NATIONAL EDUCATION POLICY CENTER

School of Education, University of Colorado Boulder
Boulder, CO 80309-0249
Telephone: 802-383-0058

NEPC@colorado.edu
http://nepc.colorado.edu

Response of Moshe Adler to the Authors' Reply

June 24, 2014

Below is my response to the May 2014 “Response to Adler (2014) Review” by Chetty *et al*. My original points (e.g., “Adler Concern No. 1”), as quoted in the response from Chetty *et al*., are in roman type, followed by the responses of Chetty *et al*., which are italicized; my subsequent responses follow, also in roman.

Concerns About Part II: Teacher Value-Added and Student Outcomes in Adulthood

> **Adler Concern No. 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; this claim is not true. They had 220,000 observations (p. 15), which is a more than sufficiently large sample for their analysis.
>
> ***Chetty et al. 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 Reply: In the original version of their paper Chetty *et al.* had 61,639 observations of 30-year-olds. Chetty *et al.* analyzed that very large sample, and with this sample they discovered that they could not reject the hypothesis that teacher VA (value-added) does not affect income at age 30. Statistical significance is not something to be brushed aside. Regardless of the magnitude (which is discussed below), the existence of this result should have been reported. In the version that will be published by the American Economic Review (AER), Chetty *et al.* do not divulge the existence of this result.

But the magnitude issue is important here as well. In their response Chetty *et al.* inflate their results by a factor of 10.On page 39 of the original version of their research they state: "A 1 SD increase in teacher VA in a single grade increases earnings at age 28 by $182 [a year]." This is also what Figure 6 of their paper shows: "1 sd TVA [teacher value-added] = $182." (This result is statistically significant.) In the sample of 30-year-olds, a 1 SD increase in teacher value-added increases earnings at 30 by $206 a year. (This result is not statistically significant.) Both magnitudes are small, if not negligible. In their response Chetty *et al.* report figures that are 10 times larger, $1,815 and $2,058 respectively. These values would be the results of increasing the value-added of a teacher by 10 standard deviations, which is, of course, not possible because such a teacher does not exist.[1]

Chetty *et al.* inflate their results by a factor of 10 throughout their paper. Before I published my review of this paper I wrote to Jonah Rockoff about this problem as it applies to test scores. His written response is quoted and discussed below, where I examine the inflation of the effect of teacher value-added on test scores.[2] The fact that they repeat this misrepresentation after it was called to their attention is troubling.

The result that Chetty *et al.* found is very small for all ages and statistically insignificant for age 30, which invalidates their main claim that having a high value-added teacher in one year increases income over a lifetime. They should have divulged this result, and in their response to me they should not have justified the omission by misrepresenting its magnitude.

In the second part of their response to Adler Concern No. 1, Chetty *et al.* write:

> *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*

[1] The figures $1,815 and 2,058 are not exactly 10 times $182 and $206; the discrepancies are due to the rounding of the figures $181.5 and $205.8

[2] See also Page 5 of: Adler, M. (2013). Findings vs. interpretation in "The Long-Term Impacts of Teachers" by Chetty et al. *Education Policy Analysis Archives, 21*(10). Retrieved 6/10/2014, from http://epaa.asu.edu/ojs/article/view/1264.

> *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 Reply: A year passed between the original version and the second version of their paper, and Chetty *et al.* added one additional year (2011) of observations to their data set. In the version that will be published in the AER, Chetty *et al.* had not 61,639 observations but 220,000 observations at age 30. With these many observations, more than three times as many as they had in the first version, they "dropped the estimates at age 30 in the interest of space since there is inadequate data at age 30 to obtain precise estimates." Are 220,000 observations "inadequate data" even given the repeated observations and correlated errors across students within a classroom? In my review of their first version, I calculated that all they needed were 8,124 observations. Chetty *et al.* dismiss this calculation, but where is their calculation? Two hundred and twenty thousand observations is a large sample, and not to include the results of the statistical tests with this many observations in order to "save space" continues the pattern of not reporting results that may contradict the authors' main claim in their research.

There is also a problem with the way Chetty *et al.* treat their observations of the 29-year-old and 31-year-old cohorts. Chetty *et al.* do not mention these age groups at all in their paper but, since they have 28-year-olds and 30-year-olds in their data set, they must have observations for 29-year-olds as well; in addition, in their data for the first version they had observations for 30-year-olds, and the addition of one year to that data set means that in the data set for the second version they have observations for 31-year-olds.

