THREE ESSAYS IN LABOR ECONOMICS AND THE ECONOMICS OF EDUCATION

By

Brian Stacy

A DISSERTATION

Submitted
to Michigan State University
in partial fulfillment of the requirements
for the degree of

Economics – Doctor of Philosophy

2014

ABSTRACT

THREE ESSAYS IN LABOR ECONOMICS AND THE ECONOMICS OF EDUCATION

By

Brian Stacy

In the first chapter of my dissertation, I examine the robustness of typical teacher quality measures to alternate ranking systems factoring in the dispersion of value-added. The typical measure used by researchers and school administrators to evaluate teachers is based on how the students' achievement increases after being exposed to the teacher, or based on the teacher's "value-added". When teacher value-added is heterogeneous across her students, then the typically used measure reflects differences in the average value-added the teacher provides. However, researchers, administrators, and parents may care not just about the average value-added, but also its variance. Encouragingly, ranking systems factoring in the dispersion produce similar rankings as the ranking system based only on the mean.

In the second chapter, I examine the effect of measurement error in the dependent variable on quantile regression, because unlike OLS regression, even classical measurement error can generate bias. I examine the pattern and size of the bias using both simulation and an empirical example. The simulations indicate that classical error can cause bias and that non-classical measurement error, particularly heteroskedastic measurement error, has the potential to produce substantial bias. Using restricted access Health and Retirement Study data containing matched IRS W-2 earnings records, I examine whether estimates of the returns to education statistically differ using a precisely measured and mismeasured earnings variable. I find that returns to education are over-stated by roughly 1 percentage point at the median and 75th percentile using earnings reported by survey respondents.

In the third chapter, my coauthors and I investigate how the precision and stability of a teacher's value-added estimate relates to student characteristics. We find that the year-to-year stability of teacher value-added estimates can depend on the previous achievement level of a teacher's students.

The stability level of the estimates are typically 25% to more than 50% larger for teachers serving initially higher performing students. We offer a policy simulation demonstrating that teachers who serve low-achieving students may be differentially likely to be the recipient of sanctions in a high stakes policy based on value-added estimates.

I would like to dedicate this dissertation to my future wife, Tina Plerhoples, to my parents, Kathy and Richard, to my brother and sister, Mark and Katie, and to my other family and friends who have helped me in countless ways throughout the years. Some have helped directly on the dissertation, but just as importantly many others have encouraged me, helped me, or simply made my life more enjoyable throughout my time as a graduate student.

ACKNOWLEDGEMENTS

I would like to thank my committee chair, Steven Haider, for the tremendous effort he put into helping me develop, produce, and revise the contents of this dissertation. I would also like to greatly thank my other committee members, Scott Imberman, Mark Reckase, and Jeff Wooldridge for the help, guidance, and patience they also showed throughout the process of producing this dissertation. Several others also greatly aided me on these essays: Quentin Brummet, Steve Dieterle, Cassie Guarino, Tina Plerhoples, all the members of the MSU VAM project group, as well as numerous seminar and conference participants. I would like to thank Dan McCaffrey and JR Lockwood who both taught me a great deal about teacher value-added methods and many other topics, while I was a summer associate at the RAND corporation.

I would also like to acknowledge that the research reported here was supported by the Institute of Education Sciences, U.S. Department of Education, through Grants R305D100028 and R305B090011 to Michigan State University. The opinions expressed are those of the authors and do not represent the views of the Institute or the U.S. Department of Education.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

RANKING TEACHERS WHEN TEACHER VALUE-ADDED IS HETEROGENEOUS

1.1 Introduction

Teacher quality measures based on student achievement data are increasingly being utilized by researchers in topics ranging from the impact of teacher quality on later life outcomes, to the impact of teacher quality on housing prices, to the quality of teachers who transfer or leave the teacher labor force (see, e.g., Chetty et al. (2011), Imberman and Lovenheim (2013), or Boyd et al. (2008) for examples of each). Additionally, federal education policies, such as the Teacher Incentive Fund and the Race to the Top, have sparked substantial demand for rigorous measures of teacher quality by administrators who wish to identify the most and least effective teachers. The most commonly used measures of teacher quality are value-added measures that attempt to isolate a teacher's contribution to student learning in a year.

Some studies make the simplifying assumption that teacher value-added is identical for all students.[1] With this assumption, a "teacher effect" can be estimated for each teacher, which reflect differences in the value-added provided. Other studies explicitly explore heterogeneity in teacher value-added and find evidence that teacher value-added is different for different students.[2] With

[1]The assumption of a constant value-added is explicitly stated in Chetty et al. (2011) for instance, but implicitly assumed in many structural models of achievement used in value-added estimation.

[2]For instance, Dee (2004) examines whether assigning a student to a teacher of the same race improves student achievement using experimental project STAR data, and finds an increase for both black and white students. One year with same race teacher increases achievement 2 to 4 percentile points. Aaronson et al. (2007) computes teacher value-added separately for students with high and low prior year test scores and finds that the correlation between the two is .39. A similar exercise is done by Condie et al. (2014). Loeb et al. (2014) examines whether teachers quality depends on whether a student is an English learner. Lockwood and McCaffrey (2009) examine heterogeneity in teacher value-added by interacting value-added with predicted achievement and find modest interaction effects with the interactions explaining around 10% of the total variation in teacher effects across teachers.

heterogeneity, the "teacher effects" that are typically estimated reflect differences in the mean value-added provided. From here on I will refer to these measures as "value-added means".

Despite the recognition that teacher value-added can be heterogeneous, little work has been done examining teacher quality beyond the value-added means.[3] Teachers may differ in the variance of the value-added they provide, and this information may be important for researchers and administrators forming and using teacher quality ratings. For example, an individual may view a teacher that produces large learning gains for a few students and small gains for the rest differently from a teacher that produces moderate gains for all students. Examining the variance of value-added in addition to the mean can distinguish between these two cases.

In this paper, I examine the sensitivity of teacher rankings to alternate rankings that factor in the variance of teacher value-added. I estimate "value-added variances", which reflect differences across teachers in the variance of the value-added a teacher provides. These can be identified using the same random assignment conditional on observables assumptions made to identify value-added means. I then use this additional information to create alternate rankings, which I compare to the rankings based solely on value-added means.

Using administrative data linking students to teachers from a large, diverse, anonymous state, I find little evidence of a systematic mean-variance trade-off in teacher value-added. The value-added means and variances are in fact negatively correlated (math: -.328, $p<.001$, reading: -.206, $p<.001$). I also find that there are larger differences across teachers in terms of the mean than the variance. As a result, teacher rankings systems incorporating both value-added means and value-added standard deviations are highly correlated with a system only comprised of value-added means. The correlations are above .9 in most cases.

[3] Some exceptions include the papers listed in the footnote above.

1.2 Framework for Evaluating Teacher Quality

A convenient framework for ranking teachers is the potential outcomes framework.[4] For our purposes, the potential outcomes are the potential achievement outcomes if a student is assigned to any of the teachers in the population. Let i denote a randomly drawn student from the population. Let $A_i(j)$ be the achievement level of student i if they are assigned to a particular teacher j.

Administrators and researchers are typically interested in identifying how students would perform if they were assigned to one teacher compared to another. The primary difficulty in making this type of causal inference is that, if there are J potential teachers, it is only possible to observe one of the J potential outcomes for a student.

The key assumption used to make causal inferences about teachers is that assignment of teachers to student is random conditional on X_i, which is a set of observable characteristics of students. With this assumption, even though we do not observe all $A_i(j)$ for each student, we can use the observed outcome, A_i, to estimate teacher effects. This assumption of selection only on observables is sometimes referred to as ignorability (or unconfoundedness) conditional on X_i.[5] This assumption implies that principals base assignment on observable characteristics of students, such as prior year test scores, but do not assign on unobservable factors that affect achievement.

The ignorability assumption has been hotly debated in the value-added literature.[6] Important

[4] See Rosenbaum and Rubin (1983),Rubin (1974), Rubin et al. (2004), or Imbens and Wooldridge (2008) for further background.

[5] See Imbens (2000) for further discussion.

[6] The assumption is not directly testable. However, Rothstein (2010) develops an indirect falsification test based on the idea that future teachers cannot impact contemporaneous test scores, so evidence of a relationship is evidence of a violation of the assumptions. Rothstein finds that the falsification test rejects, suggesting estimates of teacher effects may be biased. However, Goldhaber and Chaplin (2012) and Guarino et al. (2014) both find that such falsification tests may over reject. Also, Guarino et al. (2012) produce simulation evidence that estimators flexibly controlling for prior year test scores and teacher fixed effects are fairly robust across a variety of nonrandom assignment scenarios. Chetty et al. (2011) find that value-added measures controlling for prior year achievement and demographics predict changes in school level achievement when teachers switch schools and predict long term outcomes such as earnings and college attendance. Finally, Kane et al. (2013) examine whether value-added estimates are biased using a large randomized experiment in which students were randomly assigned to teachers within schools. The authors

for my purposes, this assumption is necessary for estimating value-added means. And without further identifying assumptions, we can estimate value-added variances.

A typical way of estimating teacher effects is to estimate the parameters in the following equation for the conditional mean of achievement:[7]

$$E(A_i|X_i,T_i) = (X_i - \mu_X)\beta + T_{i1}\gamma_1 + \cdots + T_{iJ}\gamma_J. \tag{1.1}$$

X_i is a set of control variables. T_{i1} is an indicator variable equal to 1 if assigned to teacher 1 and 0 otherwise, T_{i2} is an assignment indicator for teacher 2, and so on, and γ_j is the teacher effect for teacher j.[8] Under the ignorability assumption, the estimates of γ_j are consistent estimates of the value-added means.

