Making Decisions with Imprecise Performance Measures: The Relationship Between Annual Student Achievement Gains and a Teacher's Career Value-Added

By Douglas O. Staiger, Thomas J. Kane

This Draft: December 23, 2013.

I. Introduction

There is continuing confusion and debate over whether value-added measures are sufficiently reliable to be used in high-stakes personnel decisions. Critics often point to the year-to-year volatility in value added measures as *prima facie* evidence against their use. They ask, how can we sanction teachers who are in the bottom quartile of value added this year when we know that value added for nearly two thirds of these teachers will no longer be in the bottom quartile when measured again next year? On the other hand, the results of the MET project and others have highlighted that even unreliable performance measures such as value added can identify substantial and lasting differences across teachers.

Our goal in this chapter is to reconcile these two views. Despite the fact that value-added measures are unreliable by conventional standards and unstable over time, they are strong predictors of an individual teacher's career performance that can be used to improve decision making. Much of the confusion is due to an over-interpretation of seemingly low year-to-year correlations in value-added measures. For most decisions, year-to-year volatility in annual performance is the wrong statistic for judging the informational value of value-added data. A retention decision, for instance, rests on a different relationship, the correlation between a single year's performance (or performance to date) and a teacher's career performance. We propose a way to infer the year-to-career correlation using the year-to-year correlation. We also test that method using data from several urban school districts which have 7 or more years of data on teacher's value-added. We show that the year-to-career performance correlation can be estimated with a simple calculation, that the estimate corresponds with the actual correlation observed between a single year of value added and a teacher's multi-year average, and is substantially stronger than the year-to-year correlation in performance.

In addition, we study the usefulness of value-added data in a retention decision. To do so, we model the decision problem faced by a supervisor. When analyzed in that way, it becomes clear that every retention decision involves two teachers—the incumbent teacher and a prospective new hire. Although the latter is usually anonymous, a principal's or supervisor's decision requires comparing the likelihood that either of the teachers will turn out to be high performing. A common conceptual error is to focus on the degree of uncertainty surrounding an individual teacher's likely career performance. Yet it is the teacher's performance relative to the potential replacement that matters. In the worst case scenario, the supervisor would have to hire a rookie every year to fill the slot. In that context, the right decision rule would be to ask whether the teacher was likely to be more or less effective than an infinite series of novice teachers. We find that even one year of value added data can substantially reduce the chance of making a mistake.

Unlike many debates in education, there is surprisingly little dispute about the underlying facts. In most studies, the correlation in test-based measures of teaching effectiveness between one school year and the next ranges between .35 and .50 in elementary grades, and is somewhat higher in middle school grades (where value added is based on multiple classrooms per teacher). Such fluctuations are due to a number of factors— such as the finite number of students in their classrooms in a given year. For instance, an elementary teacher will have between 15 and 25 students. With samples that small, a few unusually rowdy or studious students can make a difference from year to year.

A correlation as low as .35 can produce seemingly troubling statistics in terms of year to year changes. For instance, only about a third of teachers ranked in the top quartile (highest 25%) of value added based on one academic year's performance would appear in the top quartile again the next year. Moreover, ten percent of bottom quartile teachers (bottom 25 percent) one year would appear in the top quartile the next.

Such instability in measures of performance is not unique to teaching. In a wide range of settings, ranging from using SAT scores to predict college GPA (Camara and Echternacht (2000)), surgical mortality rate at hospitals (Dimick et. al. (2009)), to the batting averages and earned-run-averages for major league baseball players (Schall and Smith (2000)), annual performance measures show similarly low correlations yet are regularly used for high stakes evaluation[1].

There are three key questions which we should be asking when trying to interpret value-added:

- **Does a teacher's value-added one year predict value-added over his or her career?**

It would be troubling if the measures were so volatile that one year's performance does not predict future performance. But this is not true. Despite the fact that annual performance of all teachers varies widely from year to year, this variation is not enough to hide large differences across teachers in their underlying career performance. As the evidence presented below demonstrates, value added from one year of teaching predicts large differences in performance over the teacher's career. For example, teachers ranked in the bottom 25% based on a single year of value added will typically perform worse than an average rookie teacher over the remainder of their career. Averaging value added over two years of teaching predicts even larger differences in career performance.

- **Would our impression of a given teacher's performance change wildly from one year to the next? Would we simply be whipsawing and confusing school administrators and teachers by providing them with annual performance data?**

This too would be troubling-- if it were true. But beliefs about teacher performance are *cumulative*. When provided with two years of value-added data, an administrator should use the *average* over the two years rather than focusing solely on the most recent year. Why? Because the two-year average is a better predictor of career performance. Yet, despite the volatility in single-year measures, teacher rankings based on cumulative estimates of teacher value added change very little from year to year. For instance, in the districts we look at, less than 1 percent of those who were in the top quartile of performance after only one year of data would be in the bottom quartile over two years. Less than 4 percent would be in the bottom quartile over four years. Despite the volatility, there is a low probability that someone averaging performance over multiple years would change their minds about who is more and less effective.

[1] See Goldhaber and Hansen (forthcoming) and Sturman et. al. (2005) for a summary of other examples.

- **Can value-added be used to improve decision-making?**

Would a single year measure of performance lead to too many mistakes? It is impossible to say without knowing the decision to be made and the costs of different types of mistakes. Using the example of a principal deciding whether to renew a new teacher's contract for a second year, we show that not renewing the bottom 25% of teachers based on one year of value added would increase the chance of having a more effective teacher in the classroom (and thereby reduce mistakes), even if the principal had to replace the incumbent teacher with a string of newly hired rookie teachers.

Our conclusion is that value-added measures are useful despite their volatility. Test-based measures of a teacher's effectiveness from one year do predict their effectiveness over their careers. Moreover, cumulative estimates of teacher value added change very little over time, and predict large career performance differences. Finally, decision-making improves when value-added is used in an appropriate manner.

The first section of this chapter develops a statistical model of teacher impacts on student test scores in order to look at the issue theoretically. In this section, we (1) define what we mean by a teacher's underlying long-term (career) performance, (2) show that the correlation with underlying long-term performance is different from year-to-year correlation in annual performance, (3) show how to easily calculate the correlation between any observed measure and a teacher's underlying long-term performance, and (4) argue that the correlation with underlying long-term performance summarizes both the predictive power and risk of misclassification for any annual performance measure.

The second section provides evidence from three large districts in order to look at the issue empirically. Rather than using data from the MET study, which was only available for a small number of years, we use historical data from three large anonymous districts with up to 9 years of value added data for each teacher (required for estimating career value added). We report on three types of analyses: (1) comparing differences in career value added from ranking teachers on one year versus ranking them on career, and showing what fraction of the total career differences you can capture with one year, (2) showing how rankings on one year of value added misclassify teachers in terms of their career value added (versus next year value added), and (3) comparing the stability and predictive power of cumulative measures of value added versus one-year measures of value added.

