Prior Test Scores Do Not Provide Valid Placebo Tests of Teacher Switching Research Designs

Raj Chetty, John Friedman, and Jonah Rockoff

September 2014

Abstract

Chetty, Friedman, and Rockoff (2014) [CFR] evaluate the degree of bias in teacher value-added (VA) estimates using a "teacher switching" research design, regressing changes in mean test scores across cohorts on changes in mean teacher VA. Recent studies (Kane, Staiger, and Bacher-Hicks 2014, Rothstein 2014) have found that regressing changes in mean scores in the *prior* grade on changes in mean VA also yields a positive coefficient, a result we confirm in our data. At first glance, this "placebo test" seems to suggest that CFR's quasi-experimental teacher switching design is invalid: how can changes in current teacher quality affect students' past test scores? The problem with this placebo test is that teacher value-added is itself estimated from prior test scores. Hence, changes in current teacher VA will be mechanically correlated with changes in prior scores even when a research design is valid. We show that simple variants of the lagged score placebo test that remove this mechanical effect uncover no evidence of a relationship between changes in teacher VA and lagged test scores. We conclude that (1) the correlation between changes in teacher VA and prior test scores is not an informative placebo test and (2) the original CFR research design and results are valid.

Chetty, Friedman, and Rockoff (2014) [CFR] evaluate the degree of bias in teacher value-added (VA) estimates by regressing changes in mean test scores across cohorts on changes in mean teacher VA. They find that the coefficient in such regressions is close to 1 and conclude that the degree of forecast bias in VA estimates is small.

A surprising feature of the CFR design is that regressing *prior* scores on changes in mean VA also yields a statistically significant, positive coefficient (Kane, Staiger, and Bacher-Hicks 2014, Rothstein 2014). At first glance, this correlation appears troubling for the validity of the design, as changes in current teacher quality cannot cause changes in past test scores. In this note, we explain why the correlation between changes in prior (lagged) scores and current teacher VA is consistent with the assumptions underlying CFR's research design and in fact to be expected given the way in which VA is estimated.

We make two points, one conceptual and one empirical. Conceptually, we show why CFR's teachers switchers design will in general reveal "effects" of changes in teacher VA on prior scores even if the data generating process is one in which the research design is valid. Because teacher VA is estimated using data from students in the same schools in previous years, teachers will tend to have high VA estimates when their students happened to do well in prior years. Regressing changes in prior test scores on changes in teacher VA effectively puts the same data on the left- and right-hand side of the regression, mechanically yielding a positive coefficient.

Empirically, we show that accounting for such mechanical effects eliminates the correlation between changes in lagged scores and current teacher VA. In particular, excluding teachers who follow students across grades and controlling for school-year-subject shocks to test scores reduces the coefficient on lagged scores to zero, while leaving the coefficient on mean current scores unchanged. In addition, placebo tests that do not directly re-use the same data used to estimate VA – e.g., estimating VA only using data from future years or looking at effects on test scores in a different subject – uncover no evidence of a correlation with prior test scores.

We conclude that (1) lagged scores do not provide a valid placebo test of the CFR research design and (2) correcting for mechanical biases in the lagged score placebo test yields a "placebo" coefficient of zero without changing the estimates of forecast bias reported by CFR. Hence, the correlation between changes in prior scores and current teacher VA does not raise concerns about the validity of CFR's teacher switching research design or the conclusions that follow from it.

The note is organized as follows. Section I documents the correlation between changes in lagged scores and current VA. Section II shows that removing teachers who follow students across grades and school-year-subject level shocks eliminates this correlation. Section III implements a set of alternative placebo tests that do not re-use the data used for estimation of VA. Section IV concludes.

I. Correlation Between Changes in Lagged Scores and Current VA

Table 1 shows estimates from six separate regressions. Each cell in the table shows the coefficient and standard error on the change in mean VA across cohorts from a separate regression. In the first row, the dependent variable is the change in the current score; in the second row, the dependent variable is the change in the lagged score.

Column 1 of Table 1 replicates the baseline specification in CFR. The coefficient in Col 1, Row 1 for current scores matches the 0.974 estimate in CFR Table 4, Col 1. The coefficient in Col 1, Row 2 for lagged scores shows that we find a "placebo effect" of changes in current teacher VA on *lagged* scores of 0.22. Kane, Staiger, and Bacher-Hicks (2014) document results very similar to those in Column 1 for both current and lagged scores in LAUSD and Rothstein (2014) documents analogous results in North Carolina.