However, in this case, γ_j does not fully characterize the impact of assigning students to a teacher. Teachers may also differ in the variance of the value-added they provide. Teacher value-added may vary for a few reasons. For instance, a teacher's pedagogical style may work well with some students and not others. Also, some teachers may relate better with some students than others, for instance if they are of the same race or gender, which could lead to differences in the value-added provided. Teachers may also deploy more resources at some students than others.[9] Some others may be less able to cater their instruction to the needs of all students in a classroom.

Define σ_j^2 as the value-added variance for teacher j. With γ_j and σ_j^2 we can get a more complete measure of teacher quality than looking at the mean alone. In order to estimate σ_j^2, assume that the conditional variance of achievement has the following function form:

find no evidence of bias in estimators that control for a student's prior achievement scores and demographics.

[7]For instance see Rothstein (2009) or Harris et al. (2011). This achievement model is sometimes motivated using the education production function framework. For more details, see Hanushek (1979) or Todd and Wolpin (2003)

[8] In my parameterization, γ_j is normalized so that it is teacher j's mean level of achievement produced for the average student. A value of zero for γ_j indicates that a teacher produces a mean achievement level of zero for the average student.

[9]Neal and Schanzenbach (2010) find evidence that teachers may target resources at students in the middle of the achievement distribution because of proficiency requirements.

$$Var(A_i|X_i,T_i) = exp(T_{i1}\psi_1 + \cdots + T_{iJ}\psi_J + (X_i - \mu_X)\delta). \quad (1.2)$$

Note that X_i is centered around its mean, μ_X, in (1.1) and (1.2). After centering X_i around its average, one can interpret γ_j and $\sigma_j^2 = exp(\psi_j)$ as teacher j's mean and variance of achievement produced conditional on having the average student.[10] Intuitively, if a teacher produces a larger variance in achievement for the average student (σ_j^2) than another teacher, then this reflects a larger variance in the value-added provided by that teacher, and likewise with the mean.

In order to estimate γ_j and σ_j^2, I use the following procedure based on least squares. I first estimate the parameters in Equation (1.1) using an OLS regression of the student's observed achievement score on $X_i - \bar{X}$ and T_i. Then I form residuals from this initial regression and estimate (1.2) using non-linear least squares of the squared residuals on $X_i - \bar{X}$ and teacher indicators.[11]

1.3 Data

The data come from an administrative data set in a large and diverse anonymous state. Basic student information such as demographic, socio-economic, and special education status are available. 3,341,109 student year observations are available for students in grades 3-6 from years 2001-2007. The data include achievement scores in reading and math on a state criterion referenced test. The test scores are vertically scaled, so that test scores in grades 3-6 are on the same scale. The benefit of the vertical scale is that if, for instance, a student scores a 500 in 4th grade and a 500 in 5th

[10] For clarity, a value of zero for γ_j indicates that a teacher produces a mean achievement level of zero for the average student, and a value of zero for σ_j^2 indicates that a teacher produces a variance of achievement of zero for the average student.

The exponential function is chosen to model the conditional variance instead of a linear function, because a linear function would not guarantee that the predicted conditional variance is positive. Using the exponential function to model a conditional variance dates back in the econometrics literature to Harvey (1976).

[11] To see why the non-linear least squares regression using the squared residuals can consistently estimate the parameters in the conditional variance, note that $Var(A_i|X_i,T_i) = E(\varepsilon_i^2|X_i,T_i)$ by definition, where $\varepsilon_i = A_i - E(A_i|X_i,T_i)$. Because the OLS residuals converge in distribution to ε_i, as noted in Harvey (1976), using the squared residuals in place of ε_i^2 in the NLS regression still produces consistent estimates of the parameters in $E(\varepsilon_i^2|X_i,T_i)$.

grade, then you can interpret this as meaning the student gained no knowledge from 4th to 5th grade. Student-teacher links are available for value-added estimation.

The analysis focuses on mathematics and reading student achievement in grade 6. Grade 6 is chosen for two reasons. First, conditioning on a larger number of previous test scores increases the plausibility that assignment of students to teachers is unrelated to student unobservables. Second, teachers in grade 6 often teach multiple sections in a given year, which increases the number of student observations. The larger number of student observations is important for the precision of the estimates.

I impose some restrictions on the data. Students that cannot be linked with a teacher are dropped, as are students linked to more than one teacher in a school year in the same subject. The analysis focuses on traditional public school students, so students in charter schools are dropped. I also drop teachers with less than 12 student observations because accurately estimating value-added means and variances requires a large number of student observations. In all around one third of the student observations are not used in the analysis. Student level characteristics of the final data set are reported in the Table 1.1. The students in the final sample tend to be somewhat higher achieving, more white, and less likely to be free-and-reduced price lunch or limited English proficient than the students in the original sample.

Table 1.2 reports summary statistics aggregated to the teacher level. There are 5,987 math and 6,606 reading teachers in the sample. There are on average 114.58 and 105.013 student observation per teacher for math and reading teachers respectively. This is important for the precision of the estimates of γ_j and σ_j^2. Student characteristics aggregated to the teacher level are also reported.

1.4 Results

The controls included are similar to other papers in the literature (e.g. Chetty et al. (2011)). The vector of covariates, X_i, includes cubic functions of lagged and twice lagged math and reading scores, indicators for whether the student is a minority, the student's free-and-reduced price lunch status, the student's limited English proficiency status, and gender.

In order to increase the precision of the estimates, I pool student observations across all available years and include year dummies as additional controls. Estimation is done separately for math and reading teachers.

Similar to Rothstein (2009), I standardize test scores so that grade 6 test scores have a population mean of zero and a standard deviation of one. Using the same standardization in each grade keeps the vertical scale intact.[12] Therefore, one test score unit translates into an increase of one standard deviation in achievement for sixth graders.

Based on this, I estimate the measures for the value-added means (γ_j) and the value-added variances (σ_j^2) for the 6,249 mathematics teachers and 6,836 reading teachers. As reported in Table 1.3, the standard deviation of the estimates of γ_j across teachers is .207 in mathematics and .155 in reading.[13] Additionally, going from the teacher at the 50th percentile in the estimated distribution of γ_j to a teacher at the 75th percentile in mathematics increases mean value-added by .13 test score standard deviations. Going from the 50th to 75th percentile in reading increase mean value-added by .092 standard deviations.

The differences across teachers for σ_j^2 are more modest. σ_j^2 has a standard deviation across teachers of .086 in mathematics and .106 in reading. Going from a teacher at the 50th percentile in the estimated distribution of σ_j^2 to a teacher at the 75th percentile increases the variance of value-added by .043 test score standard deviation units. This is 4.3% of the variance of overall achievement. Going from the 50th percentile to the 75th percentile in reading means increasing variance of value-added by .054.

In order to provide some information about the precision of the estimates, I report estimates

[12] With this standardization, grade 5 math test scores have a mean of -.152 and standard deviation of .928. Grade 4 math test scores have a mean of -.763 and standard deviation of .979. Grade 5 reading test scores have a mean of -.209 and standard deviation of .981. Grade 4 reading test scores have a mean of -.413 and standard deviation of .960.

[13] These estimates are in line with what other researchers have found for the standard deviation across teachers for the mean. Kane and Staiger (2010) find a standard deviation adjusted for sampling variation of .143 for mathematics teachers. Aaronson et al. (2007) find an adjusted standard deviation of .193 for mathematics teachers and .113 for reading teachers. Rothstein (2009) finds an adjusted standard deviation of .107 for reading teachers.

of γ_j and σ_j^2 along with their standard errors for select teachers in Table 1.4.[14] Estimates and standard errors for teachers at the 10th, 25th, 50th, 75th, and 90th percentiles of γ_j (top panel) and σ_j^2 (bottom panel) are reported. Additionally, Figures 1.1 and 1.2 show the 95% confidence intervals and standard errors plotted on the number of student observations for a randomly selected subsample of teachers for math and reading.[15] The OLS estimates of γ_j are in the top left. The NLS estimates of σ_j^2 are in the top right.[16] An average standard error at each number of student observations for the OLS estimates of γ_j is displayed in the bottom left, and the standard errors for the NLS estimates of σ_j^2 are in the bottom right.[17] The red lines are when the number of students are 25, 50, and 100 student observations.

One thing to notice from this analysis is that, as more student observations are available for each teacher, the estimates all become more precise. This is evident in Figures 1.1 and 1.2 by noticing that both the confidence intervals and standard errors shrink as the number of student observations increase. Also, the magnitudes of the standard errors do not differ much for γ_j and

[14] I use a bootstrapping technique to produce standard errors for the estimates of γ_j and σ_j^2. In order to keep the number of student observations per teacher fixed for every bootstrap replication, I do sampling with replacement within teachers. To be clear, if there are N observations and N_j observations corresponding to teacher j in the original data set, to produce N observations for each bootstrap sample, draw N_j observations for teacher j, where the N_j observations are randomly drawn with replacement from the set of students assigned to the teacher, and repeat this procedure for all teachers. 100 bootstrap replications were performed. Since estimation of σ_j^2 involves two steps (first forming residuals after an OLS regression of current achievement on the covariates and teacher indicators then NLS of the squared residuals on the covariates and teacher indicators) each bootstrap iteration involves estimation of both steps. The sampling with replacement of teachers done in this paper is similar to bootstrapping approach done in Winters et al. (2012).

[15] The randomly selected subsamples of 584 mathematics teachers and 743 reading teachers were used instead of the entire sample, because the bootstrapping procedure was very time intensive.