The third section looks at the issue of using value added in the context of a simple decision-making problem under uncertainty. We consider the problem facing a principal in deciding whether to renew a new teacher's contract for a second year. This problem can be thought of as having to choose between two teachers (the incumbent teacher and the prospective new hire), with the goal of choosing the teacher who will have a higher effectiveness in the classroom over their career. Using this framework, we show that most teachers who are in the bottom quartile in their first year of value added will have lower value added over their career than a typical new hire. The bottom 25% of teachers based on one year of value added will have career performance worse than if their position was filled every year by a new rookie teacher. In contrast, using a "legal" standard of only removing incumbent teachers if one is 95% certain that they are below average results in mistakenly retaining teachers who have up to a 90% chance of having worse value added over their career than if one had filled the position with a new rookie each year. Relative to the current practice of identifying only a small fraction of teachers as ineffective, our evidence suggests a more aggressive policy of identifying at least the bottom quarter of teachers as ineffective.

II. A Statistical Model of Teacher Impacts

Suppose teachers did not differ in the degree to which their impacts on students improved or declined over their careers. Rather, suppose their measured impact on students were simply fluctuating randomly around their long-term average.[2] If that were the case, we ask, what would the current (albeit imperfect) measure of a teacher's effectiveness tell us about their long-term effectiveness?

The above scenario would have two important implications: first, it would mean that the estimated impact of a teacher in any given year or with any given group of students is a noisy estimate of a teacher's long-term impact on student achievement; and, second, it would mean that one could estimate just how much noise there is in any given year by observing the correlation in estimated impacts from any two years. In particular, the correlation between any current measure of value added and the expected long-term effectiveness is simply the square root of its correlation with another single-year of value-added. The only requirement is that the two annual value-added measures are estimated for different classrooms of students, so that the errors are independent.

To see why the square root of the year-to-year correlation is an estimate of the year-to-career correlation in value added, consider the following simple model. Suppose the short-term measure of a teacher's value added can be expressed as $T_{ij} = \alpha_i + \varepsilon_{ij}$, where i is a subscript for teacher and j is a subscript for school year or classroom of students, α_i is the effect of that teacher over the long-term, ε_{ij} is a transitory error in measurement which is uncorrelated for different values of j and unrelated to α_i. (This simple model supposes that there's a "fixed" teacher effect, which does not drift or evolve over time.) Then, the correlation between $T_{ij}$and α_i can be written as:

$$\rho_{T_{ij},\alpha_i} = \frac{\sigma_{T_{ij},\alpha_i}}{\sigma_{T_{ij}}\sigma_{\alpha_i}} = \frac{\sigma_\alpha^2}{\sigma_{T_{ij}}\sigma_{\alpha_i}} = \sqrt{\frac{\sigma_\alpha^2}{\sigma_{T_{ij}}{}^2}} = \sqrt{\rho_{T_{ij},T_{ik}}}.$$

The latter term is the square root of the correlation between the short-term measure for two different school years or distinct groups of students, j and k.

The intuition for this result is fairly simple. The year-to-year correlation in value added is based on two noisy estimates of a teacher's underlying career performance, i.e. it is the correlation between one noisy measure from this year and another noisy measure from the next year. Because both this year's and next year's value-added measures are noisy, the correlation between the two tends to be low. However, the year-to-career correlation should be greater than the year-to-year correlation because it is based on only one noisy estimate, i.e. it is the correlation between one noisy measure from this year and the teacher's actual career performance (not a noisy estimate from next year). Thus, for example, if the year-to-year correlation is 0.36, taking the square root implies a larger year-to-career correlation of 0.6. The year-to-year correlation is misleading in that it suggests that this year's value added is only weakly related to a teacher's future performance, while in fact it is only weakly related to the teacher's noisily measured performance from next year and is more strongly related to the teacher's career performance.

[2] The key assumption here is that the errors in measurement are independent across different time periods and with different groups of students, and the errors are unrelated to a teacher's long term effectiveness. This also implies that there is no gradual "drift" in effectiveness. It is straightforward to allow for drift, e.g. to let underlying teacher effectiveness follow a statistical model such as an auto-regressive process. This would have little impact on the analyses we present here.

The estimated correlation of annual value added with long-term effectiveness captures the two things we most care about in any measure: predictive power and the risk of putting teachers in the wrong categories using the measure.

Predictive Power. Suppose we were to use an annual measure of value added to identify teachers with more effective and less effective practice. The difference in expected long-term student achievement gains for those in the two groups will be proportional to the correlation with long-term value-added. This is what we mean when we say it is a measure of predictive power. For example, if we rank teachers into quartiles based on one year of value added data, a year-to-career correlation of 0.6 implies that the difference in career value added between top and bottom quartile teachers will be 60% as large as it would be if we ranked teachers into quartiles based on actual career value added. Thus, a year-to-career correlation of 0.6 implies that we can capture 60% of the potential differences in career value added with just one year of value added data.

Miscategorization. The estimated correlation with long-term effectiveness is also an indirect measure of the risk of misclassifying those with different long-term effectiveness based on short term measures. Under the assumption that the distribution of effectiveness is bell-shaped, the difference in the *probability* that a teacher in either group is above or below some threshold in long-term effectiveness is proportional to the correlation with long-term effectiveness.[3] In fact, the difference in the probability that a top or bottom quartile teacher on any given measure has above (or below) average value-added in the long-term is approximately equal to the measure's correlation with long-term effectiveness. For example, if we rank teachers into quartiles based on one year of value added data, a year-to-career correlation of 0.6 implies that a top quartile teacher will have a 60% greater chance than a bottom quartile teacher of having above average value added over his or her career.

The hypothetical scenario in which teachers underlying effectiveness does not change over their careers is unlikely to be exactly true. However, the evidence in Chetty et. al. (2013) and Goldhaber and Hansen (forthcoming) suggest it is not far off as an approximation. Therefore, it will be an empirical question as to whether this approximation is a good guide in practice – i.e., whether the square root of the year-to-year correlation in value added is approximately equal to the year-to-career correlation is an empirical question to which we now turn.

III. Evidence from Three Large Districts

To explore the implications of volatility, we used actual data on teacher-level "value-added" from three large districts. Estimating a teacher's career value added requires many years of value added data. Therefore, rather than using data from the MET study, which was only available for a small number of years, we use historical data from three large anonymous districts with up to 9 years of value added data for each teacher. Each of the districts had a large sample of teachers in grades 4-8 teaching ELA and math, for whom we could estimate between 6 and 9 years of value added data. We used standard methodology for calculating teacher value added using student achievement, including statistical controls for each student's performance on state tests from the prior school year as well as controls for gender, race, ethnicity, free or reduced price lunch status and the means of all the above characteristics for the other students in the class. In each year, we average value added estimates over

[3] We have demonstrated this by simulation.

all the classrooms taught by a teacher (typically one classroom in elementary grades, and 2-5 classrooms in middle grades).