The number of observations for each age group is large. The 28-year-olds of the first version are the 29-year-olds of the second version, and the 30-year-olds of the first version are the 31-year-olds of the second version. In the first version they had 376,000 observations for 28-year-olds and 61,639 observations for 30-year-olds, so these become the number of observations for 29- and 31-year-olds respectively in the second version. The first figure, 376,000 observations, was considered "large enough" by the authors in the first version, so the (same) number of observations of 29-year-olds in the second version should also be "large enough." Why do Chetty *et al.* then claim in the second version that "the oldest age at which the sample is large enough to obtain informative estimates of teachers' impacts on earnings turns out to be age 28"?

> **Adler Concern No. 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.
>
> ***Chetty et al. 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 Reply: Chetty *et al.* do not address the problem. The reported increase in income was $286 a year. This constitutes 1.34% of average income, but only because the calculation of average income includes the workers with zero income. An increase of 1.34% in the income of a person who earns an income is meaningful. But to somebody who does not earn anything this number is meaningless. 1.34% of what? Observations with zero income should not be included in the calculation of the percentage increase of income because they inflate the increase while at the same time rendering it meaningless.

In addition, the inclusion of zero-income observations means that the result of the analysis is random. Consider the following example. Suppose there are three workers; two who had a low value-added teacher and one who had a high value-added teacher. Of the first two, one does not report any income and the other earns $11 per hour. The third worker who had a high value-added teacher earns $9 per hour. The correct conclusion from these data is that those who had a high value-added teacher had (a) a lower probability of not reporting income and (b) a roughly 20% lower wage. Because we expect that a high (or low) value-added teacher would have a similar effect on all workers, we would estimate that among non-reporters also, the incomes of those who had low value-added teachers were 20% higher than the incomes of those who had high value-added teachers. But including non-reporters in the calculation and attributing zero income to them yields a very different result. It yields a finding that a good teacher increases a worker's wage by 53%.

The reason that including people who do not report earning a wage has such a dramatic effect on the calculation is that the distance between zero and most of the positive wages—even for non-filers—are far greater than the differences between the positive wages. The large differences dominate the calculations. Of course, the example above depends on the non-filers having had low value-added teachers. Otherwise the result that value-added has a positive effect on income would have meant that among the employed workers, those who had higher value-added teachers earned more. But this is precisely the problem with including non-filers: It hinges on the random proportion of low and high value-added scores among their teachers.

An additional problem is that many of the people to whom Chetty *et al.* assigned zero income probably had positive income. Chetty *et al.*'s procedure was as follows: "For non-filers, we define total income as just W-2 wage earnings; those with no W-2 income are coded as having zero total income. 29.6% of individuals have 0 total income in our sample."[3] Non-reported income and zero income are not the same. As one example, consider the following from *The New York Times*: "Various estimates put the tax cheat rate at 80 to 95 percent of people who employ baby sitters, housekeepers and home health

[3] WP 19424, page 9

aides."[4] A similar situation probably exists for many other occupations, including handymen and tutors, for example. In all these cases employers do not to fill out W-2 forms, and some of the workers may not report self-employment income either, but this does not mean that these workers are not getting paid. Assigning zero income to them biases the results of the regression as well as of the calculation of the percent increase, as we saw above.

Top coding, or assigning an income of $100,000 to any worker who earned more than that sum, further biases the results. Top coding is only justified when there is a need to preserve subjects' identities (which is why the Bureau of the Census uses it). But this is not the case here. According to the authors they "cap earnings in each year at $100,000 to reduce the influence of outliers" (page 9); however, as used here it assigns to subjects values that they do not have, which is misleading (it amounts to falsifying the data). Top coding biases both the results of the regression and of the calculation of the percent increase income due to an increase in teacher value-added. Moreover, Chetty *et al.* contradict themselves: the top coding had no effect, and we did it to reduce the effect of not doing it.

> **Adler Concern No. 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). Because the data sets are not identical, a slight discrepancy between estimates is to be expected. But since the discrepancy is so large, it suggests that the correlation between teacher value-added and income later in life is random.
>
> ***Chetty et al. 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* [sic] *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 Reply: This explanation is purely speculative. The authors could determine the actual cause of the dramatic difference between the two studies simply by applying their new method to the data of the first version. Otherwise the large discrepancy in results suggests that the correlation between teacher value-added and income later in life is largely if not completely random.

[4] Lieber, R. (2009, January 23).Doing the right thing by paying the nanny tax. *The New York Times*.