[16] I also try a procedure based on Normal quasi-MLE to estimate γ_j and σ_j^2. I parameterize the mean and variance of the normal distribution so that

$$D(A_i|X_i,T_i) = Normal((X_i-\mu_X)\beta + \mathbf{T_i}\gamma, exp(\mathbf{T_i}\psi + (X_i-\mu_X)\delta)) \tag{1.3}$$

The estimates were similar in the two approaches, although the QMLE results were slightly more efficient. Since the QMLE is more complex and more computationally difficult to implement, I chose to present the results for the simpler two-step estimator.

[17] The average standard error at each number of student observations was formed using a polynomial smoother.

σ_j^2.

In the bottom panels of Figures 1.1 and 1.2, I also include cutoffs for whether the measures are accurate enough to distinguish the very best and very worst teachers, which is often a goal of forming teacher quality measures. The upper blue line represents the standard error necessary to say with 95% accuracy that a teacher ranked in the bottom 10% is not in the top 10%. The lower blue line represents the standard error necessary to say that a teacher ranked in the bottom 25% is not in the top 75%.[18] Ideally, the average standard errors should be below the cutoffs. 12 student observations are enough to distinguish teachers at the 90th and 10th percentiles for both γ_j and σ_j^2 in mathematics and reading. When going to the tougher requirement of distinguishing teachers at the 75th and 25th percentiles, 12 student observations is only enough in the case of mathematics for γ_j. 50 student observations is enough in reading at the 75-25 difference with γ_j, and 100 observations is enough in reading for σ_j^2. More than 200 are necessary in mathematics to distinguish between the 75th and 25th percentiles in σ_j^2. This is partially due to the smaller difference in the value-added variances for mathematics teachers at the 75th and 25th percentiles compared to for instance the value-added variances for reading teachers (a gap of .079 versus .147 in reading).

Overall, 12 student observations are enough to distinguish the very worst from the very best for both γ_j and σ_j^2, but in some cases it may be difficult to distinguish teachers toward the center of the distribution without large numbers of student observations. In the remaining analysis, I will continue to use all teachers with more than 12 student observations, but will also explore the results when only teachers with more than 100 student observations are included as a sensitivity check.

[18] I form the blue lines by calculating the difference in γ_j and σ_j^2 at the 90th and 10th percentiles and the 75th and 25th percentiles. For math and γ_j, the 90-10 difference is .48 and the 75-25 difference is .254. In reading and γ_j, the 90-10 difference is .374 and the 75-25 difference is .183. In math and σ_j^2, the 90-10 difference is .176 and the 75-25 difference is .079. In reading and σ_j^2, the 90-10 difference is .223 and the 75-25 difference is .147. I then form the standard error necessary at each of the gaps, by dividing the gap by 1.96.

1.4.1 Correlation between γ_j and σ_j^2

A worry in using only estimates of γ_j in rankings is that teachers that produce high value-added means may be leaving some students behind, producing small gains for these students. In order to examine whether this is the case, in Table 1.3 I report the correlation between $\hat{\gamma}_j$ and $\hat{\sigma_j^2}$, which is -.328 for mathematics and -.206 for reading. Scatterplots for the estimates of γ_j and σ_j^2 are also shown in Figures 1.3 and 1.4. In both cases the correlation is statistically different from 0 at the 1% level.[19] This indicates, contrary to the initial fear, that teachers with higher levels of mean value-added tend also to have a lower variance in value-added. This suggest that, if having a low variance is a good thing, teachers rated favorably along one dimension are more likely to be rated favorably along the other.

This also means that a ranking system that incorporates both $\hat{\gamma}_j$ and $\hat{\sigma_j^2}$ will tend to produce similar rankings as a ranking system that only focuses on mean value-added. Also, because there are fewer differences in σ_j^2 across teachers, then rankings that incorporate information on the teacher's effect on the variance may not differ much from a ranking based solely on the mean effect.

1.4.2 Do Teacher Rankings Change When We Add Information on Value-Added Variances under Plausible Teacher Ranking Functions?

Principals or administrators may be interested in ranking teachers at least in part on the variance of value-added. A teacher that produces a given mean level of value-added, but with a high variance, may generate more complaints from parents than a teacher that produces a similar mean level and a lower variance. Administrators may also have asymmetric payoffs, for instance if they are penalized for having a certain number of students fall below basic proficiency levels, that may make them rate the slightly lower mean, lower variance teacher more highly.[20]

[19] The standard errors for the significance test for the correlations are calculated by bootstrapping.

[20] There may be cases where individuals would prefer a higher variance. For instance, if a school's sole focus was to produce a few super star students, they would want teachers to have a

In the following section I produce teacher rankings under a variety of ranking schemes. I use value-added standard deviations in the ranking function rather than variances, because standard deviations are expressed in the same units as the mean, whereas the variance is expressed in squared units.[21] I use the following simple ranking function:[22]

$$r_j = q\hat{\gamma}_j - (1-q)\hat{\sigma}_j.$$

where r_j is teacher j's ranking and q is a weight put on the value-added mean and value-added standard deviation. I will compare three alternate ranking systems to the rankings based only on γ_j:

Baseline Ranking: Teacher rankings are based solely on the estimate of γ_j

25% on σ_j: Teacher rankings based 75% on estimate of γ_j and 25% on estimate of σ_j

33% on σ_j: Teacher rankings based 67% on estimate of γ_j and 33% on estimate of σ_j

50% on σ_j: Teacher rankings based equally on estimate of γ_j and σ_j

I produce Spearman rank correlations between the baseline ranking system and the three alternate rankings systems in in Table 1.5. The rank correlations are above .94 in mathematics, and above .88 in reading. All rankings are above .96 when less than 1/3 of the weight is placed on the value-added standard deviation and above .98 when less than 25% of the weight is placed on σ_j. Thus, incorporating σ_j into teacher rankings isn't likely to dramatically alter the rankings for most teachers under a variety of alternative ranking systems compared to ranking teachers solely

large variance.

[21] The value-added standard deviations are estimated by taking the square root of the estimated value-added variances.

[22] There are many other potential objective functions, which may not translate exactly into a mean-variance trade off. For instance, a principal may want to maximize the number of students that pass a proficiency level, and suppose that principal wants to assign a teacher to a classroom of students that is initially far below the proficiency level. The principal in this case may want a teacher that produces a large variance in value-added to get more students up to that proficiency level. However, I chose the ranking function in this paper for its simplicity.

on their value-added mean. This result is likely driven by the negative correlation between the value-added mean and standard deviation and the more modest variation in σ_j compared to γ_j.

To add some comparison to the numbers, Goldhaber et al. (2013) compare teacher rankings, based on value-added means, under alternate sets of control variables. The authors find that the correlation between estimates that control for student test scores and demographics and estimates that control for additional peer characteristics is around .99. The correlation between estimates that control for school fixed effects and estimates that do not is only .65. This suggests that the decision to include information on the value-added standard deviation is slightly more consequential than the decision to include peer variables, and much less important than the decision to include school fixed effects.

One caveat is that, even though the correlations are strong, for particular teachers changing the ranking system can have a large impact. In order to provide a rough sense of how far a teacher may be moving using the different rankings, in the bottom panel of Table 1.5, I report the fraction of teachers that move in the rankings $\pm$ 10% of teachers. This corresponds to a move of 625 spots in the rankings for math teachers and 684 spots for reading teachers. Particular teachers can move quite a bit in the rankings in some of the alternate ranking schemes. 22% of teachers move in the rankings $\pm$ the equivalent of 10% of teachers in the case where 50% of the weight is put on the standard deviation in math. However, in the case where 25% of the weight is put on the standard deviation, only 2% of teachers move the equivalent of $\pm$ 10% of teachers.

1.5 Sensitivity Checks

I perform a number of sensitivity checks for the analysis, which are reported in Table 1.6 for mathematics and Table 1.7 for reading. I discuss the results for mathematics in detail below, but the results for reading are similar. Overall, the results for the sensitivity checks are similar to the baseline results.

As I discussed in section 4, the estimates of γ_j and σ_j^2 are not precise enough to distinguish between teachers at the 75th and 25th percentiles teachers when teachers have only 12 student

observations, except in the case of γ_j for mathematics. The imprecision that this reflects could potentially affect the results. In row 2 of Table 1.6, I report results when the sample is restricted to include only teachers with more than 100 observations. I report the correlation between the estimates of γ_j and σ_j^2, the Spearman rank correlations between a system were teachers are ranked only on the mean and a system where 50% of the weight is put on the estimate of σ_j, and the percentage of teachers that move $\pm$ 10% of teachers in the rankings under the alternate ranking system. Overall, the results are similar to the baseline case. The Spearman rank correlations are slightly higher and the percent moving 10% in the rankings is slightly lower when this restriction is imposed compared to the baseline.

As discussed in Goldhaber et al. (2013), there is considerable disagreement about the conditioning variables that are needed for ignorability. It is common to include classroom level peer characteristics or school indicator variables in the regressions. In row 3 of Table 1.6, results for the estimates of γ_j and σ_j^2 when classroom level peer variables are included.[23] The correlation between the estimates of γ_j and σ_j^2 when the classroom level variables are included is -.244. The spearman rank correlation between the alternate ranking system and the ranking based only on the mean is .906, and the percent moving more than 10% is 34%. The correlation is slightly lower, and the percent moving 10% is slightly higher than the baseline case. This may be due to the additional noise in the estimates created by trying to identify the coefficients on the classroom peer variables.