Comparing year-to-year and year-to-career correlations

In the first row of Table 1, we report the year-to-year correlations for the 3 districts in ELA & math. These correlations are typical of what is seen in the literature, ranging from .25 to .62, with higher year-to-year correlations in math than in ELA. The second row of Table 1 reports the implied year-to-career correlations – i.e. the square root of the first row. These are predictably larger than the year-to-year correlations, and range from .50 to .78. More importantly, they are almost identical to what we get when we calculate actual year-to-career correlations, reported in the third row of Table 1. These are based on correlating single-year value added with a teacher's average value added over their entire career. All of the teachers in the sample have 6-9 years of value added data, with the average career value added being based on 6.7 to 7.6 years of data (in the bottom row of the table). Thus, the square root of the year-to-year correlation is an excellent guide to the correlation one will actually observe between one year of value added and value added over the teacher's career.

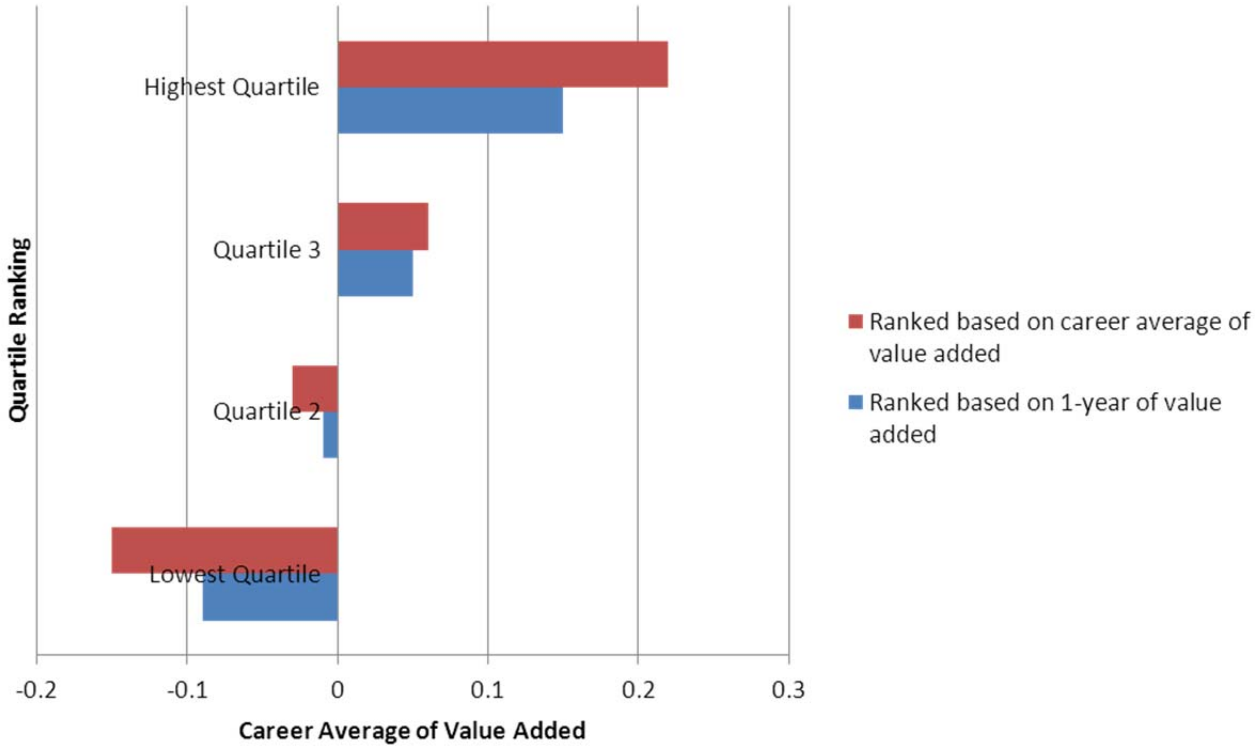
Table 1: year-to-year versus year-to-career correlations

	Math			ELA		
	District 1	District 2	District 3	District 1	District 2	District 3
Correlation of value added:						
Year-to-Year	0.42	0.62	0.47	0.27	0.48	0.25
Implied Year-to-Career (square root)	0.65	0.78	0.69	0.52	0.69	0.50
Actual Year-to-Career	0.65	0.78	0.70	0.55	0.71	0.57
Number of Teachers	2832	3984	377	2640	4197	370
Average number of years per teacher	7.2	6.7	7.5	7.1	6.7	7.6

The results in Table 1 highlight the importance of the distinction between year-to-year and year-to-career correlations in value added. All of the debate has focused on the low year-to-year correlations. However, Table 1 demonstrates that the actual year-to-career correlations are much higher. Thus, annual value added measures are a fairly powerful predictor of a teacher's career performance despite low year-to-year correlations in value added. Moreover, as theory would suggest, in the absence of data on career value added, the square root of the year-to-year correlations is a useful way of estimating the year-to-career correlation.

In Figure 1, we show the average difference in career value added for math teacher's sorted into quartiles based on one year of value added in District 1 (blue bars). For comparison, we show the difference in career value added when we sort the same math teachers into quartiles based on their career value added (red bars) – the best we could do. One year's ranking identifies 65% of the eventual difference in career value added that we could eventually identify, i.e., the difference in career value added between the highest & lowest quartile ranked based on 1 year of value added is 65% of the best case (ranking based on career value added). This is perfectly consistent with a year-to-career correlation in value added of 0.65 for math in District 1. Results are very similar for other subjects & districts. Thus, the year-to-career correlation is an excellent guide to the predictive power of one year of value added.

Figure 1. One Year's Ranking Identifies 65% of Eventual Difference in Career Value Added for Math in District 1, Consistent with



Do rankings on one year of value added misclassify teachers in terms of their career value added?

In Tables 2-4, we provide evidence of how badly a single year of value added misclassifies teachers in terms of their career performance. In Table 2, we show how teachers ranked in the bottom 25% based on one year of value added rank on career value added. For each district and subject, we report the percent of these teachers, all of whom were ranked in the bottom 25% based on 1 year of value added, who ranked in each quartile based on their career value added. For example, for math in District 1, 55% of the teachers ranked in the bottom quartile on 1 year of value added turned out to be in the bottom quartile over their entire career, and 82% (55% + 27%) were below average over their career. In contrast, only 5% of these teachers ended out being in the top quartile over their career. Results for other districts and subjects are similar. Note that for math in district 1, the difference between the percentage below average (82%) and the percentage above average (18%) is 64%, which is very close to the year-to-career correlation reported in Table 1 as predicted by our statistical model. Thus, while there is some misclassification, rankings based on one year of value added have only modest amounts of misclassification that are in line with the simple estimate of year-to-career correlation.

Tables 3 and 4 repeat this analysis, but limiting the sample to teachers who performed in the bottom 10% and 3% based on a single year of value added. These tables might be more representative of real-world practice, where only a small percentage of the worst-performing teachers are being identified for dismissal or as needing improvement. Teachers ranked in the bottom 10% or 3% of one-year value added are even more likely to be in the bottom quartiles over their career. Of the teachers ranked in the bottom 3% based on one year of value added, only 2-12% (depending on district and subject) rank above average on career value added, and 3% or less rank in the top quartile over their career. Thus, teachers in the tails of the distribution on 1-year value added are relatively unlikely to be mis-categorized in terms of their career performance.

