Abstract Title Page

Title: Evaluating the Performance of Repeated Measures Approaches in Replicating Experimental Benchmark Results

Authors and Affiliations:
Kevin McConeghy, University of Illinois Chicago
Coady Wing, Indiana University
Vivian C Wong, University of Virginia

Abstract Body

Background / Context:
Randomized experiments have long been established as the gold standard for addressing causal questions. However, experiments are not always feasible or desired, so observational methods are also needed. When multiple observations on the same variable are available, a repeated measures design may be used to assess whether a treatment administered at a known time results in changes in the outcome. The counterfactual is the post-intervention change that would have occurred if no treatment had been introduced. Given that one never observes the counterfactual outcome, researchers may employ a number of strategies to estimate the counterfactual. For example, they may assume that the pretest provides a valid estimate of the counterfactual outcome and compare pre-post changes in the outcome. However, validity threats, such as history and maturation, render most applications of this design "causally uninterpretable" (Shadish, Cook, and Campbell, 2002). Despite these weaknesses, the inclusion of thoughtfully constructed design features may bolster the causal interpretation of this approach. When a non-equivalent comparison group is available, a differences-in-differences (DD) approach may be used to rule out alternative explanations that coincidence with the introduction of the policy or program. The counterfactual here is the difference in the pretest-posttest measure between the treatment and comparison groups that may be subjected to the same confounders. When a pre-intervention time series is available, the researcher also may adjust for baseline trends to rule out maturation effects through an interrupted time series (ITS) design. In cases where a pre-intervention time series is available for both treatment and comparison groups, a comparative interrupted time series (CITS) design may be used to rule out history and maturation effects together.

Despite the popularity of repeated measures approaches for assessing policy impacts, questions remain about the empirical performance of these approaches in field settings. In WSC designs, the quasi-experimental approach is evaluated by comparing QE results with those from a benchmark design that shares the same treatment group. The purpose is to determine whether the QE approach can replicate results from the experiment. Thus far, at least five studies have examined the performance of repeated measures approaches – four in education contexts, and one in environmental policy (Bilfuclo, 2012; Ferraro & Miranda, 2014; Fortson et al., 2012; Somers, Zhu, Jacob, and Bloom, 2013; St. Clair, Cook, & Hallberg, 2013). Combined, these studies suggest that although we have reasons to be optimistic about the performance of such approaches, these results do not generalize to every domain of study in which repeated measures designs has been examined. As such, more work is needed to uncover the contexts and conditions under which repeated measures approaches produce unbiased results in field settings.

Setting:
The study employs experimental data from the Cash and Counseling Demonstration Project (Carlson, Foster, Dale, & Brown, 2007), which evaluated the effects of a "consumer-directed" care program on Medicaid recipients' outcomes. The data include monthly Medicaid expenditures for 12 months prior to the intervention (pretest), and 12 months after the intervention (posttest). Medicaid participants in Arkansas, New Jersey, and Florida were randomly assigned to treatment and control conditions, where the treatment consisted of Medicaid recipients selecting their own services using a Medicaid-funded account and the control consisted of local agencies selecting services for Medicaid recipients.

Purpose / Objective / Research Question / Focus of Study:
The purpose of this WSC is to examine the following three methodological questions:

(1) Does the simple interrupted time series (ITS) produce unbiased treatment effects, relative to an experimental benchmark?
(2) Do the comparative ITS and DID approaches produce unbiased treatment effects, relative to an experimental benchmark?
(3) Do the use of multiple in-state and out-of-state non-equivalent comparison groups rule out plausible threats to validity in the comparative ITS and DID designs?

For each comparison, we assess bias for short-run effects (4 months post-intervention), medium-run effects (9 months), and long-run effects (12 months).

Significance / Novelty of study:
The present study makes several contributions to the methodological and evaluation literature. Prior evaluations of the ITS design has utilized aggregate level data with fewer pretest time points (e.g. 2-3 pretest time points) and relatively stable and linear pre-intervention functions. The Cash and Counseling data allows us to evaluate the performance of repeated measures design approaches with 12 pre-intervention time points are available, and highly non-linear pre-intervention trends. We also examine the performance local and non-local comparison groups that are non-equivalent at baseline. Finally, we assess the performance of the design when treatment take-up was not immediate but gradually diffuse, allowing us to assess bias in short-, medium-, and long-run effects.