In row 4, I show results from when I estimate value-added variances using a linear functional form rather than an exponential functional form and while keeping the covariate set identical to the baseline specification.[24] I estimate σ_j^2 in an OLS regression of squared residuals from the regression to estimate (1.1) on $X_i - \bar{X}$ and teacher indicator variables. The correlation between the

[23] The peer variables I include are: average prior year math and reading scores, proportion free and reduced-price lunch, and proportion limited English proficient. These coefficients are identified using within teacher variation in classroom composition.

[24] Note that the estimates of σ_j^2 are not guaranteed to be positive using this approach. However, in practice there are only a few instances where σ_j^2 is estimated to be negative for a teacher. In the case with the linear variance but no school fixed effects, only .2% teachers have negative estimates. In the specification, reported below, with school dummy variables and linear variance only .1% teachers have negative estimates.

estimate of σ_j^2 and the estimate of γ_j is -.348. The rank correlation is .919, which is similar to the rank correlation from the baseline specification of .947.

In the final row, I report results from a specification with school dummy variables. Due to computation issues related to finding convergence in the non-linear least squares algorithm when school and teacher indicator variables were both included, I again change the functional form for the variance from an exponential function of the parameters to a linear function. I estimate σ_j^2 in an OLS regression of squared residuals from the regression to estimate (1.1), which also had school indicator variables included, on $X_i - \bar{X}$, school indicator variables, and teacher indicator variables.[25] In this case, the correlation between the estimates of γ_j and σ_j^2 is -.310, and the Spearman correlation drops slightly to .860 compared to .947 in the baseline specification. The percent that move $\pm$ 10% also increases to 35%.

1.6 Summary and Conclusions

Researchers and administrators interested in teacher quality typically produce a single measure of teacher quality. If teachers are having heterogeneous impacts on their students, this measure reflects differences across teachers in the mean value-added they provide, but only examining the effect for the mean may offer an incomplete characterization of a teacher's quality. This paper offers an empirical strategy for identifying measures of value-added variances, and examines how rankings change when this information is added.

There are several important findings in this paper. I find evidence that there are modest to moderate differences across teachers in the size of the value-added variance, but the differences across teachers for σ_j^2 are smaller than differences across teachers for γ_j. Teacher rankings based on the mean and the variance are negatively correlated, with a correlation around -.25. As a result, teacher rankings that include value-added variances tend to be highly correlated with rankings that only include value-added means under some plausible ranking schemes. Typically the correlation

[25] I used the user written felsdvreg package in Stata to estimate the coefficients on the teacher and school indicator variables. The coefficients are identified by teachers switching schools.

is above .9. A positive conclusion from this paper is that rankings using measures of value-added means are fairly robust to adding information on the value-added variance.

This paper also shows that value-added variances can be calculated at fairly low cost. Researchers already computing value-added means by regressing test scores on covariates and teacher indicator variables can estimate value-added variances using the two step approach used in the paper. These estimates could be useful for researchers who wish to study the factors that affect the variance in teacher value-added for instance. More research could be done on this topic. The methods and findings in this paper can serve as a starting point.

APPENDIX

APPENDIX

TABLES AND FIGURES

Table 1.1: Student Level Summary Statistics

Variable	Mean	Std. Dev.
Original Sample		
Number of Student Obs	923,247	
Math Standardized Scale Score	0	1
Reading Standardized Scale Score	0	1
White	0.492	.5
Free and Reduced Price Lunch	0.486	0.5
Limited English Proficiency	0.18	0.384
Female	0.508	0.5
Sample After Restrictions		
Number of Student Obs	685967	
Math Standardized Scale Score	0.074	0.962
Reading Standardized Scale Score	.09	0.956
White	0.497	0.5
Free and Reduced Price Lunch	0.479	0.5
Limited English Proficiency	0.177	0.382
Female	0.512	0.5

Table 1.2: Teacher Level Summary Statistics

Variable	Mean	Std. Dev.
Math Teachers		
Number of Mathematics Teachers	5987	
Student Obs for Math Teachers	114.58	126.303
Student and Teacher Characteristics Aggregated to Teacher level		
Average Prior Year Math Score	-.203	.547
Fraction Free Reduced Price Lunch	0.527	0.257
Fraction Limited English Proficient	0.186	0.215
Fraction White	0.462	0.299
Teacher Experience	7.826	8.85
Reading Teachers		
Number of Reading Teachers	6606	
Student Obs for Reading Teachers	105.013	119.82
Student and Teacher Characteristics Aggregated to Teacher level		
Average Prior Year Reading Score	-0.337	0.611
Fraction Free Reduced Price Lunch	0.521	0.256
Fraction Limited English Proficient	0.181	0.22
Fraction White	0.471	0.299
Teacher Experience	7.711	8.832

Table 1.3: Standard Deviation and Correlations for γ_j and σ_j^2

Statistic	Mathematics	Reading
Std Dev $\hat{\gamma}_j$	0.207	.155
Std Dev $\hat{\sigma_j^2}$	0.086	.106
Correlation $\hat{\gamma}_j$ and $\hat{\sigma_j^2}$	-.328	-.206
Number of Teachers	6249	6836

Controls included in estimation of γ_j and σ_j^2 include a year dummy, cubic functions of lagged and twice lagged math and reading scores, indicators for minority status, free-and-reduced price lunch status, limited English proficiency status, gender, and teacher indicator variables.

Table 1.4: Estimates and Standard Errors of γ_j and σ_j^2 for Select Teachers

γ_j

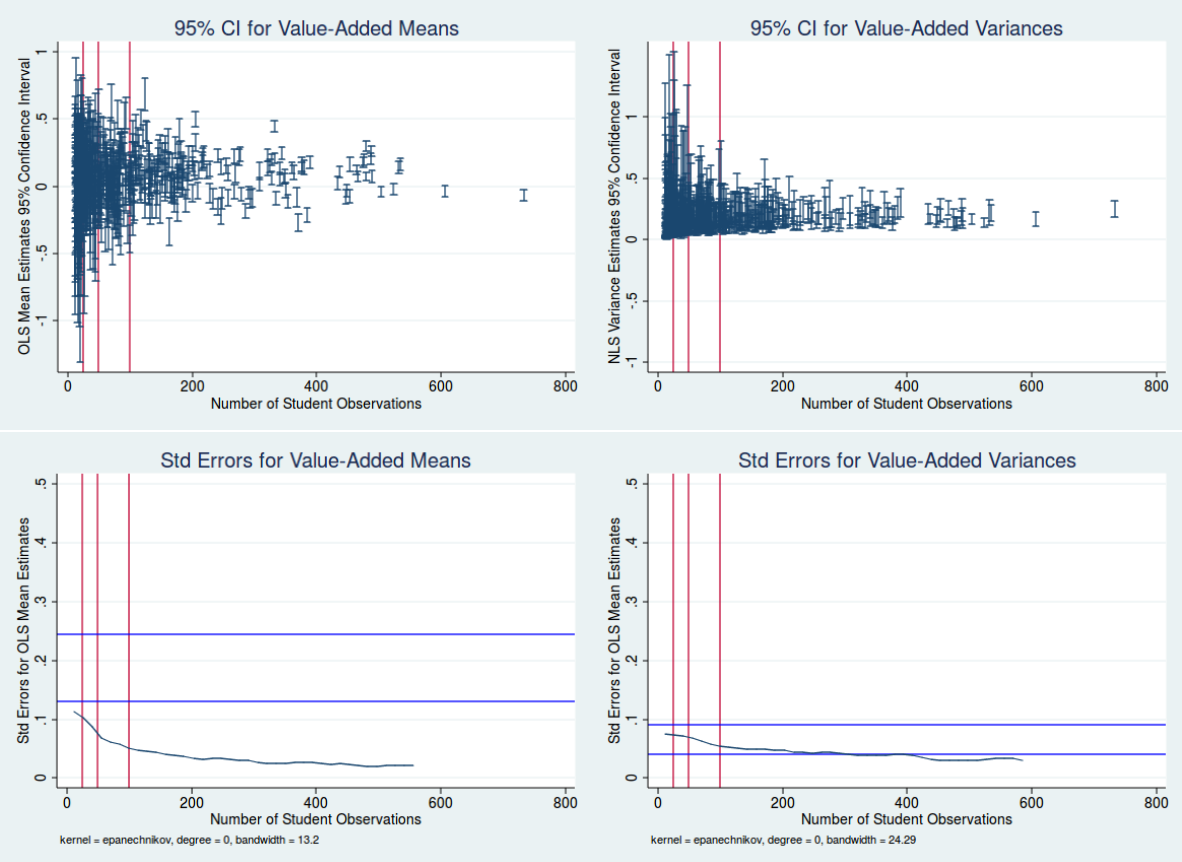
Select Teachers	Mathematics		Reading	
10th Pctl	-.165	(.106)	-.084	(.094)
25th Pctl	-.061	(.081)	.006	(.061)
50th Pctl	.066	(.056)	.092	(.091)
75th Pctl	.194	(.066)	.187	(.066)
90th Pctl	.314	(.086)	.272	(.109)

σ_j^2

Select Teachers	Mathematics		Reading	
10th Pctl	.096	(.035)	.171	(.049)
25th Pctl	.133	(.049)	.220	(.059)
50th Pctl	.181	(.043)	.270	(.073)
75th Pctl	.238	(.059)	.329	(.079)
90th Pctl	.301	(.072)	.405	(.122)
Observations	584		743	