Table 2: misclassification rates for teachers in bottom 25% on one year of value added data

	Teacher Ranked in Bottom 25% Based on One Year of Value Added Data					
	Math			ELA		
	District 1	District 2	District 3	District 1	District 2	District 3
% of Teachers falling in each quartile of ***career average*** value added (6+ years)						
bottom quartile	55%	65%	59%	48%	60%	50%
2nd quartile	27%	25%	27%	27%	26%	27%
3rd quartile	13%	8%	11%	17%	10%	16%
top quartile	5%	1%	3%	8%	4%	7%

Table 3: misclassification rates for teachers in bottom 10% on one year of value added data

	Teacher Ranked in Bottom 10% Based on One Year of Value Added Data					
	Math			ELA		
	District 1	District 2	District 3	District 1	District 2	District 3
% of Teachers falling in each quartile of ***career average*** value added (6+ years)						
bottom quartile	67%	81%	70%	61%	76%	61%
2nd quartile	22%	15%	20%	22%	16%	23%
3rd quartile	8%	4%	8%	12%	6%	11%
top quartile	3%	1%	1%	5%	2%	5%

Table 4: misclassification rates for teachers in bottom 3% on one year of value added data

	Teacher Ranked in Bottom 3% Based on One Year of Value Added Data					
	Math			ELA		
	District 1	District 2	District 3	District 1	District 2	District 3
% of Teachers falling in each quartile of ***career average*** value added (6+ years)						
bottom quartile	75%	91%	81%	71%	86%	70%
2nd quartile	18%	7%	12%	17%	9%	18%
3rd quartile	4%	2%	5%	8%	4%	11%
top quartile	3%	0%	2%	3%	1%	1%

Comparing the stability and predictive power of cumulative measures of value added.

1. Does early-career performance predict later performance

In Tables 5 (district 2) and 6 (district 1), we limit the sample to those teachers with value-added data during their 1st through 4th year of teaching (in district 3 samples were too small for this analysis). We first sort teachers into quartiles using their value-added during their first year of teaching. We also do so using their value-added averaged over their first two years of teaching. The first column reports the mean value-added of each group during their third and fourth year of teaching. In District 2, those with math value-added in the top quartile during their first year of teaching led students to a performance .14 standard deviations *above* similar students during their third and fourth year of teaching. Those with value-added in the bottom quartile during their first year had students with value-added gains .14 standard deviations *below* similar students during their third and fourth year. In other words, value-added data--even from the first year of teaching—*does* help predict student achievement gains in future years. In fact, the stakes involved in being assigned a 3rd or 4th year teacher who had performed in the top vs. bottom quartile during his or her first year of teaching are quite large—roughly a quarter of a standard deviation (approximately a quarter of the black-white achievement gap).

The predictive value increases somewhat by averaging over the first two years of teaching. For instance, those who were in the top quartile after two years had students with gains .17 standard deviations above similar students during their third and fourth years, while those who were in the bottom quartile after two years watched their students lag behind -.18 standard deviations during their third and fourth year. Instead of a .28 standard deviation difference, there's a .35 standard deviation difference between those assigned a top or bottom quartile teacher as ranked at the end of their first two years of teaching.

All these differences were somewhat smaller in reading. (Researchers commonly find larger teacher effects on math achievement.) They are also somewhat smaller in District 1 (Table 6) than in District 2 (Table 5). However, the same two findings hold true: First-year teacher performance does predict future performance. And combining data over the first two years increases the predictive value somewhat.

2. Does cumulative performance change over time?

In the subsequent columns in Tables 5 and 6, we report the percent of teachers appearing in each quartile of value-added when another year's worth of value-added data are added. As noted above, one gains predictive power by averaging the measures over more than one year. How much would those measures change when averaging in another year? Among those who were in the first quartile at the end of their first year of teaching, 71 percent were in the top quartile over the first two years of teaching. Less than 1 percent appeared in the bottom quartile over the first two years of teaching. In fact, only 5 percent of those who started out in the bottom quartile at the end of their first year would appear in the top half over two years.

The impact of adding another year's worth of data is even smaller after two years. 82 percent of those in the bottom quartile after two years would appear in the bottom quartile after three years. Less than 1 percent of those in the bottom quartile after two years appeared in the top half over three years.

Overall, Tables 5 and 6 suggest that accumulated value added estimates, averaged over a teacher's career to date, are better predictors of future value added and are considerably more stable than single year value added estimates.

IV. Using Value Added: Decision-Making under Uncertainty

Would a single year measure of performance lead to too many mistakes? It is impossible to say without knowing the decision to be made and the costs of different types of mistakes.

More than two centuries ago, Daniel Bernoulli wrote a famous paper on the dilemma facing an 18th century merchant when deciding whether or not to insure a ship's cargo in winter given the probability of an accident. (Bernoulli (1738)) Since that time, a rich theory of decision-making under uncertainty has been elaborated. In this section, we apply a simple decision-theoretic model to employment decisions at schools.

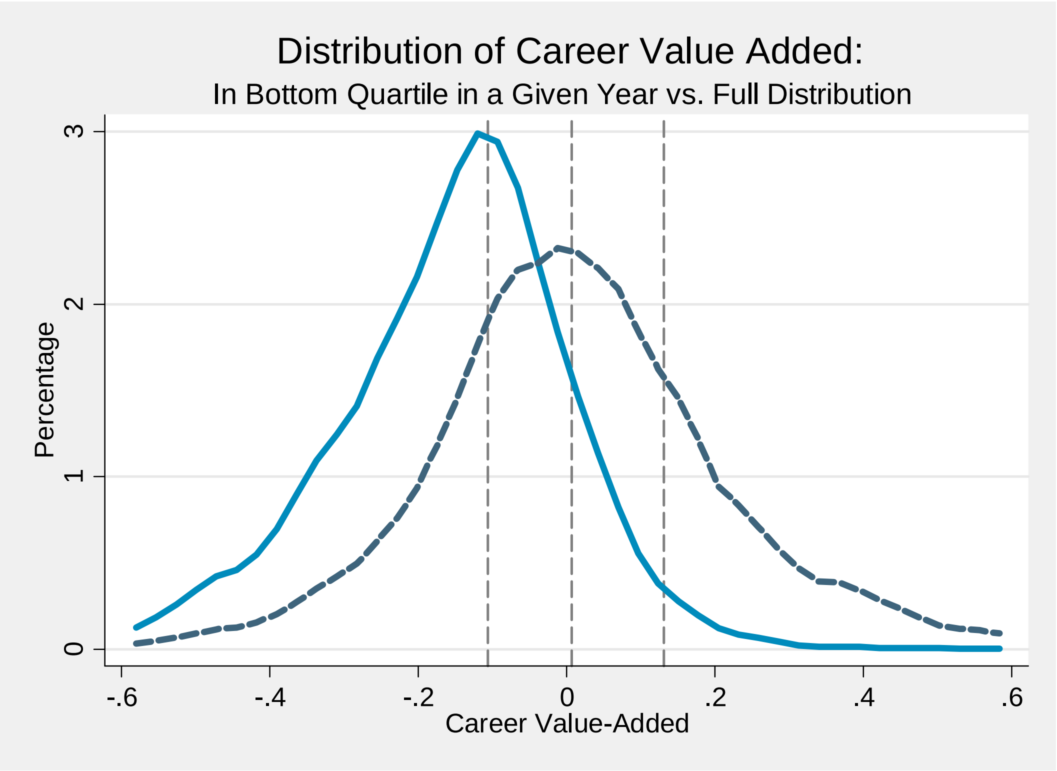
Consider the following hypothetical example: Suppose that an elementary school principal learns that an experienced teacher she recently hired completes her first year with *measured* effectiveness in the bottom quartile. It is only a single year of teaching and, as we've seen, one year is an imperfect signal of a teacher's likely career performance. The principal faces a dilemma: Should she renew the teacher's contract for a second year?