Statistical, Measurement, or Econometric Model:

Experimental Benchmark. One feature of this experiment was that treatment assignment occurred on a rolling basis, where enrollment took place over the course of several months from late 1998 until mid-2003. To address the fact that study participants were enrolled at different times, we centered the monthly time variable for all participants at the month prior to when treatment was introduced. Thus, the first follow-up month was coded 1, the second follow-up was coded 2, and so forth through the 12^{th} follow-up period. To estimate intent to treatment (ITT) effects for the RCT, we conducted longitudinal regression of month *t* outcomes on group specific trends and included a treatment/control dummy in the model. In essence, this is a DD approach applied to randomly assigned treatment and control groups. Time varying effects were computed using group specific year effects.

Observational arm. We assessed the performance of repeated measures approaches by constructing a series of QE designs using data from the Cash and Counseling experiment. To generate the simple ITS design, we selected the "WSC treatment sample" as nonelderly adults who were randomly assigned to the treatment condition and then deleted information from nonelderly adults in the experimental control group in the same state. We used two model specifications to estimate treatment effects – the linear model and the saturated model. The linear model assumes a pre- and post-intervention linear trend. For the saturated model, we allowed a more flexible specification of the model by imposing no functional form assumptions on the treatment response group.

To construct the QE designs for the DD and CITS approaches, we undertook the following steps. First, we selected the following three subgroups to be the WSC target samples: (1) nonelderly adults, (2) elderly and nonelderly females, and (3) all adults in each of the three states. Next, we deleted control group data for the WSC target groups in the state. Then, we constructed non-equivalent comparisons for the three WSC target samples. For the WSC nonelderly group, we used data from in-state elderly experimental controls; for the WSC female group, we used control group data from in-state men; and for the adult treatment sample, we used control group information from all adults in the other two states in the Cash and Counseling experiment. Finally, we deleted experimental treatment group information for our constructed non-equivalent comparisons. Table 1 summarizes the ITS and non-equivalent comparison group approaches described above, and Table 2 describes sample sizes for each treatment and comparison case.

To estimate ITT effects for repeated measures approaches with comparison time series, we used four model specifications. For the differences-in-differences approach ("DD"), we compared changes in pre- and post-intervention means for treatment and comparison groups, which we estimated using a regression model with an interaction term between the "post-intervention" and "treatment" indicators. We also examined the performance of three specifications of the CITS approach that allow for varying degrees of flexibility in the linear trend. In the "Trend" model, we assume that treatment and comparison groups have different intercepts and slopes, but that pre- and post-test trends are the same within each group. In the "pre-post" specification, treatment and comparison groups are assumed to have the same slopes, but we allow for slopes to change post-intervention, and for the treatment group to have a different intercept jump. Finally, for the "Group-specific trends" model, we allow for treatment and comparison groups to have their own intercepts and slopes pre- and post-intervention.

Assessing correspondence between experimental and non-experimental results. To assess correspondence between experimental and non-experimental results, we began by creating a measure of standardized absolute bias in the QE and RCT results: $\hat{B}(t) = \frac{/\hat{\tau}(t)_{QE} - \hat{\tau}(t)_{RCT}/}{s}$. This was calculated by taking the absolute difference in the estimated QE and RCT results, and then dividing the difference by the standard deviation of the pretest outcome. Our second correspondence metric was the root mean squared error (RMSE) of the QE and RCT results, such that: $RMSE(t) = \sqrt{\frac{1}{N}\sum_{i=1}^{500}(\hat{\tau}(t)_{QE} - \hat{\tau}(t)_{RCT})^2}$. To implement this approach, we constructed 500 bootstrap replicates of each estimate, centered the QE bootstrap estimate around the experimental benchmark result, and squared the deviations. Next, we computed the average of the squared deviations across bootstrap replicates to calculate the mean squared error, and took the square root of the mean squared error. The benefit of this correspondence metric is that it takes account of both bias and efficiency of the QE approach (as compared to the RCT result) in the same framework.

Finally, we devised a method for interpreting and synthesizing the multiple QE estimates produced across states. This was because for each time period (4 months, 9 months, and 12 months), the QE design produced two treatment effect estimates for the simple ITS design ("linear" and "saturated" models), three DD estimates (with in-state elderly, in-state male, and out-of-state comparisons), and nine CITS estimates (3 non-equivalent comparisons x 3 models

(Trend, Pre-Post, Group-specific Trends). Thus, for each of the three states, the QE design produced 84 treatment effect estimates, or 252 QE estimates across all three states.