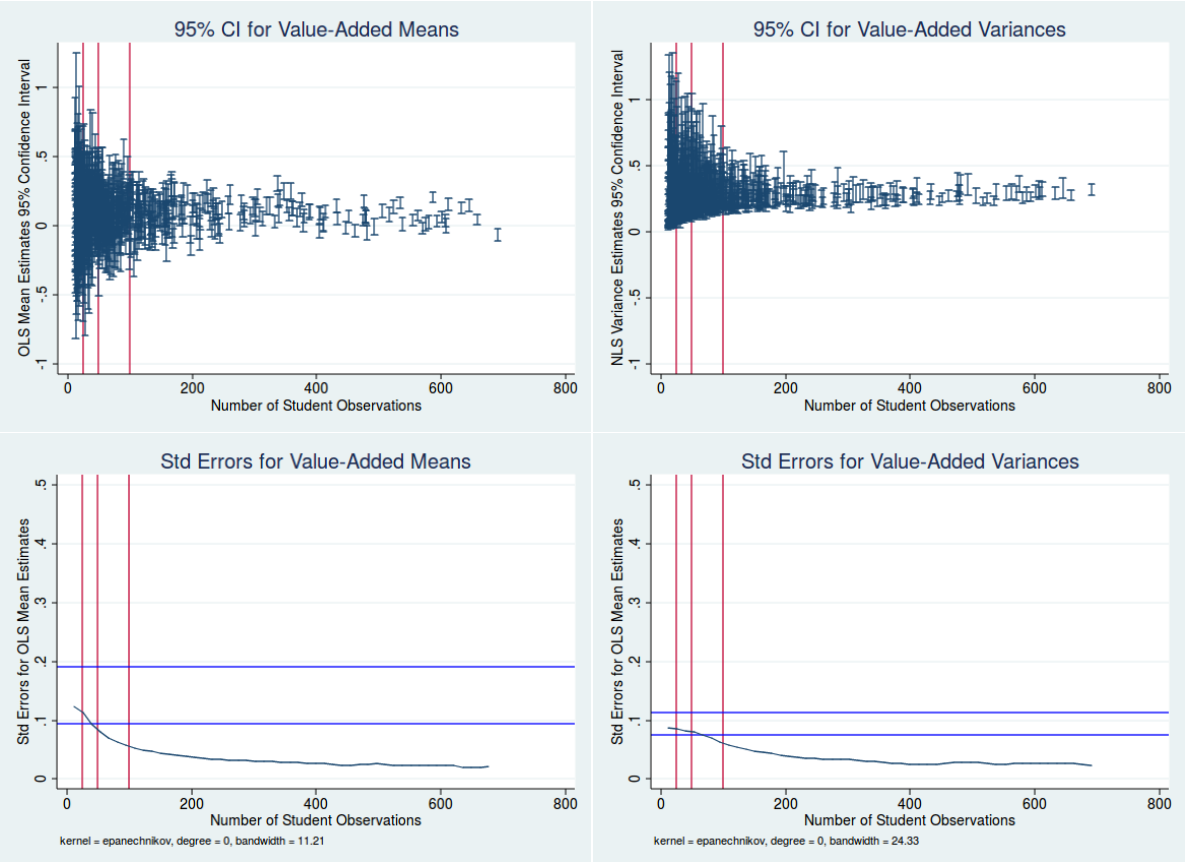
10th Pctl refers to a teacher at the 10th percentile. 25th Pctl refers to a teacher at the 25th percentile and so on. Controls included in estimation of γ_j and σ_j^2 include a year dummy, cubic functions of lagged and twice lagged math and reading scores, indicators for minority status, free-and-reduced price lunch status, limited English proficiency status, gender, and teacher indicator variables.

Figure 1.1: Plots of 95% CI and Standard Errors on the Number of Student Observations for Math Teachers



The OLS estimates of γ_j are in the top left. The NLS estimates of σ_j^2 are in the top right. Average standard errors at each number of student observations, formed using a polynomial smoother, for the OLS estimates of γ_j are in the bottom left, and the average standard errors for the NLS estimates of σ_j^2 are in the bottom right. The red lines are when the number of students are 25, 50, and 100 student observations. The blue lines represent the standard error necessary to statistically reject at the 5% level that a teacher at the 25th percentile is not above the 75th percentile, and that a teacher in the 10th percentile is not above the 90th.

Figure 1.2: Plots of 95% CI and Standard Errors on the Number of Student Observations for Reading Teachers



The OLS estimates of γ_j are in the top left. The NLS estimates of σ_j^2 are in the top right. Average standard errors at each number of student observations, formed using a polynomial smoother, for the OLS estimates of γ_j are in the bottom left, and the average standard errors for the NLS estimates of σ_j^2 are in the bottom right. The red lines are when the number of students are 25, 50, and 100 student observations. The blue lines represent the standard error necessary to statistically reject at the 5% level that a teacher at the 25th percentile is not above the 75th percentile, and that a teacher in the 10th percentile is not above the 90th.

Figure 1.3: Scatterplot of Estimates of γ_j and σ_j^2 for Mathematics

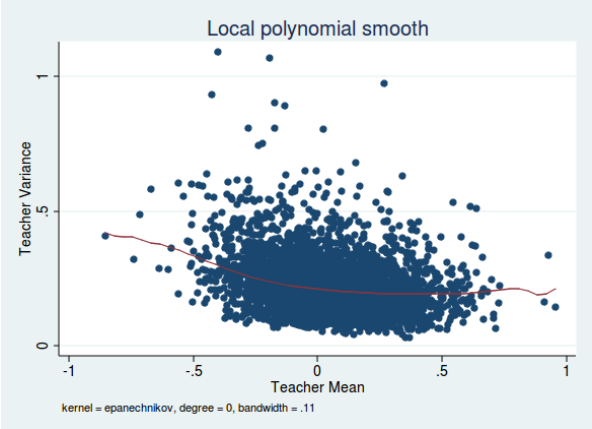


Figure 1.4: Scatterplot of Estimates of γ_j and σ_j^2 for Reading

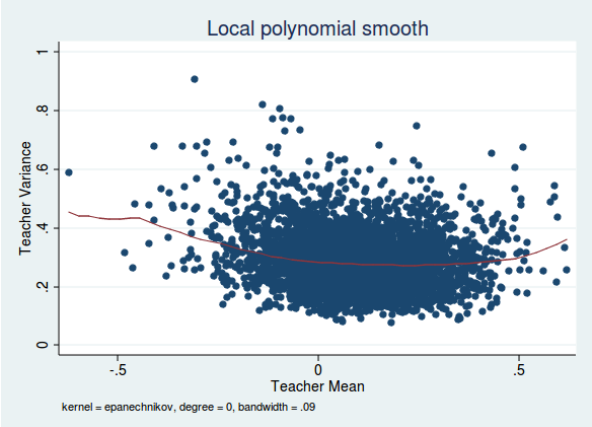


Table 1.5: Comparison of Ranking System Composed of $\hat{\gamma}_j$ and Alternative Ranking Systems Including σ_j

Subject	**25% on σ_j**	**33% on σ_j**	**50% on σ_j**
Spearman Rank Correlation with $\hat{\gamma}_j$			
Mathematics	.993	.985	.947
Reading	.982	.964	.881
Percentage Moving in Rankings 10% of Teachers			
Mathematics	2%	6%	22%
Reading	8%	16%	37%
Math Observations	6249		
Reading Observations	6836		

Controls included in estimation of γ_j and σ_j^2 include a year dummy, cubic functions of lagged and twice lagged math and reading scores, indicators for minority status, free-and-reduced price lunch status, limited English proficiency status, gender, and teacher indicator variables.

Table 1.6: Sensitivity Checks for Mathematics Teachers

Specification	Corr $\hat{\gamma}_j$ and $\hat{\sigma}_j^2$	Spearman 50% on $\hat{\sigma}_j$	Moving $\pm$ 10%
Baseline	-.328	.947	22%
Teachers with $\geq$ 100 Student Obs	-.317	.969	15%
Classroom Level Variables	-.244	.906	34%
Linear Variance	-.348	.919	28%
School Dummy Variables with Linear Variance	-.310	.860	35%

Controls included in baseline estimation of γ_j and σ_j^2 include a year dummy, cubic functions of lagged and twice lagged math and reading scores, indicators for minority status, free-and-reduced price lunch status, limited English proficiency status, gender, and teacher indicator variables.

Table 1.7: Sensitivity Checks for Reading Teachers

Specification	Corr $\hat{\gamma}_j$ and $\hat{\sigma}_j^2$	Spearman 50% on $\hat{\sigma}_j$	Moving $\pm$ 10%
Baseline	-.206	.881	37%
Teachers with $\geq$ 100 Student Obs	-.228	.924	31%
Classroom Level Variables	-.172	.858	40%
Linear Variance	-.221	.844	42%
School Dummy Variables with Linear Variance	-.166	.823	39%

Controls included in baseline estimation of γ_j and σ_j^2 include a year dummy, cubic functions of lagged and twice lagged math and reading scores, indicators for minority status, free-and-reduced price lunch status, limited English proficiency status, gender, and teacher indicator variables.

CHAPTER 2

LEFT WITH BIAS? QUANTILE REGRESSION WITH MEASUREMENT ERROR IN LEFT HAND SIDE VARIABLES

2.1 Introduction

Quantile regression, which allows a researcher to examine the effects of covariates on different points of the conditional distribution of the outcome variable, is an important tool for empirical research. For instance, such methods have been used to examine the returns to schooling (Buchinsky (1994)), inter-generational earnings (Eide and Showalter (1999)), birth weight (Abrevaya and Dahl (2008)), and empirical finance (Chernozhukov and Umantsev (2001)). See Koenker and Hallock (2001) for a review.