First, note that there is no such thing as a low-stakes decision. Whether she retains the teacher or replaces the teacher, there are consequences for two adult teachers (the incumbent who may be looking for work and a prospective replacement teacher who will be relieved to have finally found a job) and twenty-five youngsters who will be in the classroom.

Second, note that the principal is not assessing the performance of one teacher, but two: the incumbent teacher and, implicitly, the prospective replacement teacher. If the principal knows the potential replacement, she could compare the two teacher's recent performance. However, even if the principal does not know the potential replacement, she is not completely in the dark. Even if the principal will be required to take whomever the district's human resource department sends her, the distribution of possible outcomes is the distribution of career value-added for all teachers. The expected value of the teacher's career effectiveness is simply the mean career effectiveness across teachers. And the probability of different outcomes is reflected in the distribution of career effectiveness across all teachers.

In Figure 2, we report the distribution of career average achievement gains for those who appeared in the bottom quartile in a given year. We compare it to the distribution of career average achievement gains for all teachers. The vertical lines represent the 25th, 50th and 75th percentiles in career value-added for all teachers.

Figure 2. Distribution of Career Value Added for Teachers Ranked in the Bottom Quartile Based on One Year of Value Added Compared to All Teachers



What is the probability that a principal and her students will end up having a highly effective teacher in the future? That depends on the *difference* in the probability that the incumbent and the replacement teacher are highly effective. According to Table 1, if the incumbent teacher was in the bottom quartile one year, there is only a 4 to 8 percent chance that the teacher will turn out to be in the *top quartile* at the end of 6 or more years. But the likelihood that a randomly drawn replacement teacher will be in the top quartile is considerably higher—25 percent.

Which decision maximizes the chance that students have access to a top quartile teacher? If the principal keeps the incumbent and foregoes the opportunity to hire a replacement, she has a 4 to 8 percent chance that the teacher is highly effective. On the other hand, if the principal chooses the replacement teacher, she has a 25 percent chance of having a top quartile teacher. Therefore, the more concerned a principal is about the prospect of losing a great teacher, the more likely he or she will be to hire the replacement teacher and replace the incumbent.

Whenever a school leader has to make a decision based on a single year of data, they run the risk of falsely identifying a great teacher as ineffective. However, once a teacher has a poor track record—even on an imperfect measure—that teacher has *lower* odds of being a great teacher than a replacement teacher drawn at random. On the other hand, if a teacher has a strong track record, then they have higher odds of being a great teacher than an unknown replacement teacher.

The Expected Impact if the Principal Must Hire a Novice Teacher as the Replacement

How might the principal's decision differ if she knew that there were no experienced teachers available, that the only available replacements will be a novice teacher right out of graduate school? In that case, the appropriate comparison would be to expected effectiveness of the average *novice* teacher. Many researchers have studied the difference in effectiveness between the average novice teacher and other experienced teachers. Most of that research suggests that the students assigned to the average novice teacher lose .06 to .08 standard deviations in achievement during the teacher's first year of teaching relative similar students assigned to the average teacher.

Note that the principal's best prediction of the bottom quartile incumbent teacher's achievement gain (about -.10 standard deviations based on Figure 2) is still lower than the expected achievement gain of the average novice (-.06 to -.08) standard deviations. The principal could expect to raise 4th grade performance by.02 to .04 standard deviations *next year* by replacing the teacher and taking their chances with a novice. Admittedly, that's not a large difference. However, the principal could expect within two or three years that the average novice's achievement gains will converge toward that of the average teacher and the gap would be back to .10 standard deviations.

In sum, at least in terms of *expected* impact on students, the incumbent teacher in our example has a serious disadvantage with respect to any potential replacement teacher. On average, when ranked on just one year of value added data, teachers in the bottom quartile will typically perform worse than a novice teacher over their entire career. Even a single year of performance in the bottom quartile means that a teacher is a worse bet than an unknown teacher with a clean slate, even if that replacement is a novice teacher. While this conclusion may be surprising to some, it derives directly from the strong year-to-career correlation in value added (along with large differences in career performance across teachers). An alternative strategy for the principal would be to target bottom quartile teachers for professional development and in-service training. However, such training would have to be much more effective than traditional professional development, given that bottom quartile teachers otherwise would perform worse than rookies and have little chance of being highly effective over their career.

Presumed Average Until Proven Below Average?

Many analysts have sought to apply the framework of classical hypothesis testing to making employment decisions with imperfect information. (Schochet and Chiang (2010), Hill (2009), Baker et. al. (2010)) They argue that a high-stakes decision can be justified only when a teacher's performance is *statistically significantly* different from average. In effect, they would establish a "no deny zone" (and, presumably, a "no bonus zone") by adding and subtracting two times the standard error of measurement to the average teacher's performance. Within that range, no teacher would be statistically different from the average teacher. As a practical matter, given the size of standard errors of measurement, such a span would include most of the distribution of teachers, and leave only the extreme tails uncovered.

However, many employment decisions—such as the retention decision— do not fit the classical hypothesis testing framework. The classical hypothesis test was designed for a specific type of decision: when the costs of rejecting a true hypothesis are paramount and the cost of failing to reject a false hypothesis are secondary. In many areas of science, it makes sense to assume that a medical procedure does not work, or that a vaccine is ineffective, or that the existing theory is correct, until the evidence is very strong that the starting presumption is wrong. That is *why* the classical hypothesis test places the burden of proof so heavily on the alternative hypothesis, and preserves the null hypothesis—in our case, that the incumbent teacher is presumed effective—until the evidence is overwhelmingly to the contrary.

In the case of a retention decision, that would be inappropriate. To be sure, there are costs to failing to retain an effective teacher (that is, mistakenly rejecting the presumption that a teacher is effective). A poor decision with an incumbent teacher can have a negative effect on morale. Parents may complain. An incumbent teacher may be more likely to pursue legal action than a prospective teacher who was passed over. However, these costs are not overwhelmingly larger than the cost of retaining an ineffective teacher—a decision whose costs are born primarily by the students. This is especially true in the case of a tenure decision, when an ineffective teacher is granted decades of job security in teaching children. Indeed, the classical hypothesis testing framework would be especially inappropriate in a tenure decision, given that the cost of failing to reject a false hypothesis (that the teacher is effective) is likely to be larger than the cost of rejecting a true hypothesis.