II. Do Lagged Scores Provide a Valid Placebo Test?

In general, a placebo test requires that there is no channel through which the treatment variable might covary with the placebo outcomes when the research design is valid. In our application, the treatment variable (teacher VA) is estimated from student test scores in other school years. This basic feature of teacher VA estimates makes it unlikely that using the same test scores in other years will satisfy the basic criterion for a valid placebo test.

More specifically, there are a variety of data generating processes in which the switchers design is unbiased, yet lagged score changes are correlated with changes in current VA. In this section, we focus on two such examples that are particularly important in our data: teachers following students across grades within a school (as occurs in our data), and school-year shocks (as documented by Kane and Staiger 2002).

Example 1: Teachers switch grades and sometimes teach members of the same cohort of students, or even the same students, as in previous years. This creates a direct relationship between VA estimated from prior years' data and prior test scores.

A teacher moving to the following grade (and teaching the same cohort, or even the same students, in year *t* as in year *t-1)* creates a correlation between changes in teacher VA and lagged scores. This is because noise in students' lagged scores directly enters VA estimates and because teachers have direct treatment effects on prior scores.

It is easiest to see these issues with an example. Consider the relationship between changes in math test scores and math teacher VA from 1994 to 1995 in 5th grade in a given school. Suppose a teacher with high estimated VA leaves 5th grade after 1994 and is replaced by a teacher with average VA; assume there are no other changes in the teaching roster. We know that the high-VA teacher who departed did *not* teach the children who were in grade 5 in 1995 when they were in 4th grade in 1994 (because she taught 5th grade in 1994). However, she may have taught the children who were in grade 5 in 1994 when they were in 4th grade in 1993. As a result, the high VA of the departing teacher is positively correlated with lagged test scores of the cohort that reaches 5th grade in 1994, but not the test scores of those who reach 5th grade in 1995. This effect makes lagged (4th grade) test scores fall on average across the two cohorts. Since (by construction) teacher VA is also falling in this example, there is a positive correlation between changes in lagged (4th grade) scores across the two cohorts and mean teacher VA.

It is important to distinguish between two separate channels that drive this correlation. The first channel is fluctuations in student test scores that are not related to the persistent component of teacher value-added, i.e., noise in student test scores. The teachers in 5th grade in 1994 could have higher estimated VA simply because her students in 4th grade test in 1993 did particularly well by chance (e.g., because the curriculum in the school happened to be well aligned with the test questions that year). This creates a mechanical correlation between lagged scores and VA estimates but has no bearing on our estimate of forecast bias using current test scores.

Second, the correlation could be driven by teacher treatment effects. If the 5th grade teachers in 1994 were of truly high quality, they would affect the performance of 4th graders in 1993 (because some of them taught 4th grade in 1993), but *not* the 4th graders in 1994 (because we know they are teaching 5th grade in 1994). Note that, in contrast to the first channel, the direct treatment effect of teachers in prior grades could potentially create bias in our estimate of forecast bias, as having better teachers in prior school years can increase current scores. The degree of bias will depend upon the rate of fade-out in the sample where a teacher teaches the same child twice. The magnitude of this bias is an empirical question that we analyze below.

The most direct way to address the problem of teachers following students across grades is to avoid identifying the change in mean VA from teachers who switched into the current grade from the previous grade. To do so, we define a modified version of the change in VA that excludes teachers who switch to the current grade from the previous grade when constructing mean VA in the current and previous cohort. We then instrument for the actual change in mean VA with the no-follow estimate. The first-stage coefficient is 0.95, as expected given that very few teachers follow their students across grades

Column 2 of Table 1 shows that eliminating the small piece of variation due to teachers who follow students reduces the coefficient on the lagged score by approximately 50%, to 0.116. Hence, roughly 50% of the correlation between lagged scores and current teacher VA appears to be directly driven by teachers following students across grades.

The coefficient on the *current* score is 0.936 in Column 2. The reduction in the coefficient across the specifications may be due to netting out the treatment effect of having better teachers in prior years (the source of bias noted above). However, the change in the coefficient falls within one standard error of the point estimate, so the magnitude of this bias appears to be negligible relative to other sources of error.

Next, we turn to a second source of correlation between lagged scores and changes in VA.

Example 2: There are large school-year level fluctuations in test scores, as documented in Kane and Staiger (JEP 2002), and these fluctuations vary by subject. This again generates a mechanical positive correlation between VA estimates and lagged performance.

Shocks to test scores at the school-year-subject level create correlations between changes in lagged scores and current VA even if no teacher teaches the same child twice. For example, suppose that 1993 was a particularly good year in a school-subject cell because the math curriculum that year lined up closely with the problems on the end-of-year tests. As a result, lagged scores in 1994 will be unusually high relative to lagged scores in 1995. Moreover, average estimated teacher VA will be higher in 1994 than in 1995 because of teacher turnover. Thus changes in VA from 1994 to 1995 will be correlated with changes in lagged scores.