Despite it's popularity as an empirical tool, a relatively small literature exists on the effects of measurement error on quantile regression estimation, and within this literature, most of the work has been concentrated on measurement error in independent variables.[1] Almost no research has been done on the issue of bias in quantile regression estimation caused by measurement error in the dependent variable, except for a brief discussion in a footnote in Hausman (2001) and in Chen et al. (2005), who only examine the issue in the context of censored quantile regression at the median.

This lack of research is surprising because, unlike OLS, even classical measurement error in the dependent variable can cause quantile coefficient estimates to be biased.[2] Moreover, many other realistic types of measurement error, such as mean-reverting and heteroskedastic measurement error, complicate matters quickly.[3]

In this paper, I examine bias in the quantile regression estimator caused by measurement error

[1] See Angrist et al. (2006) for example.

[2] Hausman (2001) mentions this fact and that the bias tends to be in the direction of the median coefficient estimate.

[3] Bound and Krueger (1991), Bound et al. (1994), and Pischke (1995) have found evidence that the measurement error is mean reverting, and Hausman (2001) reports that heteroskedastic measurement error may exacerbate bias.

in the dependent variable using simulation and an empirical example. In the simulations, I examine the cases of classical measurement error, mean-reverting measurement error, and heteroskedastic measurement error. My results confirm that the introduction of classical measurement error when the underlying error term is symmetrically distributed can bias the quantile regression estimator towards the coefficient at the median.[4] My results further show that, in cases when the regression error is asymmetric, the estimator can be biased as well, but no clear pattern emerges. The simulations also show that mean reverting and heteroskedastic measurement error can potentially cause bias.

In the empirical application, I examine quantile regression estimates of the returns to education using both reported earnings from the Health and Retirement Study and matched IRS W-2 records, which I assume to be accurate. I find that estimates of the returns to education at the median and 75th percentile are overstated by around 1 percentage point (a bias of around 12-15%) using reported earnings instead of the more accurate W-2 records. These differences are statistically significant at the 5% level. For context, this bias is similar in magnitude to the upward bias caused by omitted ability in the OLS estimator that has been found by others.[5] Also, the pattern of the estimates suggests that the returns to education are less heterogeneous than previously thought.

2.2 Model and Estimator

This section will provide a brief overview of quantile regression. For more details, one can read Koenker and Bassett (1978), Koenker (2005), or Wooldridge (2010) among many other sources.

The goal of quantile regression is typically to examine the effects of covariates on different points of the conditional distribution of the outcome variable. It is common to model conditional quantiles using a model that is linear in parameters. In which case, we can express the τth conditional quantile of y_i as

[4] This finding is also reported in Hausman (2001), although no simulation results are presented.

[5] Upward ability bias in the OLS estimator of the return to education is also around 10-15%, as reported in Card (1999).

$$Q_\tau(y_i|\mathbf{x_i}) = \mathbf{x_i}\beta_0(\tau), \tag{2.1}$$

where $\mathbf{x_i}$ is a vector of covariates, and β_0 is a vector of population parameters.

It can be shown that $\beta_0(\tau)$ satisfies the condition that

$$\min_{\beta \in \Re^K} E[(\tau - \mathbf{1}[y_i - \mathbf{x_i}\beta < 0])(y_i - \mathbf{x_i}\beta)], \tag{2.2}$$

where $\mathbf{1}[\cdot]$ is the indicator function. Assuming that $\beta_0(\tau)$ uniquely satisfies Equation (2.2), the parameters can be consistently estimated under some weak regularity conditions by finding values that satisfy the sample analog.

In many cases, instead of observing the dependent variable y_i, the researcher observes the variable measured with error, call it Y_i. As is well-known, such measurement error causes no bias in the OLS estimator if it follows the classical assumptions.[6] Heuristically, this is the case, because the measurement error is simply absorbed into a composite error term.

This fortunate outcome is not generally the case for the quantile regression estimator for the following reason. Let $u_i = y_i - \mathbf{x_i}\beta$ be the quantile regression error term. In the case of no measurement error, it can be shown that the first order conditions for (2.2) are:

$$E(\mathbf{x_i'}(\mathbf{1}[u_i < 0] - \tau)) = 0. \tag{2.3}$$

When measurement error in the dependent variable is introduced, the first order conditions are:

$$E(\mathbf{x_i'}(\mathbf{1}[u_i + e_i < 0] - \tau)) = 0. \tag{2.4}$$

Because the expected value operator does not pass through the indicator function, the first order conditions are not the same, so there is no guarantee that the parameters that solve (2.4) also solve (2.3) even under classical measurement error.[7]

[6] I define classic measurement error as measurement error that is independent of the true value of the dependent variable and the covariates.

[7] Note that in the OLS case with classical measurement error, the first order conditions with and

2.3 Simulation Evidence of Bias in Quantile Regression

Since no closed form solution exists for the quantile regression estimator, it is difficult to examine bias caused by measurement error in the dependent variable analytically. In order to study the issue further, I produce simulation evidence on how various forms of measurement error affect the quantile coefficient estimates.

My data generating processes consist of a dependent variable with a single explanatory variable. In order to generate different parameters at different quantiles, a random coefficients model is used. My baseline data generating process is meant to be a very simple model of returns to schooling and takes the following form:

$$y_i = \alpha_o + x_i\beta_0 + \gamma x_i\eta_i + \omega_i, \tag{2.5}$$

where η_i and ω_i are independent of one another and x_i and have a standard normal distribution. Note that throughout the discussion of the simulations, I will refer to ω_i as the regression error and e_i, which I will define below, as the measurement error. The regressor x_i, which can be thought of as years of schooling, has a binomial distribution with $n = 16$ and $p = .75$ producing a distribution with a mean of 12 and standard deviation of 3.[8] In my simulation, β is set to .075, α is set to 5, and γ is set to .04. The parameter choices are meant to roughly mimic what is found in previous literature and in my HRS/W-2 data.[9]

In addition to the baseline, I also examine the performance of the estimator under a number of alternate data generating processes. In the second data generating process, I make the effect of x_i negative rather than positive. In the third, I report simulation results in which the effect at the

without classical measurement error are the same. This is true since the expected value operator does pass through linear functions.

[8]These parameters were chosen to give a very basic approximation to the distribution of number of years of schooling. I have also produced simulation results where x_i has a uniform, a normal, and a Poisson distribution. The general patterns are the same as described below.

[9]The choice of .075 is meant to be reflective of the estimates of the mean return to education found previously by other authors, which typically are in the .07-.10 range. For an overview of the mean returns to education literature, see Card (1999).

10th percentile is the largest. In the fourth, I examine the bias caused by measurement error when the effect of x_i is much more heterogeneous. In the fifth, I examine the case where the distribution of ω_i, in equation (2.5), is the Student's t distribution with 3 degrees of freedom instead of the standard normal distribution. This distribution is a symmetric distribution that has a thicker tail than the standard normal distribution. In the sixth, I examine the results when the distribution of ω_i has an asymmetric, lognormal distribution. In this case, $\omega_i = exp(Z_i)$, where Z_i has a standard normal distribution. For this case, the effect at the mean and median will no longer be identical.

I examine three cases of measurement error: classical measurement error, mean-reverting measurement error, and heteroskedastic measurement error.[10] In each case, the measurement error is additive with the form:

$$Y_i = y_i + e_i. \tag{2.6}$$

Simulations were done using Stata. 1,000 simulation repetitions were performed. Each repetition contained 10,000 simulated observations. In the tables, I report the quantile regression estimates at the .10, .25, .50, .75, and .90 quantiles. For purposes of comparison, I also report OLS estimates.

2.3.1 Simulation Results Under Classical Measurement Error

For the simulations with classical measurement error, I assume that e_i is independent of y_i and x_i. Also, e_i is normally distributed with a mean of zero. The variance is chosen so that the reliability ratio is

$$\frac{Var(y_i)}{Var(y_i) + Var(e_i)} = .8. \tag{2.7}$$

[10] Classical measurement error is a case in which the measurement error is independent of the covariates. Mean-reverting measurement error is a case where there is a negative correlation between ω_i and e_i. As discussed in Kim and Solon (2005), one way to interpret mean-reversion found in the measurement error in earnings records is that when workers are asked to report their earnings for the year, the workers under report transitory earnings and shade toward their usual earnings. Heteroskedastic measurement error is a case where the variance of the measurement error depends on the covariates in x_i.

This reliability ratio is approximately the value calculated by Bound and Krueger (1991) for men's reported income in the CPS data. In addition, I examine classical measurement error when the reliability is .6 as a more extreme case.

In row (1) of Table 2.1, I report the results for the baseline specification with normally distributed regression error and classical measurement error with a reliability of .8. In column (2), the quantile regression estimator at the .10 quantile is shown to be biased towards the median coefficient in this simulation. The estimate is .056, while the true value of the parameter is .053 (a bias of roughly 6%) and the median coefficient is .075. The estimator at the .25 quantile is also biased towards the median, but to a lesser degree. The median estimator is unbiased. The estimator at the .75 quantile is slightly biased again towards the median coefficient, and the estimator at the .90 quantile is also biased toward the median coefficient, by an amount nearly symmetric with estimator at the .10 quantile. This pattern is consistent with the pattern reported in the footnote in Hausman (2001) that quantile regression estimators at the tails of the distribution are biased towards the true parameter at the median.

In row (2), I report the estimates when the reliability is .6. The estimates follow the same pattern as those in row (1), but the results show a stronger bias towards the true parameter at the median.

In row (3), I report results in which the coefficient on x_i is negative. In row (4), I report results in which the effect at the 10th percentile is the largest. In both of these cases, the finding that under classical error the estimator at the tails are biased toward the median coefficient holds.