In Figure 3, the horizontal axis reports the percentile of each teacher's value-added from one year. The vertical axis reports two types of statistics. The blue lines report the top and bottom points for the 95 percent confidence interval for a representative teacher in each percentile. The red lines report the 25th, 50th and 75th percentile of the career value-added of the teachers in each percentile. (These data represent actual career average value-added for teachers in the three districts for whom we could calculate value-added for more than 7 years. These are not simulations.) We have also noted in the graph the average value-added of a novice teacher in the district, which was -.08 student level standard deviations. Value-added is reported relative to the achievement gain achieved by the average teacher in the district. (By definition, the average teacher's value-added is 0.)

First, focusing on the upper and lower bounds of the 95 percent confidence interval (the blue lines), note that the 95 percent confidence interval for teachers in the bottom percentile includes zero. (The upper bound of the 95 percent confident interval is above zero for every percentile.) In other words, although there may be teachers within the bottom 1 percent of teachers who are, the average teacher in the bottom 1 percent is not "statistically significantly" worse than average. Similarly, only a few percentiles of teachers at the top of the distribution in single-year value-added are "statistically significantly" better than average (that is, the curve for the lower bound of the 95 percent confidence interval rises above zero.)

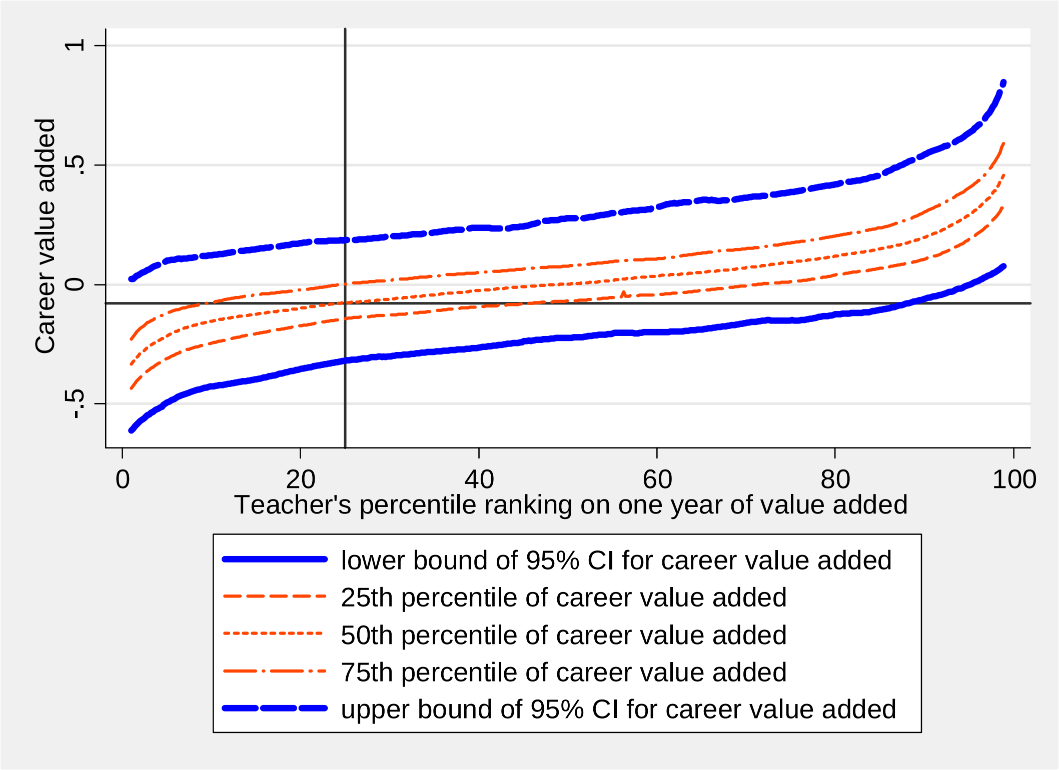
Second, focusing on the distribution of actual career value added in each percentile (the red lines), note than almost 75 percent of the teachers in the bottom 1 percent on single-year value-added had career value-added below -.25. Neal and Johnson (1996) found that an entire year of schooling produces a .25 standard deviation in test gains for the typical student. In other words, 75 percent of the teachers in the bottom one percent caused students to fall behind by the equivalent of a whole year while the students were in their classrooms. At the other extreme, for teachers with single-year value added in the top 1 percent, 75 percent of such teachers had career value added above .25. In other words, the average student in their classes were achieving two years worth of achievement gains in a single year.

Above, we proposed using "better than the average novice teacher" as the threshold for a retention decision. Figure 3 further illustrates the implications of such a rule. Note that median career value-added is equal to the average value-added of a novice teacher for teachers in the 25th percentile. In other words, 25 percent of teachers would fail the "better than the average novice" test based on a single year of value-added, depending on the district. For teachers with single year value-added at that threshold, we could not reject the hypothesis that the teacher was equivalent to the average teacher. (Zero is contained within the 95 percent confidence interval for such teachers.) However, even though we could not pass the classical hypothesis test threshold, 75 percent of such teachers had career value added worse than the average teacher! (The 75th percentile of the career-value added for teachers at that point in the horizontal axis is less than 0.)

What is a "mistake" in the context of a tenure decision, and how do the two rules compare in terms of mistakes made? A supervisor would be making a mistake whenever a tenure offer is made to a teacher whose career value-added is below that of a novice teacher. In such cases, students would have been better off if the principal committed to hire a rookie teacher to fill the slot every year. (This is a conservative estimate. In principle, a supervisor could do even better by promoting and retaining the successful future rookies.) When a teacher's single year value-added is at the 25th percentile, the likelihood of making a mistake is 50-50. When a supervisor offers tenure to a teacher below the 25th percentile, the chances of a mistake rise. For instance, at approximately the 10th percentile, more than 75 percent of teachers will have career value-added worse than the average novice. Yet, the 10th percentile teacher is not "statistically significantly" different from the average teacher.

An example from another field may be useful. Suppose you have had a heart attack and an ambulance arrives to transport you to a hospital emergency room. You can go to one of two hospitals, Hospital A or Hospital B. At Hospital A, the mortality rate for heart attack patients is 75 percent. At Hospital B, the mortality rate is 20 percent. Of course, these mortality rate estimates are subject to fluctuation. But suppose you knew that among hospitals of this size who had 75% mortality, you had a 90% chance that they were better than average over the long run, while the hospital with 20% mortality only had a 10% chance of being better than average Suppose neither rate is "statistically significantly" different from average. In other words, there was not a sufficient number of admissions at either hospital to make the evidence overwhelming that Hospital A is better than average or that Hospital B is worse than average. Would you truly be indifferent which hospital the ambulance driver chose?

Figure 3. Distribution of Career Value Added for Teachers Ranked at Each Percentile Based on One Year of Value Added.



IV. Conclusions

Our results challenge the claim that year-to-year volatility in value added measures is *prima facie* evidence against their use. While value-added measures are unreliable by conventional standards and unstable over time, they are strong predictors of an individual teacher's career performance that can be used to improve decision making.