One simple way to address this problem is to include school-subject-year fixed effects in the quasi-experimental regressions. Such specifications net out any correlated fluctuations in scores and mean VA across all students in a given school-subject-year. In Column 3 of Table 1, we control for school-year-subject shocks by including school-year-subject fixed effects in the specification implemented in Column 2. This reduces the lagged score coefficient to a statistically insignificant -0.023, while leaving the current score coefficient at 0.942.[1] These results indicate that whatever shocks drive the lagged score coefficient in Column 1 do not violate the assumptions of the original CFR quasi-experimental design.

One may be tempted to control for changes in lagged scores to address the issues raised above rather than eliminating the correlation between changes in VA through the two corrections above. This approach yields invalid estimates of forecast bias for two reasons. First, because some teachers follow students across grades, lagged scores are endogenous to teacher VA, and including an endogenous control naturally generates bias. Second, when one controls for lagged scores, one effectively applies the cross-sectional correlation between lagged and current scores to translate the correlation between lagged scores and teacher VA to a predicted changes in current scores. However, to the extent that the correlation between teacher VA and lagged scores is driven by noise in student's scores – as suggested by the results above – the persistence of these shocks will not be the same as the persistence estimated from the relationship between lagged and current scores in the cross-section. Intuitively, if lagged scores correlate with VA estimates purely due to transitory shocks, then the predicted persistent impact on current scores would be 0.

[1] In contrast, if we identify the model based on the variation in mean VA between school-year-subject cells, we find a large, significant coefficient on lagged test scores, while the coefficient on current scores remains very similar to the baseline estimate. This confirms that much of the correlation between teacher VA and lagged scores is due to shocks at the school-year-subject level.

Because of these issues, the most definitive approach to assessing whether the correlation between VA and lagged scores generates bias is to directly eliminate the source of the correlation as we do above and evaluate whether the estimates of forecast bias change.[2]

Our interpretation of the preceding findings is not that the original CFR design needs to be modified to include school-year-subject fixed effects in order to be valid. Rather, the lesson is that the lagged score placebo test is not informative about violations of the identification assumptions in the original design. We present further evidence to substantiate this interpretation in the next subsection.

III. Alternative Placebo Tests that do not Re-Use Data

Our main conceptual point is that correlating changes in mean teacher VA with prior test scores is not a valid placebo test because it re-uses the lagged scores to construct VA. In Table 2, we implement three variants of the lagged score correlation test that separate the data used to estimate VA and implement the placebo test. As expected, we find that these three tests yield "placebo" coefficients that are not significantly different from 0.

In Column 1 of Table 2, we use only *future* scores to estimate each teacher's value-added. For instance, we estimate a teacher's value-added in 1995 using only data from 1996 forward, omitting any data from 1994 or before. We then replicate the specification in Column 2 of Table 1 (dropping teachers who follow students, but not including school-year-subject fixed effects) using this alternative measure of value-added. Having eliminated the direct relationship between VA and test scores in prior years, we find no statistically significant relationship between lagged scores and changes in VA in this specification.

In Column 2 of Table 2, we instrument for the change in mean VA across cohorts ($\Delta Q_{sgt} = Q_{sgt} - Q_{sg,t-1}$) with the mean VA from the second cohort (Q_{sgt}) and replicate the baseline specification in Column 1 of Table 1. Importantly, the variation in this specification does not include those teachers leaving the grade, whose VA is most likely estimated from the same lagged score data used in the placebo test. Once again, in this "cleaner" version of the placebo test, we find no correlation between changes in mean VA and changes in lagged test scores.

[2] In the CFR dataset, the two solutions proposed above are adequate to eliminate the correlation with the lagged scores. In other datasets, it may be necessary to implement additional corrections to obtain a valid placebo test using lagged scores. For instance, some teachers may not have switched grades but may have the same student twice because of grade repetition. Others may have taught the same students two years ago rather than last year. Since students switch schools, introducing current school-by-year fixed effects may not fully account for the effect of shocks in previous years that affect teacher VA estimates. These examples underscore the broader point of this note: the variety of channels through which lagged scores may be correlated with current VA estimates makes it difficult to use lagged scores for placebo tests of VA estimates.