Next, I examine the bias caused by measurement error when the effect of x_i is much more heterogeneous. In this simulation, the bias at the tails is much larger in magnitude than the baseline simulation. As shown in Table 2.2, the bias at the 10th and 90th percentiles is still towards the median, but the magnitude of the bias is around .06 instead of .003 in the baseline simulation (a bias of around 14% instead of 6%). The simulations do not prove this, but they may hint that as the effects at different quantiles become more heterogeneous, the bias becomes larger with classical measurement error. With larger differences between the effect in the tails and the effect at the

median, there may be more room for bias towards the median.

In rows (6) and (7), I change the distribution of the regression error. In row (6), I report results were the distribution of ω_i, in equation (2.5), is the Student's t distribution with 3 degrees of freedom instead of the standard normal distribution. Despite this difference, the results look very similar to the orginal simulation in row (1). The results still display bias towards the median in the case of classical measurement error. Finally, I examine the results when the distribution of ω_i is asymmetric with a lognormal distribution in row (7). An important thing to note is that in this case the coefficients at the tails of the distribution are not necessarily biased towards the median coefficient. Also, the estimator at the median is noticably biased by around 1 percentage point, which was not the case with the symmetric distributions.

A few key points emerge. First, under some alternate simulation parameters and distributions, when the error term, ω_i, is symmetrically distributed, the quantile regression estimator at the tails tend to be biased towards the median coefficient when there is classical measurment error in my simulations. I conjecture that this is true generally for symmetric distributions, but the simulations do not prove this. Second, when the effects across the conditional distribution are relatively more heterogeneous using my data generating process and normally distributed errors, the bias at the tails can be larger. Third, when the error term, ω_i, is asymmetrically distributed, bias still exists and the direction is less clear when there is classical error. The estimator for the coefficient at the median may also be biased.

2.3.2 Simulation Results Under Mean-Reverting Measurement Error

In Tables 2.3 and 2.4, I report estimates when mean-reverting measurement error is added to the dependent variable. In these cases, the measurement error has the following form:[11]

[11] As discussed in Kim and Solon (2005), one way to interpret mean reversion in the measurement error is that when workers are asked to report their earnings for the year, the workers under report transitory earnings and shade toward their usual earnings. In my simulation, this is reflected with a negative correlation between ω_i and e_i.

$$E(e_i|\omega_i) = -.3\omega_i. \tag{2.8}$$

The parameters are chosen to match what is found in Bound and Krueger (1991) for measurement error in log earnings and matches what is found in my HRS data discussed below. As a more extreme case, I also examine, in row (2), mean-reverting measurement error of the form:

$$E(e_i|\omega_i) = -.45\omega_i. \tag{2.9}$$

In row (1) and (2), results are reported for the baseline specification with normally distributed regression error. In these cases, the estimators at the tails of the distribution are biased away from the true parameter at the median in this simulation. In row (1), the bias is -.002 at the .10 quantile (a bias of roughly 4%) and the bias is .001 at the .90 quantile. In row (2) the bias is more pronounced, with a bias of -.004 at the .10 quantile (a bias of roughly 7.5%) and a bias of .004 at the .90 quantile. The OLS estimator and the estimator at the median are unbiased by this form of mean reverting measurement error in this simulation. Rows (3) and (4) show a similar picture as in row (1). The results show a slight bias at the tails of the conditional distribution away from the median coefficient. In rows (5) and (6), the bias is towards the median, meaning that there appears to be no general result for mean reverting error regarding bias towards or away from the median coefficient. Again, in the case with more heterogeneous effects in row (5), the magnitude of the bias for the estimators at the tails of the distribution is much larger than for instance the baseline case in row (1) (a bias of around 13.2% versus 4% in the case of the .10 quantile). Finally in row (7), there again is no clear patter to the bias when the regression error is asymmetric.

2.3.3 Simulation Results Under Heteroskedastic Measurement Error

In the cases of heteroskedastic measurement error, reported in Tables 2.5 and 2.6, the measurement error has the following form:

$$Var(e_i|x_i) = .25exp(-.1x_i + .01x_i^2). \tag{2.10}$$

The parameters are chosen to match what is found empirically in my HRS/W-2 earnings data.[12] In row (2), I again examine a more extreme case that takes the form:

$$Var(e_i|x_i) = .25exp(-.1x_i + .02x_i^2). \tag{2.11}$$

Heteroskedastic measurement error has the potential to produce bias at the tails that is considerably larger than previous cases. The results in row (1) show heteroskedasticity producing a bias of -.019 (a bias of 36% compared to 6% in the baseline case with classical measurement error) in the case of the estimator at the .10 quantile and a bias of .019 in the case of the .90 quantile. The estimators at the .25 and .75 quantiles are also biased, but to a lesser degree. The median estimator does not appear to be biased by heteroskedasticity in the case with the normally distributed regression error. In row (2), which are based on added measurement error with a more extreme form of heteroskedasticity, we see severe bias at the tails of the distribution. The bias at the .10 quantile is -.15 (a bias of 283%), and the bias at the .90 quantile is .186. The estimators at the .25 and .75 quantiles are also substantially biased, while the estimator at the median is largely unbiased. In the other simulation scenarios in rows (3) through (7), substantial bias exists in many cases as well.

Overall, the simulation evidence suggests that the quantile regression estimator can be biased by classical measurement error under a variety of distributions and data generating processes. The bias can potentially be made worse when there is non-classical measurement error, particularly in the case of heteroskedastic measurement error. In this next section, I offer an empirical example showing bias.

2.4 Quantile Returns to Education as an Application

In my empirical example, I use reported earnings in data from the Health and Retirment study benchmarked against what I maintain are more reliable IRS W-2 records data. Bound and Krueger (1991) find that reported earnings in Current Population Study data contains substantial measure-

[12] More details on the data can be found in section 4.1. More details on the approach to estimating the parameters can be found in section 4.2.

ment error when bench-marked against more reliable Social Security earnings records data. Since quantile regression is often applied to income data, the effect of measurement error in these income variables on quantile coefficient estimates is important to understand. I am following a convention in the literature, for instance Chen et al. (2005), maintaining that the administrative earnings records are more reliable.[13]

Buchinsky (1994) has an excellent paper examining the returns to education using quantile regression. I will closely follow the specification in that paper. The regressions are based on the familiar Mincer (1974) equation.

$$log(y_i) = \beta_0 + S_i\beta_1 + E_i\beta_2 + E_i^2\beta_3 + B_i\beta_4 + \varepsilon_i \tag{2.12}$$

where $log(y_i)$ is the log of annual earnings, S_i is years of schooling, E_i is experience, and B_i is an indicator variable for being African American.

I will follow Buchinsky (1994) and estimate parameters for a reduced form equation that does not factor in omitted ability. In addition, I will not address the issue of measurement error in reported years of schooling. The focus of this analysis will be on measurement error in earnings.

2.4.1 Data

The Health and Retirement Study is a survey of over 26,000 Americans (and their spouses) over the age of 50. The purpose of the study was to examine the transition of individuals from the labor force into retirement. The study collects information on income, employment, demographics, as well as on the participants health, retirement assets, and health care expenditures.

Participants are asked to report their total wage earnings, labor force status, age, experience and education level.[14] Importantly for my analysis, many HRS respondents also consented to having

[13]There is good reason to think that the actual dependent variable of interest is permanent income, since the income in any one year may not be an accurate reflection of the return to an additional year of education (see Haider and Solon (2006)). Constructing a measure of permanent income and examining how estimates using this measure compare to using reported annual earnings may be a topic of future research.

[14]In order to keep things as similar as possible to Buchinsky (1994) I use potential experience,

their survey records matched with their W-2 earnings records, which allows me to match reported earnings with the respondent's W-2 records. Haider and Solon (2000) show that the respondents who consented have observable characteristics which are similar overall to the complete sample. The total wage earnings from the W-2 data comes from the box described as, 'Wages, tips, and other compensation'. Income from self employment or income contributed to 401(k) pensions is not included. Income above $250,000 is top coded.

I make a number of sample restrictions in the analysis. I use only the first wave of the study, which took place in 1992-93, since many of the workers, particularly in later waves of the survey, are not prime working age. In the 1992-93 survey, workers are surveyed about earnings in 1991. I exclude women from the analysis in order to avoid sample selection issues with female participation in the labor force. My main set of results includes all workers that have at least $2500 in self-reported and W-2 earnings in 1991 dollars. Summary statistics of the final sample are reported in Table 2.7.

2.4.2 Characteristics of Measurement Error in Log Earnings

I define the measurement error as the difference between log reported earnings and the more accurately measured log W-2 earnings.[15] In this section, I provide an overview of measurment error in my self reported earnings data.[16] Given that non-classical measurement error, and in particular heteroskedastic measurement error, has the potential to exacerbate bias caused by measurement

defined as age minus years of education minus, as my measure of experience instead of years reported working.

[15]To be more clear, the measurement error for observation i, e_i is constructed as:

$$e_i = log(sv\_earn_i) - log(irs\_earn_i),$$

where $log(sv\_earn_i)$ is the log of survey earnings, and $log(irs\_earn_i)$ is the log of IRS W-2 earnings.

[16]Bricker and Engelhardt (2008) also study measurement error in HRS earnings data using the HRS/IRS W-2 matched earnings records. They find evidence of a negative correlation between the measurement error and the true earnings variable. They also find a positive correlation between the measurement error and the education level of the respondent.

error in the dependent variable, I also examine the relationship between the measurement error and the true earnings variable and covariates.

