I think we have many lost opportunities when we are going out into the field.

Take school improvement grants, for example. I would venture to say that there is probably not a measurable whisker of difference in the quality of many of those applications that were funded and not funded. We could probably draw a pretty big circle around a lot of these applications. What we chose to do was to fund down the slate until we ran out of money. We could have had a different attitude towards that. We could have a different attitude when we decide to change class size. We could change it progressively over time. Then, these studies would not be as expensive. We can then use the same administrative data that we are collecting for other purposes.

We have a lot of opportunities lost because we decide to give all the students in seventh grade a computer all in the same year. There is likely to be an implementation failure and other things. It would take us about 18 months, very low budget, and higher implementation if we did some randomized assignment to a phase-in of that policy or program. And we would know how effective that was, how much it costs, and what kinds of implementation challenges there were.

Therefore, I think we need to think differently about how we implement change and stop doing everything all at once.

Howard Bloom: Howard Bloom of MDRC. In a way this is going to be more a comment than a question and the question will justify the comment, I hope.

I have been attracted to interrupted time series since the 1970s and I am a periodic user—not a steady user. The logic of it is always attractive to me. It just makes sense. So, I applaud Tom on this presentation. I have also been a fan of yours since that time, because this kind of thinking is about scientific thinking.

You have a cause and an effect, with alternative hypotheses, things you think about when you try to rule them out one by one. With an RCT you sort of jump right past that stuff and sort of assume you have gotten past most of it. Now, obviously, people who know our work know we care about it completely. This is really important. I have seen interrupted time series work where there are buttons where you can show the jump in the time series and control series, the conditions of the two are similar. We might even show that different kinds of situations with different treatment programs, as you were saying, give different differences as we predicted.

You can really carry out a logical conversation that justifies the belief that this is causal. Recently, I saw this situation first hand, where we were trying to do some research where people were arguing: "Well if we can't do this, we're not going to help people doing this other stuff." So I do worry about that. Not quite as much, I suspect, as Rob. But getting the conditions under which these kinds of designs work and demonstrating empirically when they work and when they do not work, will be, I think, absolutely essential.

Therefore, my question, which I hope justifies my comment, is: How as a community and how as individual scholars do we set an agenda that shows: "When does it work? When doesn't it work?" Conceptually, theoretically, and empirically?

Thomas Cook: I agree that that is the agenda and that there is a large section of work which is trying to empirically test theories of the conditions under which matching works and does not work. There'll be a presentation about that tomorrow.

So, I agree with the agenda. But there's hardly anybody working on interrupted time series right now. There is no community of scholars. There is none of that social support and fun. (Science is about having fun as well as doing important work, right?)

I am working on it, with all the things that I have on my plate. The elucidation of the conditions under which time series works—whether it's through a Rubin causal model perspective, whether it's through a Campbell threats to validity issue, and then the empirical demonstration that you get the same results as randomized clinical trials—under the right conditions, which you don't under the wrong conditions—which is the proof at the end. That's a ten year agenda of quite a few people.

As I began the remarks today, all I want is for people to go away with an impression that maybe this is interesting. And if you do think it's interesting, come see me. Maybe we can work something out. Shameless…

Judy Gueron: What I am struggling with is your first bullet: "modest effects." Most studies find only modest effects. You have concluded by saying that RCTs are still the best thing. I think of my work over these close to many decades. There was one program (Youth Entitlement) that guaranteed a job to any youth in eighteen communities and it doubled the employment rate of minority youth. You had a hell of a treatment. We don't usually need to test such big treatments. I mean we have some comparison size, but in a sense that was superfluous. If you missed that finding, you would have been a moron.

That, however, is not the world we usually operate in. The problem is that our treatments are not that powerful. We are not doing these massive interventions. Some of what you're saying is that those massive interventions are often the ones where we haven't been able to do an RCT. When you think of the struggles to try to see the effect of the welfare reform of 1996 after the caseloads fell by half. It was clear from the studies that there was no question that caseloads fell by half. But the questions were: How much of the decline was because of the economy? How much was because of welfare reform? How much was because of the simultaneous EITC? Those were the questions that people grappled with, not that the caseload fell by half. So if it's those big policy changes, these alternative explanations had a range of estimates that were all over the place on how much each explanation contributed to the decline in caseloads.

Douglas Besharov: A closing thought. If past is prologue, you could look at what MDRC did over thirty years, what Mathematica did, what Abt did, that is to say, building not just experience, but an infrastructure to do complex field randomized trials. One could look at the full text of what Tom described and say you know that is really not just the work of solitary academic researchers in one university some place (and I'm afraid I'm in this category). This endeavor may need long-term, institutional support.

Thank you, Tom. Thank you, panel. Thank you all very much.