> **Adler Concern No. 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.
>
> ***Chetty et al. 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 Reply: The authors divulge that in the second version they had 220,000 observations for 30-year-olds and, for reasons that were explained above, we know that they had something like 376,000 observations for 29-year-olds and 61,639 observations for 31-year-olds; however, for all these ages they do not report their results at all. There is good reason to be troubled by their insistence that, even though they did not provide their results for 29-, 30-, and 31-year-olds in the second version, and in spite of their result in the first version that for 30-year-olds the null hypothesis cannot be rejected, they can still reject the null hypothesis that teacher value-added has no effect on income later in life.

Furthermore, as was shown above, because of the inclusion of observations with zero income and the top coding of incomes above $100,000, the figure of 1.34% increase is both biased upward and meaningless.

> **Adler Concern No. 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.
>
> ***Chetty et al.'s 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.*

Adler Reply: Chetty *et al.* claim that studies by David Deming and James Heckman *et al.* buttress the validity of their results when they do not, as I pointed out. But instead of acknowledging this misrepresentation, they simply do not respond to it. This poor use of research cannot be ignored, however, first because the fade-out casts doubt on the value of teacher value-added, and also because the programs that Deming and Heckman *et al.* studied do improve people's lives.

Chetty *et al.* are of course correct that they (and all other researchers) found a fade-out of the effect of teacher value-added on test scores. But just because the fade-out is "generic" does not mean that it is not a problem; on the contrary, it means that it is not a random result. It is an established pattern. Further, the fade-out of the effect of teacher value-added is a central issue in assessing the value of measuring teacher value-added. If teacher value-added does not have a lasting effect even on test scores while the child is in school, why would it have a lasting effect on a person's income throughout his or her life? This is a thorny question, and Chetty *et al.* pretend that Deming and Heckman *et al.* provide the answer to it.

In fact Deming and Heckman *et al.* investigated the long-term effects of Head Start and of the Perry Preschool Project, a high-quality pre-school program. They discovered that these programs improve people's lifetime performance, but they also discovered a fade-out of the effect of these interventions on cognitive results. Chetty *et al.* cite this fade out as evidence that the fade-out in their study does not matter. But to Head Start and the Perry programs, test scores are of no consequence at all. These programs deal with the psychological and emotional needs of children and of their parents, as well as issues of nutrition and health. These programs are beneficial not because they involve the hiring and firing of teachers or other personnel but because they increase the resources that are available to children and their families. The citing of these studies by Chetty *et al.* is particularly galling because Heckman *et al.* specifically draw attention to the distinction between the concerns of education economists and the concerns of these particular early education programs. They explain:

> The literature in the economics of education assumes the primacy of cognitive ability in producing successful lifetime outcomes. . . From this perspective, the success of the Perry program is puzzling. Although the program initially boosted the IQs of participants, this effect soon faded. . . Consistent with this evidence, we show negligible effects of increases in IQ in producing program treatment effects. Although Perry did not produce long run gains in IQ, it did create persistent improvements in personality skills. The Perry program substantially improved Externalizing Behaviors (aggressive, antisocial, and rule-breaking behaviors), which, in turn, improved a number of labor market outcomes, health behaviors, and criminal activities.[5]

While it is easy to understand why the Perry Project and Head Start lead to success in adulthood despite the fade-out in test scores, it does not follow that short-duration improvements in elementary school test scores would lead to economic success in adulthood. Citing early childhood programs as evidence that the fade-out of test scores in programs that are designed to increase test scores does not matter, is troubling.

[5] Heckman, J., Pinto, R., & Savelyev, P. (2012). Understanding the mechanisms through which an influential early childhood program boosted adult outcomes, 3. Retrieved March 1, 2014, from https://economics.sas.upenn.edu/sites/economics.sas.upenn.edu/files/u21/0_PerryFactorPaper_AER_2012-09-07_sjs.pdf.

It should also be noted that contrary to their claim, Chetty *et al.* do not show that the fade-out stabilizes after four years. They did not check the effect of value-added on test scores beyond year 4, and between year 3 and year 4 the effect declines. In year 4 the effect of a 1 SD increase in teacher value-added is a negligible 0.0221 SD, 13% lower than the 0.0255 of year 3. (Note that Chetty *et al.* report this effect as 0.221SD, but this is due to the tenfold inflation, as discussed below).

Concerns About Part I: Evaluating Bias in Value-Added Estimates

> **Adler Concern No. 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.
>
> ***Chetty et al. 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 Reply: The model used for measuring teacher value-added actually does not measure teacher value-added; it measures a residue. The model explains students' test scores by factors such as students' prior achievement, parents' income, and the performance of other students in the classroom. But none of these factors can entirely predict a child's performance. After accounting for all known and easily measurable social and economic factors, a residue still remains. Some would attribute this residue to the unpredictability of life itself. But the economists who advocate the use of this model, including Chetty *et al.*, attribute the residue to the teacher.