Our analysis has three key ***implications for practice***:

- Year-to-year instability in value added and other teacher performance measures is misleading. One should instead focus on the correlation between annual performance measures and career performance, which is equal to the square root of the year-to-year correlation.

- Annual value added measures are a fairly powerful predictor of a teacher's career performance despite low year-to-year correlations in value added. The year-to-career correlations for value added measures are in the range of .5-.8.

- The classical hypothesis testing framework, which presumes that a teacher is "average until proven below average," identifies too few teachers as ineffective. Instead, we suggest using whether a teacher's expected career value added lies below a given threshold, such as the performance of a typical novice. Based on this standard, teachers ranked in the bottom 25% of annual value added would be expected to perform below the novice level over the course of their careers.

Schools are unaccustomed to differentiating between teachers. It would be difficult to implement a new teacher evaluation system even if performance could be measured perfectly. The manifest imperfection of the measures makes a difficult implementation even more difficult.

Not surprisingly, many districts have chosen to use teacher performance measures cautiously. The "average until proven below average" criterion is designed to protect the interest of incumbent teachers, just as "innocent until proven guilty" is designed to protect the liberty of the accused in our legal system. However, if the paramount goal were to raise student achievement, to maximize the chance that all students have an effective teacher, and to be fair to both prospective novice teachers as well as incumbent teachers, school systems would be using a different standard. For example, promoting only those teachers who have expected effectiveness higher than the average novice teacher would lead to very different decisions and a different set of outcomes for kids. Most teachers with single year value-added in the bottom quartile will not be "statistically significantly" different from average achievement, yet will perform at levels below a novice teacher throughout their careers. The classical hypothesis testing framework which presumes that a teacher is "average until proven below average" is simply inappropriate for such decisions.

References:

Eva L. Baker, Paul E. Barton, Linda Darling-Hammond, Edward Haertel, Helen F. Ladd , Robe rt L. Linn, Diane Ravitch, Richard Rothstein, Richard J. Shavelson, and Lorrie A. Shepard "Problems with the Use of Student Test Scores to Evaluate Teachers" Economic Policy Institute Briefing Paper No. 27, August 2010.

Camara, W.J. & Echternacht, G. (July 2000). The SAT I and high school grades: Utility in predicting success in college. New York, NY: The College Board.

Chetty, Raj, John N. Friedman, Jonah E. Rockoff, "Measuring the Impacts of Teachers I: Evaluating Bias in Teacher Value-Added Estimates" Working paper, July 2013.

Dimick, J.B., Staiger, D.O., Basur, O., & Birkmeyer, J.D. (2009). Composite measures for predicting surgical mortality in the hospital. Health Affairs, 28(4), 1189-1198.

Goldhaber, Dan, and Hansen, Michael. (Forthcoming) "Is it Just a Bad Class? Assessing the Long-term Stability of Estimated Teacher Performance." *Economica*.

Hill, Heather C., "Evaluating value-added models: A validity argument approach" *Journal of Policy Analysis and Management* Volume 28, Issue 4, pages 700–709, Autumn (Fall) 2009.

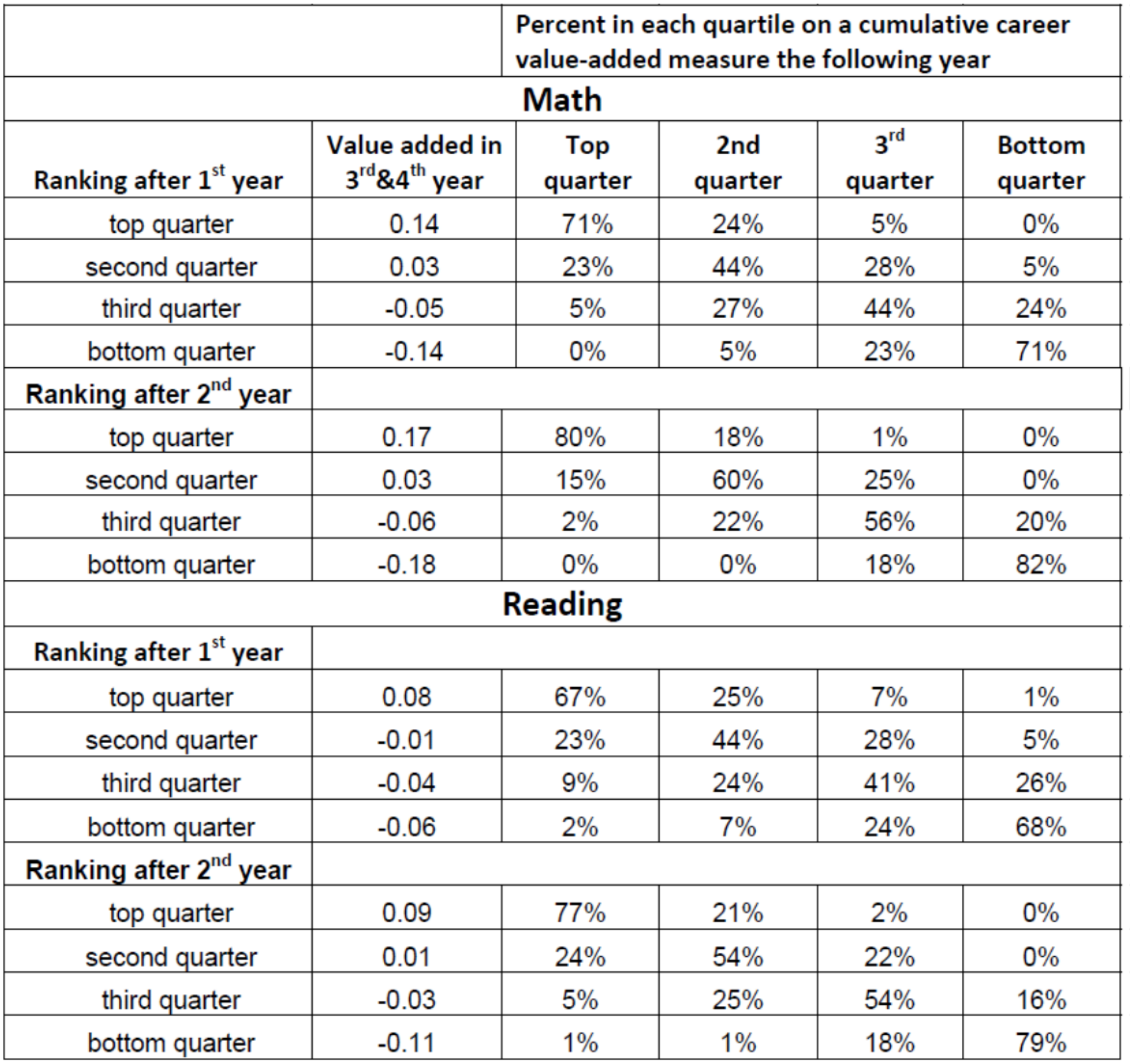
Neal, D.A., Johnson, W.R. (1996) "The Role of Premarket Factors in Black-White Wage Differences." *Journal of Political Economy*, 104(5), pp. 869-95.