To make sense of the effect estimates, we synthesized results in the following regression: $Y_i = \alpha_0 + \sum_1^r \alpha_{1r} ITS_{ir} + \sum_1^q \alpha_{1q} ElderlyC_{iq} + \sum_1^q \alpha_{2q} MaleC_{iq} + \sum_1^q \alpha_{3q} StateC_{iq} + \varepsilon_i$. Here, Y_i is standardized absolute bias or the RMSE at t = 4 months, 9 months, or 12 months for estimate i = 1….252. r indexes the two ITS estimates, q indexes the four model specifications for the repeated measures designs with non-equivalent comparisons (DD, trend, pre-post, group-specific models), and ITS_{ir}, $ElderlyC_{iq}$, $MaleC_{iq}$, and $StateC_{iq}$ are a series of dummy indicators that are equal to 1 if the estimate was produced by the corresponding comparison group, and 0 if otherwise. α_0 is interpreted as the standardized absolute bias or RMSE in the RCT (depending on the outcome used), and the coefficients are interpreted as the amount of bias or the RMSE produced by the QE approach across three states at 4 months, 9 months, or 12 months.

Findings / Results:
Table 3 provides a balance table to demonstrate that groups were equivalent at baseline, and Table 4 presents experimental benchmark results for each WSC target group, in each state, at the 4 month, 9 month, and 12 month follow-up periods. Overall, consumer-directed care resulted in significant increases in Medicaid expenditures for all subgroups at each follow-up time period. Table 5 shows results from our regression of standardized absolute bias and the root mean squared error of the QE and RCT results. Looking across the table of results, we find that the QE methods performed better at the earlier time points than at the later time points that required more extrapolation of the regression function. Second, the table shows that more complex modeling of trends tended to reduce bias, but increased the root mean squared error. In other words, the models that allowed for maximum flexibility also tended to produced less precise estimates. Third, the WSC results indicate that within-state comparisons performed better than cross-state comparison groups, and that gender-based comparisons produced less bias than age-based comparisons. Fourth, the cross-state comparisons performed relatively well when the model was adjusted for more complicated trends.

Conclusions:
As always, external validity is a concern with WSC results. The negative view holds that all WSC demonstrates is the performance of a non-experimental method in one situation. However, WSCs may also help identify contexts and conditions under which a method works better for particular subject areas, types of data, and time horizons. This WSC evaluation provides information about the performance of repeated time series methods when multiple pre-intervention data are available for modeling complex functional forms to estimate short-run, medium-run, and long-run effects. As states continue efforts to collect longitudinal data on student and school achievement, these results provide important guidance on performance of repeated measures methods in field settings when multiple pre-intervention time points are available.

Appendix A. References

Bifulco, R. Can Nonexperimental Estimates Replicate Estimates Based on Random Assignment in Evaluations of School Choice? A Within-Study Comparison. *Journal of Policy Analysis and Management, 31*(3), 729-751.

Carlson, Foster, Dale, & Brown (2007). Effects of Cash and Counseling on Personal Care and Well-being. *Health Services Research*. 42(1), 467-487.

Ferraro, P. & Miranda, J. (under review). Can Panel Designs and Estimators Substitute for Randomized Control Trials in the Evaluation of Social Programs?

Fortson, Kenneth, Natalya Verbitsky-Savitz, Emma Kopa, and Philip Gleason. "Using an Experimental Evaluation of Charter Schools to Test Whether Nonexperimental Comparison Group Methods Can Replicate Experimental Impact Estimates." NCEE Technical Methods Report 2012-4019. Washington, DC: National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, U.S. Department of Education, 2012.

Somers, M., Zhu, P., Jacob, R., & Bloom, H., (2013). The validity and precision of the comparative interrupted time series design and the difference-in-difference design in educational evaluation. *MDRC working paper in research methodology*. New York, NY.

Shadish, W.R., Cook, T.D., & Campbell, D.T. (2002). Experimental and quasi-experimental designs for generalized causal inference. Boston, MA: Houghton Mifflin.

St. Clair, Cook, & Hallberg (in press). Examining the Internal Validity and Statistical Precision of the Comparative Interrupted Time Series Design by Comparison with a Randomized Experiment.

Appendix B. Tables and Figures

Table 1. Summary of research designs and comparison groups

	Treatment participants	In-state comparison	Out-of-state comparison
Simple ITS	Nonelderly adults randomly assigned to treatment	Nonelderly adults randomly assigned treatment	
C-ITS/DID 1	Nonelderly and elderly adults randomly assigned to treatment		Elderly and nonelderly adults in other two states randomly assigned to control group
C-ITS/DID 2	Nonelderly adults randomly assigned to treatment	Elderly adults randomly assigned to control group	
C-ITS/DID 3	Elderly and nonelderly females randomly assigned to treatment	Elderly and nonelderly males assigned to control group	

Table 2. Sample size by experimental condition

	Arkansas		Florida		New Jersey	
	Treatment	Control	Treatment	Control	Treatment	Control
Nonelderly	279	277	456	458	404	413
Elderly	725	727	453	451	467	471
Children			501	501		