As Chetty *et al.* acknowledge, the value-added scores of teachers "fluctuate" from year to year. As I pointed out, McCaffrey *et al.* used data from Florida, and Koedel and Betts used data from San Diego, to examine the stability of value-added scores.[6] Their results varied from place to place, but the average result was that 13% of teachers who were at the bottom 20% of the value-added scale in one year were at the top 20% the following year, and 29% of those at the bottom 20% were at the top 40% the following year. Similar results held at the top. Twenty-six percent of teachers who were in the top quintile in one year were in the bottom 40% the following year. Only 28% percent of teachers who were at the top one year stayed at the top the following year.

Value-added scores are also not consistent across tests. The state of Florida uses two different standardized tests, and McCaffrey and his co-authors found that the two tests yielded different results. Of the teachers who were in the bottom 20% of the value-added

[6] Sass, T. (2008, November). *The stability of value-added measures of teacher quality and implications for teacher compensation policy*. Washington, DC: The Urban Institute. Retrieved March 1, 2014, from http://files.eric.ed.gov/fulltext/ED508273.pdf.

score according to one test, 5% were at the top 20% and 16% were at the top 40% according to another test.

Of course, a teacher may occasionally have a bad year or a good year or have more luck with one exam than another. But statistical analysis, including in Chetty *et al.*'s study, reveals that these fluctuations are common. If we make the assumption that a good teacher is a good teacher and a mediocre teacher is a mediocre teacher, then the high volatility of teacher value-added scores raises the suspicion that the measurement of teacher value-added does not actually measure teacher quality.

But the adherents of value-added measurements do not agree. Their response is that value-added scores contain a stable part, which is a reflection of teacher quality, and a volatile part that is determined by transitory factors. The fluctuations in teacher value-added scores are so wide, however, that they raise doubt that they measure anything real. A grocer who stocks fresh produce may slip up on occasion, but the ranking of the quality of the produce among stores is usually stable. Nevertheless, using data from six school districts, the Bill & Melinda Gates Foundation's "Measures of Effective Teaching" (MET) Project calculated the "stable" part of teacher value-added, and their results shed light on its stability.

To calculate value-added scores the project used two types of tests for English (state tests and the Stanford 9 Open-Ended tests) and two types of tests for math (state tests and the Balanced Assessment in Math (BAM) tests). It found that the correlations between the "stable" teacher value-added measured using one test or the other were low, .37 for English and .54 for math.[7] The economist Jesse Rothstein then examined how teachers would fare if evaluated by these "stable" measurements. He calculated that with a correlation of .37, a third of the teachers who have a stable measurement at the top 20% of teachers according to one test are measured as being below average according to the other. With a correlation of .54 the odds of consistency are only slight better: 30% of the teachers who are measured to be at the top 20% of teachers according to one test are measured as being below average according to the other. The instability of even the "stable" measurements of teacher value-added only buttresses the suspicion that all they measure is the unpredictability of student performance on tests.

Chetty *et al.* acknowledge the instability of the measurements of teacher value-added, but they do not deal with its implication. They calculate teacher value-added by running a regression of teacher value-added in the current year on lagged teacher value-added with 10 lags. But if teacher value-added scores do not measure teacher quality, as the instability of even the "stable" part of teacher value-added indicates, then Chetty *et al.*'s regression uses meaningless numbers to predict meaningless numbers.

> **Adler Concern No. 2:** The report inflates the effect of teacher value-added by assuming that a child's test scores can be improved endlessly.

[7] Rothstein, J. (2011). *Review of "Learning About Teaching: Initial Findings from the Measures of Effective Teaching Project."* Boulder, CO: National Education Policy Center. Retrieved June 17, 2014, from http://nepc.colorado.edu/thinktank/review-learning-about-teaching.

> ***Chetty et al. 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 Reply: On page 41 of their paper Chetty *et al*. write:

> *In our companion paper, we estimate that 1 unit improvement in teacher VA in a given grade raises achievement by approximately 0.53 units after 1 year, 0.36 after 2 years, and stabilizes at approximately 0.25 after 3 years (Chetty, Friedman, and Rockoff 2014, Appendix Table 10). Under the assumption that teacher effects are additive across years, these estimates of fade-out imply that a 1 unit improvement in teacher quality in all grades K-8 would raise 8th grade test scores by 3.4 units.*

The actual numbers in Appendix Table 10 are as follows. A single (1.0) unit increase in teacher value-added increases student test scores by 0.993 SD in the year in which the teacher is in the classroom, 0.533 after 1 year, 0.362 after 2, 0.255 after 3 and 0.221 after 4. This last figure is assumed to hold from year 4 on. The addition of all these numbers from k-8 yields 3.4 units. In other words, their response notwithstanding, Chetty *et al*. do assume that a child's test scores relative to her peers can be increased without limit.

