RESPONSE TO ADLER (2014) REVIEW OF "MEASURING THE IMPACTS OF TEACHERS"

Raj Chetty, John Friedman, and Jonah Rockoff

May 2014

In a recent National Education Policy Center (NEPC) report, Moshe Adler (2014) raises several potential concerns about the validity of Chetty, Friedman, and Rockoff's two papers (forthcoming in the *American Economic Review*) on measuring the impacts of teachers. After considering Adler's critiques carefully, our interpretation of the results is entirely unchanged. All of the concerns reflect a misunderstanding of statistical methods and the data we analyze, and we respond to each of them below.

Concerns on Paper 2 ("Teacher Value-Added and Student Outcomes in Adulthood"):

Adler Concern #1: An earlier version of the report found that an increase in teacher value-added has no effect on income at age 30, but this result is not mentioned in this revised version. Instead, the authors state that they did not have a sufficiently large sample to investigate the relationship between teacher value-added and income at any age after 28, but this claim is untrue. They had 220,000 observations (p. 15), which is a more than sufficiently large sample for their analysis.

Response: In the original version of our paper (NBER wp 17699, Table 6, Column 2), we did report estimates at age 30. The estimated impact at age 30 is $2,058, *larger* than the estimated impact at age 28 ($1,815). Why does Adler conclude that there is "no impact" at age 30 when the impact is actually larger? The reason is that Adler confuses statistical significance with magnitudes: the standard error of the estimate at age 30 is $1,953, and hence is statistically insignificant (i.e., one cannot reject the hypothesis that the effect is zero). However, this does not mean that there is no effect at age 30; rather, it means that one has insufficient data to measure earnings impacts accurately at age 30. Adler claims that the sample is "adequately large" based on a power calculation that assumes independent errors across observations and hence greatly overstates precision by failing to account for correlated errors across students within a classroom and repeat observations for students over time. The standard errors we estimate account for these issues and are a direct estimate of precision at age 30 in the data; indeed, the difference in the standard errors between age 28 and 30 is exactly what one would expect given the reduction in sample size. In the revised version of the paper, we dropped the estimates at age 30 in the interest of space since there is inadequate data at age 30 to obtain precise estimates.

Adler Concern #2: The method used to calculate the 1.34% increase is misleading, since observations with no reported income were included in the analysis, while high earners were excluded….

Response: Neither statement is correct. Observations with "no reported income" are not missing data; they are true zeroes because the data we use cover the universe of taxpayers and hence individuals with no W-2 or 1040 forms do in fact have zero taxable income. The number of individuals with zero income in our data is comparable to those in other datasets, such as the Current Population Survey (footnote 10 in paper 2). High earners are not excluded; we top code earnings for those in the top 1% (i.e., recode their income at the cutoff for the top 1%) in order to reduce the influence of outliers. Dropping this top coding has no impact on our estimates.

Adler Concern #3: The increase in annual income at age 28 due to having a higher quality teacher "improved" dramatically from the first version of the report ($182 per year, report of December, 2011) to the next ($286 per year, report of September, 2013)….Since the discrepancy is so large, it suggests that the correlation between teacher value-added and income later in life is random.

Response: The difference between our original paper and our revised paper is that we now estimate a model that permits stochastic drift in teacher quality. The model that permits drift places greater weight on more

recent test scores and thus captures more of the variance in current teacher quality. As we note in our revised paper (footnote 9 in paper 1), not accounting for drift yields smaller estimates of teachers impacts for this reason. Hence, one should expect the estimated earnings impact of teachers' true VA in a given year to increase, exactly as we find. When we analyze the impacts of current teacher VA on future earnings over a 10 year horizon, we again obtain estimates very similar to the results in our original paper, as drift in teacher quality reduces subsequent earnings impacts.

Adler Concern #4: In order to achieve its estimate of a $39,000 income gain per student, the report makes the assumption that the 1.34% increase in income at age 28 will be repeated year after year. Because no increase in income was detected at age 30, and because 29.6% of the observations consisted of non-filers, this assumption is unjustified.