Schall, T. & Smith, G. (2000). Do baseball players regress to the mean? The American Statistician, 54, 231-235.

Schochet, Peter Z. and Hanley S. Chiang (2010). Error Rates in Measuring Teacher and School Performance Based on Student Test Score Gains (NCEE 2010-4004). Washington, DC: National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, U.S. Department of Education.

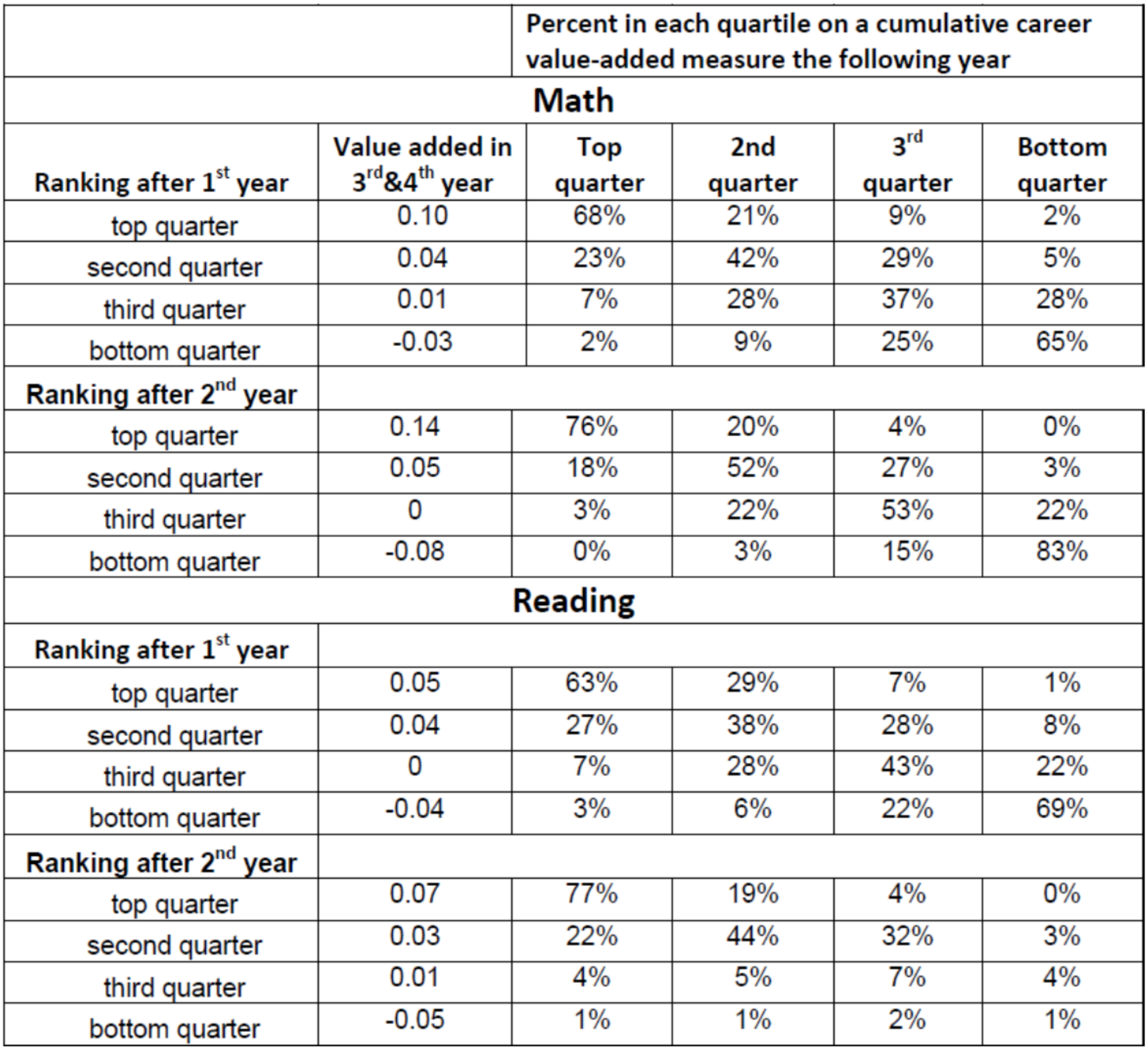
Sturman, M.C., Cheramie, R.A., & and Cashen, L.H. (2005). The impact of job complexity and performance measurement on the temporal consistency, stability, and test-retest reliability of employee job performance ratings. Journal of Applied Psychology, 90, 269-283.

Table 5. Stability of Teacher Value Added Rankings in District 2.



		Percent in each quartile on a cumulative career value-added measure the following year			
Math					
Ranking after 1st year	**Value added in 3rd&4th year**	**Top quarter**	**2nd quarter**	**3rd quarter**	**Bottom quarter**
top quarter	0.14	71%	24%	5%	0%
second quarter	0.03	23%	44%	28%	5%
third quarter	-0.05	5%	27%	44%	24%
bottom quarter	-0.14	0%	5%	23%	71%
Ranking after 2nd year					
top quarter	0.17	80%	18%	1%	0%
second quarter	0.03	15%	60%	25%	0%
third quarter	-0.06	2%	22%	56%	20%
bottom quarter	-0.18	0%	0%	18%	82%
Reading					
Ranking after 1st year					
top quarter	0.08	67%	25%	7%	1%
second quarter	-0.01	23%	44%	28%	5%
third quarter	-0.04	9%	24%	41%	26%
bottom quarter	-0.06	2%	7%	24%	68%
Ranking after 2nd year					
top quarter	0.09	77%	21%	2%	0%
second quarter	0.01	24%	54%	22%	0%
third quarter	-0.03	5%	25%	54%	16%
bottom quarter	-0.11	1%	1%	18%	79%

Table 6. Stability of Teacher Value Added Rankings in District 1.



		Percent in each quartile on a cumulative career value-added measure the following year			
Math					
Ranking after 1st year	**Value added in 3rd&4th year**	**Top quarter**	**2nd quarter**	**3rd quarter**	**Bottom quarter**
top quarter	0.10	68%	21%	9%	2%
second quarter	0.04	23%	42%	29%	5%
third quarter	0.01	7%	28%	37%	28%
bottom quarter	-0.03	2%	9%	25%	65%
Ranking after 2nd year					
top quarter	0.14	76%	20%	4%	0%
second quarter	0.05	18%	52%	27%	3%
third quarter	0	3%	22%	53%	22%
bottom quarter	-0.08	0%	3%	15%	83%
Reading					
Ranking after 1st year					
top quarter	0.05	63%	29%	7%	1%
second quarter	0.04	27%	38%	28%	8%
third quarter	0	7%	28%	43%	22%
bottom quarter	-0.04	3%	6%	22%	69%
Ranking after 2nd year					
top quarter	0.07	77%	19%	4%	0%
second quarter	0.03	22%	44%	32%	3%
third quarter	0.01	4%	5%	7%	4%
bottom quarter	-0.05	1%	1%	2%	1%