Finally, in Column 3 of Table 2, we conduct the placebo test using lagged test scores in the *other* subject. We regress changes in lagged scores on both the changes in mean VA in the own subject and the other subject, pooling all grades. We find that changes in lagged scores in the same subject are strongly related to changes in mean VA in the same subject, confirming the result in Column 1 of Table 1. But lagged scores in the *other* subject are unrelated to changes in VA. Since students' test scores are highly correlated across subjects, this result suggests that the lagged score correlation in the own subject is not due to changes in latent student ability but rather arises from correlated shocks that enter VA estimates in the same subject. More generally, this test combined with the results in Table 5 of CFR (2014a) shows that any violation of the design would have to occur through a channel that does not affect lagged (or current) scores in the other subject. We believe that plausible omitted variables are unlikely to have such properties.

IV. Conclusion

The analysis in this note yields two lessons. First, "placebo tests" of value-added models using lagged test scores will in general yield estimates that are significantly different from zero even when the underlying research design is valid. The conceptual problem with using lagged scores is that teacher VA is estimated using data from the same school-year cells that students were previously in, leading to mechanical correlations between VA estimates and students' past performance. Hence, lagged scores do not provide a robust placebo test for value-added estimates despite their appeal in other applications.

Second, correcting for these mechanical correlations – either by directly eliminating the source of the correlation as in Section II or using separate data to estimate VA and implement the placebo test as in Section III – eliminates the correlation between lagged test scores and changes in current teacher VA. While these corrections affect the results of the "placebo test", they do not significantly affect the estimates of forecast bias originally reported in CFR. We conclude that the original CFR teacher switching quasi-experimental design and estimates of forecast bias are valid.

References

Chetty, Raj, John Friedman, and Jonah Rockoff. 2014. "Measuring the Impact of Teachers I: Evaluating Bias in Teacher Value-Added Estimates" *American Economic Review* 104(9): 2593-2632.

Kane, Thomas J., and Douglas O. Staiger. 2002. "The promise and pitfalls of using imprecise school accountability measures." *The Journal of Economic Perspectives* 16.4: 91-114.

Kane, Thomas J., Douglas O. Staiger, and Andrew Bacher-Hicks. 2014. “Validating Teacher Effect Estimates using Between School Movers: A Replication and Extension of Chetty et al." Harvard University Working Paper.

Rothstein, Jesse. 2014. "Revisiting the Impacts of Teachers." UC-Berkeley Working Paper.

TABLE 1

Coefficients on Change in Mean Teacher VA in Switchers Design
Using Current vs. Lagged Score as Dependent Variable

Dep Var:	Baseline w/ Year FE's (1)	No Followers w/ Year FE's (2)	No Followers w/ School-Year-Subject FE's (3)
Δ Current Score	0.974	0.936	0.942
	(0.033)	(0.036)	(0.043)
Δ Lagged Score	0.226	0.116	-0.023
	(0.033)	(0.036)	(0.042)

Note: Each cell of this table reports estimates from a separate regression. The dependent variable in row 1 is the change in current scores, while the dependent variable in row 2 is the change in lagged scores. Column 1 is the baseline specification in CFR (2014, Col. 1 of Table 4). Colum 2 reports 2SLS estimates, instrumenting for the change in mean VA with the changes in mean VA excluding teachers who switch from the previous grade to the current grade. Column 3 replicates 2 including school-year-subject fixed effects. See notes to Table 4 in CFR (2014) for further details on these specifications.

TABLE 2

Effect of Changes in Mean Teacher VA on Lagged Test Scores:
Alternative Placebo Tests that do Not Re-Use Data

	Dependent Variable: Change in Lagged Test Scores		
	Leave Out All Prior Years	Entrants Only	Other Subject Score
	(1)	(2)	(3)
Change in Mean VA across cohorts	0.052 (0.050)	0.030 (0.076)	0.013 (0.028)

Note: This table shows estimates from three regressions of changes in mean lagged test scores on changes in mean teacher VA. Col. 1 replicates Row 2, Col 2 of Table 1, leaving out all prior years when estimating teacher VA. Col. 2 replicates Row 2, Col. 1 of Table 1, isolating the portion of the change in mean teacher VA that comes from new entrants rather than leavers. This specification reports 2SLS estimates, instrumenting for changes in mean teacher VA across cohorts ($\Delta Q_{\mathrm{sgt}} = Q_{sgt} - Q_{sg,t-1}$) with the mean VA in the second cohort (Q_{sgt}). In Col. 3, we replicate Row 2, Col. 1, of Table 1 but include changes in mean VA in both the own and other subject. We report the coefficient on the change in mean VA in the other subject in this table. The number of school-grade-subject-year cells is 59,774 in Col. 1 and Col. 2 and 58,761 in Col. 3. See notes to Table 1 and Table 4 in CFR (2014) for further details.