Response: The issue of "no increase in income at age 30" is addressed in response to comment #1 above. The assumption of a constant 1.34% increase is likely conservative, as we note in our second paper, because the impacts of teacher VA on earnings are rapidly increasing with age over the ages for which we have adequate data to estimate impacts (Figure 2b of paper 2). Extrapolating forward, one would expect the earnings gains to be larger than 1.34% after age 28.

Adler Concern #5: The effect of teacher value-added on test scores fades out rapidly. The report deals with this problem by citing two studies that it claims buttress the validity of its own results. This claim is both wrong and misleading.

Response: The fade-out pattern is not a "problem"; it is a generic empirical finding that we and others have documented in other settings, for instance in the Project STAR kindergarten classroom experiment. There are many mechanisms that could lead to fade-out of impacts on test scores but lasting impacts on later outcomes such as earnings, such as non-cognitive skills (Chetty et al. 2011). Moreover, as we demonstrate in Figure 4 of paper 2, the test score impacts do not "fade out" entirely; they stabilize at roughly 0.25 SD after four years.

<u>Concerns on Paper 1 ("Evaluating Bias in Value-Added Estimates")</u>

Adler Concern #1: Value-added scores in this report and in general are unstable from year to year and from test to test, but the report ignores this instability.

Response: It is certainly true that value-added ratings fluctuate across years. However, the statement that we ignore this instability is incorrect. We discuss the reliability of VA estimates at length in Section III of paper 1 and evaluate its impacts on the long-term gains from the use of VA measures in Section VI of paper 2.

Adler Concern #2: The report inflates the effect of teacher value-added by assuming that a child's test scores can be improved endlessly.

Response: We make no such assumption. We take the empirical distribution of test scores and assess the impacts of teachers on the test scores that are actually observed.

Adler Concern #3: The procedure that the report develops for calculating teacher value-added varies greatly between subjects within school levels (math or English in elementary/high school) and between schools within subjects (elementary or middle school math/English), indicating that the calculated teacher value-added may be random.

Response: The "procedure" for calculating value-added does *not* vary across subjects or school levels: in all cases, we use exactly the same econometric methodology. However, it is correct that the estimates vary across

subjects and school levels: for instance, the variance of teacher effects is larger in math than English. This does not indicate that teacher VA is "random"; it indicates that there are differences in the distribution of teacher quality across subjects and school levels, which is perfectly plausible and consistent with prior work. There is no reason for the distribution of math teacher quality to be identical to English teacher quality.

Adler Concern #4: The commonly used method for determining how well a model predicts results is through correlations and illustrated through scatterplot graphs. The report does not present either the calculations or graphs and instead invents its own novel graph to show that the measurements of value-added produce good predictions of future teacher performance. But this is misleading. Notwithstanding the graph, it is possible that the quality of predictions in the report was poor.

Response: The use of ordinary least squares regressions is a standard tool in econometric analysis, and every result in the paper is based on a regression analysis. We supplement these regression estimates with binned scatter plots – a simple technique to represent conditional expectation functions in large datasets non-parametrically – as is now standard in papers that study large datasets. Our methods identify teachers' mean impacts on students' outcomes; while it is true that other factors also contribute to variation in students' outcomes, this does not affect the analysis of teachers' mean impacts.

Works Cited

Adler, Moshe. 2014. "Review of Measuring the Impacts of Teachers" National Education Policy Center Report.

Chetty, Raj, John N. Friedman, Nathaniel Hilger, Emmanuel Saez, Diane Schanzenbach, and Danny Yagan. 2011. "How Does Your Kindergarten Classroom Affect Your Earnings? Evidence from Project STAR." *Quarterly Journal of Economics* 126(4): 1593-1660, 2011.

Chetty, Raj, John N. Friedman, and Jonah E. Rockoff. 2014a. "Measuring the Impacts of Teachers I: Evaluating Bias in Teacher Value-Added Estimates." Forthcoming, *American Economic Review*.

Chetty, Raj, John N. Friedman, and Jonah E. Rockoff. 2014b. "Measuring the Impacts of Teachers II: Teacher Value-Added and Student Outcomes in Adulthood." Forthcoming, *American Economic Review*.