Table 3. Balance Table for RCT

	Arkansas		New Jersey		Florida	
	Treatment	Control	Treatment	Control	Treatment	Control
Age characteristics						
Nonelderly	0.28	0.28	0.47	0.47	0.55	0.55
	(0.45)	(0.45)	(0.50)	(0.50)	(0.50)	(0.50)
Elderly	0.72	0.72	0.53	0.53	0.45	0.45
	(0.45)	(0.45)	(0.50)	(0.50)	(0.50)	(0.50)
Female	0.78	0.78	0.74	0.72	0.61	0.63
	(0.42)	(0.42)	(0.44)	(0.45)	(0.49)	(0.48)
Race						
Hispanic	0.01	0.01	0.36	0.36	0.26	0.30
	(0.11)	(0.10)	(0.48)	(0.48)	(0.44)	(0.46)
White	0.61	0.60	0.53	0.56	0.73	0.74
	(0.49)	(0.49)	(0.50)	(0.50)	(0.44)	(0.44)
Black	0.33	0.34	0.38	0.36	0.23	0.22
	(0.47)	(0.47)	(0.49)	(0.48)	(0.42)	(0.42)
Other	0.06	0.07	0.09	0.08	0.04	0.04
	(0.24)	(0.25)	(0.29)	(0.27)	(0.19)	(0.19)
Relative Health Status						
Fair	0.31	0.31	0.35	0.38	0.30	0.32
	(0.46)	(0.46)	(0.48)	(0.49)	(0.46)	(0.47)
Poor	0.47	0.51	0.45	0.40	0.26	0.26
	(0.50)	(0.50)	(0.50)	(0.49)	(0.44)	(0.44)
Compared to Past Year						
Worse	0.54	0.54	0.49	0.45	0.33	0.33
	(0.50)	(0.50)	(0.50)	(0.50)	(0.47)	(0.47)
N	1004	1004	861	869	909	908

Omnibus F-test results showed equivalence of treatment and control group on baseline characteristics for all three states

Table 4. RCT Benchmark Results

Group	Follow-up	Arkansas	Florida	New Jersey
Nonelderly	4 month	569.03*	1908.1*	618.96*
		(142.47)	(485.79)	(186.10)
	9 month	514.5*	3160.04*	1028.22*
		(113.13)	(556.88)	(196.73)
	12 month	441.19*	3298.94*	975.39*
		(145.07)	(554.07)	(220.37)
Women	4 month	488.98*	914.03*	580.78*
		(53.07)	(319.77)	(134.92)
	9 month	445.51*	1295.97*	1052.45*
		(48.20)	(330.62)	(138.96)
	12 month	345.99*	1368.8*	924.05*
		(50.57)	(340.38)	(160.09)
Full sample	4 month	516.83*	1392.71*	523.93*
		(44.29)	(302.30)	(104.24)
	9 month	454.05*	2049.14*	989.85*
		(41.46)	(332.90)	(108.57)
	12 month	368.3*	2024.69*	921.29*
		(54.36)	(326.83)	(117.79)

* indicates significance at the .05 level

Table 5. Summary of results from QE designs and comparison groups examined

Research Design	Bias M4	Bias M9	Bias M12	RMSE M4	RMSE M9	RMSE M12
Linear ITS	0.14	0.22	0.18	239.69	516.35	484.16
Saturated ITS	0.08	0.21	0.26	131.09	475.13	668.94
Within State Age Designs						
DID	0.18	0.10	0.06	342.66	191.84	91.60
DID + Trend	0.11	0.11	0.09	342.86	199.77	92.18
DID + Pre-Post Trend	0.18	0.10	0.06	338.08	189.71	90.01
DID + Group Pre-Post Trend	0.08	0.11	0.16	140.90	277.43	412.53
Within State Gender Designs						
DID	0.04	0.09	0.09	-22.65	90.81	75.71
DID + Trend	0.05	0.11	0.09	-26.53	96.20	80.54
DID + Pre-Post Trend	0.04	0.09	0.09	-27.42	95.92	79.69
DID + Group Pre-Post Trend	0.02	0.05	0.10	-5.99	129.61	341.38
Cross State Designs						
DID	0.28	0.38	0.34	398.40	741.16	671.27
DID + Trend	0.20	0.35	0.34	247.25	610.91	593.20
DID + Pre-Post Trend	0.28	0.38	0.34	409.78	752.41	682.26
DID + Group Pre-Post Trend	0.12	0.16	0.17	243.00	668.96	911.28
Constant (RCT)	-0.00	-0.00	-0.00	177.71	186.21	196.90
Observations	60	60	60	60	60	60
R^2	0.658	0.722	0.637	0.282	0.299	0.317