These numbers, which are presented in Figure 4 in Chetty *et al*. (wp No. 19424), are due to an inflation by a factor of 10. An increase of 1 SD in teacher value-added does not in fact increase student test scores by 0.993 SD. Using the Chetty *et al*. analysis, it actually leads to only a 0.0993 SD increase in student test scores in the year in which the teacher is in the classroom and 0.0221 after four years.

Chetty *et al*. have the same misleading figure (although the numbers in it are slightly different) in the first version of their research, and before I published my review of that version I contacted Jonah Rockoff, one of the coauthors, to ask about it. In particular I asked about the effect in year 3. By my calculation it should have been 0.03 instead of 0.3. Below is his response:

> This is definitely a language error on our part. Instead of "a one SD increase in teacher quality" we should have said "an increase in teacher value-added of one (student-level) standard deviation." Of course an increase of one in value added is roughly 10 teacher-level standard deviations, so your assessment of 0.03 in year three for a one teacher-level SD increase in year 0 would be correct (Jonah Rockoff to Moshe Adler, personal communication, October 12, 2012).

To clarify Rockoff's answer, it should be noted that Chetty *et al*. found that an increase of one standard deviation in teacher quality leads to an increase of about 0.1 of a standard deviation in test scores.

Rockoff's answer is problematic, however. If the authors were to change the language to reflect Rockoff's correction, it would read: "In our data, the impact of [an increase in teacher value-added of one (student-level) standard deviation] stabilizes at approximately 0.3 SD after three years, showing that students assigned to teachers with higher value-added achieve long-lasting test score gains." But this statement would not be true because the probability of finding a teacher whose value-added is 10 teacher-level standard deviations above the value-added of other teachers is practically zero. In fact, the range of teacher value-added in the Chetty *et al.* data is from about -0.18 to +0.18, or a maximum difference of about 0.36 teacher-level standard deviations.[8] Students cannot be assigned to the high value-added teachers that the Chetty *et al.* statement mentions because such high value-added teachers simply do not exist.

The fact that Chetty *et al.* continue to use this misrepresentation after it was pointed out to them is troubling.

> **Adler Concern No. 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.
>
> ***Chetty et al. 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 Reply: Chetty *et al.* use the regression that I discussed above to derive their procedure and, just as I said, the result is a procedure that is inconsistent between subjects and between schools. According to their Table 2, to calculate a teacher's value-added score in NYC, if the teacher is an English teacher in elementary school, a greater weight is placed on test scores of her or his students from 10 years ago than the test scores of the students of six years ago. For an English teacher in middle school, however, lower weights are placed on the test scores of both 10 years ago and six years ago, but these weights are equal. If it is a math teacher in elementary school, a smaller weight is placed on the test scores of 10 years ago than on the test scores of nine years ago, but if it is a math teacher in middle school, the opposite holds: A higher weight is placed on the scores of 10 years ago than on the scores of nine years ago. A similar procedure applies for each grade from 1-10. If this all appears random, it is. Why would a teacher's performance from 10 years earlier be included when calculating the quality of her teaching today if her performance over the last two or three years is known? And why would an elementary school English teacher's performance 10 years earlier matter more than performance six years earlier, while for a middle-school English teacher the two would have the same impact? Why would the performance of 10 years ago be more important than the performance of nine years ago in

[8] Chetty et al., wp 17699, Figure 6.

elementary school, while in middle school the reverse is true? And why would the importance within the same year vary so much from one school to the next or one subject to the next? The authors arrived at these numbers by using regression analysis, but what the regression reveals is that these numbers are random. All in all, the consistency (and therefore the validity) of the measure is left in question.

> **Adler Concern No. 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.

> ***Chetty* et al. *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.*

Adler Reply: Ordinary least squares regression is indeed a good tool for statistical analysis and nowhere do I dispute the validity of its use. But the reason it is a good tool is because it permits the researcher to determine how well the regression model fits the data. This determination is summarized by a number called R square, and its graphical representation is through a scatterplot of actual and predicted values. The authors claim to "augment" their regression analysis, but in fact they omit this determination and instead they include a plot that is misleading because it shows what to a lay person may appear to be a good fit, when in reality it may not be. Notwithstanding the near perfect fit that figure 2a (wp No.19423) presents, it is possible that teachers with low predicted value-added scores had high actual scores and that teachers with high predicted scores had low actual scores.