The raw summary statistics of the measurement error are reported in Table 2.8. The mean, standard deviation, and 10th, 25th, 50th, 75th, and 90th percentiles of the measurement error are included in the table. A kernel estimate of the density of the measurement error is included in Figure 2.1. The measurement error in log reported earnings has a mean close to zero and the standard deviation is .486. The measurement error also shows some rightward skewness, with the mean larger than the median.

I examine the degree of mean reversion in the measuremt error in my data by an OLS regression of the measurement error on the log of the true W-2 earnings. As discussed in Kim and Solon (2005), a coefficient of zero for the log of true earnings indicates no mean reversion in the measurement error, and a negative coefficient indicates mean reversion. Results are reported in column (1) of Table 2.9. The coefficient on the log true earnings variable is -.234, which is statistically significant at the 1% level, and is similar to the degree of mean reversion detected in Bound and Krueger (1991), Bound et al. (1994), and Pischke (1995).

Next, I examine the relationship between the measurement error and the covariates. I assume the following functional form for the conditional expectation and variance:

$$E(e_i|S_i,E_i,B_i) = \gamma_0 + S_i\gamma_1 + E_i\gamma_2 + E_i^2\gamma_3 + B_i\gamma_4, \tag{2.13}$$

$$Var(e_i|S_i,E_i,B_i) = \sigma^2 exp(S_i\delta_0 + S_i^2\delta_1 + E_i\delta_2 + E_i^2\delta_3 + B_i\delta_4). \tag{2.14}$$

I estimate the parameters in Equation (2.13) by an OLS regression of e_i on years of education, experience, experience squared, and the indicator for being black. I estimate the parameters in (2.14), which will tell us whether the measurement error is conditionally heteroskedastic, by non-linear least squares of the squared residuals, wich come from the OLS regression to estimate Equation (2.13), on the same covariates.[17]

[17]This produces consistent estimates of the parameters in the conditional variance, because

The results for the conditional mean are reported in column (2) of Table 2.9. The estimated coefficients are insignificant when experience, experience squared, and the indicator for being black are included in column (3). Overall, the estimates suggest a small or negligible effect of the covariates on the conditional mean of the measurement error.[18]

The estimates of the coefficients in Equation (2.14) are reported in column (3). The coefficient on education squared is statistically significant at the 5% level, suggesting that the measurement error is conditionally heteroskedastic. The other estimated coefficients are not statistically significant.

Overall, the measurement error displays mean-reversion and heteroskedasticity. The heteroskedasticity is particularly a cause for concern, because it had such a strong effect in the simulations. In the next section, I report estimates of the returns to education and experience using both the log of reported earnings as the dependent variable and the log of true earnings using W-2 earnings records and test for differences.

2.4.3 Estimates of the Returns to Education and Experience

In order to test whether estimates based on IRS W-2 records statistically differ from estimates based on the reported earnings, I perform the following procedure:

1. Estimate the Mincer equation in (2.12) using reported earnings and again using the (true) W-2 records at the .1, .25, .50, .75, and .9 quantiles.

2. Form the difference between the estimates using the W-2 records and the estimate using reported earnings for each quantile.

3. Repeat the procedure 1000 times sampling with replacement to produce bootstrapped standard errors for the differences between the estimates using reported earnings and (true) W-2

$Var(e_i|S_i,E_i,B_i) = E(v_i^2|S_i,E_i,B_i)$ by definition, where $v_i = e_i - E(e_i|S_i,E_i,B_i)$, and because the OLS residuals converge in distribution to v_i, as noted in Harvey (1976).

[18] These results are consistent with Bound and Krueger (1991), who do a similar analysis in Table 3 of their paper.

earnings records.

In Table 2.10, I show estimates of the returns to education and experience using true W-2 earnings in row (1) and estimates using the reported earnings records in row (2).[19] Stars in row (2) signifiy that the difference between the estimates using W-2 earnings and reported earnings are statistically different from zero. For comparison, the first column shows estimates for the mean from an OLS regression of log annual earnings on years of education, experience, experience squared, and an indicator for whether the respondent is black. Columns (2) through (6) show estimates of the quantile coefficients for the .10, .25, .50, .75, and .90 quantiles.

The return to a year of education at the 10th conditional percentile is estimated to be .046 using log reported earnings and .056 using log W-2 earnings. The return at the 25th percentile is .078 using log reported earnings and .079 for true earnings. However, neither of these differences are statistically significant. At the 50th percentile, the estimated return is .086 for reported earnings and .075 for true earnings. This difference of .0113 is statistically significant at the 5% level. Interestingly, this difference is very similar to the difference found by Chen et al. (2005), who find that using the mismeasured earnings variable biases the censored quantile regression estimate of the return to education at the median by around .014.[20] The estimate at the 75th percentile is .083 and .074 for true earnings, and this difference is also statistically significant at the 5% level. The estimates at the 90th percentile are .090 for reported and .083 for true earnings, but this difference is not statistically significant. These results suggest that returns to education may be overstated at at least the median and 75th percentile.

In the lower two panels, I show returns to experience. Since the Mincer equation in (2.12) includes a quadratic in experience, the return depends on the level of experience of the individual.

[19] I also have examined the returns using only workers with more than 7 years of education and also only workers who report working full time. The patterns are very similar to those reported below.

[20] The authors use the 1978 CPS-SSR match file, which combines reported earnings with social security earnings records. The authors do not report estimates at quantiles other than the median. They also do not report uncensored quantile regression results because of severe top coding in the social security earnings records.

I report the returns at 10 years of experience in the middle panel and 25 years of experience in the lower panel of Table 2.10. Overall, the estimates of the return to experience tend to be low compared to estimates found in the literature. This may be because the Health and Retirement study participants are older, with an average age around 55. At this age, experience may have only a small return. The mean return to experience estimated using OLS and reported earnings is statistically different from the return estimated using true earnings at the 5% level. However, none of the quantile regression estimates statistically differ using the different earnings measures.

2.4.4 Discussion

To summarize the findings, I find with 95% confidence that the effect of an additional year of education at the conditional median and the conditional 75th percentile is overstated when using the mismeasured log earnings variable. The point estimates suggest the estimator at the median is overstated by 1.13 percentage points and that the estimator of the effect at the 75th percentile is overstated by .91 percentage points. The fact that I find the estimator at the median to be biased may suggest that the conditional distribution of true log earnings or the measurement error is asymmetric, given my simulation results. I cannot say with 95% confidence that the estimators at the other quantiles are biased, but this may reflect less precision at the other quantiles. The point estimate at the 90th percentile suggest a bias of .7 percentage points, and the point estimate at the 10th percentile suggest a bias of -.8 percentage points.

The estimates of the effects of experience do not statistically differ. One potential explanation for not detecting bias for experience is that the quantile effects of experience are estimated imprecisely. This is true because there is not much variation in experience in my sample. The lack of precision in the estimates may explain the lack of statistically significant differences.

2.5 Conclusions

This paper makes several important contributions. I add to a small literature on how quantile regression estimates are affected by measurement error in the dependent variable. I show in my simulations that even under classical measurement error the quantile regression estimator may be biased by measurement error in the dependent variable. If one assumes classical measurement error and that the conditional distribution of the true dependent variable is normal, then the simulation evidence suggests that the estimator at the tails of the distribution may be biased towards the median coefficient, although a rigorous proof of this could be a useful topic of future research. In some simulations, the median is also biased, and the size of the bias depends on the amount of heterogeneity in the effects across the distribution and the amount of mean reversion and heteroskedasticity in the measurement error.

Empirically, I show that quantile regression estimator of the returns to education may be biased by measurement error in log reported earnings when compared to the more accurate W-2 earnings records. I find evidence that returns to education estimated at the median and 75th percentile are modestly over stated using reported earnings.

This paper can serve as a cautionary note to researchers using quantile regression techniques with possible mismeasured dependent variables. A bright side is that even though the estimates appear biased, the bias is not overwhelmingly large in the context of the returns to education. The largest bias seen is around 1.13 percentage points. In other contexts, however, the bias could be larger. Future research in other contexts could be useful. Also, finding a solution to the problem may be another important topic for future research.

APPENDIX

APPENDIX

TABLES AND FIGURES

Table 2.1: Simulation Results for OLS/Quantile Regression Estimates with Classical Measurement Error in Dependent Variable.

	OLS	.10	.25	.50	.75	.90
Baseline Spec: Normally Distributed Regression Error						
True Parameter	.075	.053	.063	.075	.087	.097
(1) Estimator w/ Rel .8	.075 (.00023)	.056 (.00037)	.065 (.00031)	.075 (.00028)	.085 (.00032)	.095 (.00039)
(2) Estimator w/ Rel .6	.075 (.00027)	.058 (.00044)	.066 (.00035)	.075 (.00033)	.084 (.00037)	.092 (.00046)
Negative Effect: Normal Regression Error						
True Parameter	-.075	-.097	-.087	-.075	-.064	-.053
(3) Estimator w/ Rel .8	-.075 (.00023)	-.094 (.00037)	-.085 (.00031)	-.075 (.00028)	-.065 (.00032)	-.055 (.00039)
Largest Effect at .10 Quantile: Normal Regression Error						
True Parameter	.075	.088	.082	.075	.068	.061
(4) Estimator w/ Rel .8	.075 (.00022)	.087 (.00036)	.082 (.0003)	.075 (.00027)	.069 (.0003)	.063 (.00037)

The first row of each panel reports the true value of the parameter. The second row reports the estimated coefficient. Rel .8 refers classical measurement error with a reliability ratio of .8. Rel .6 refers classical measurement error with a reliability ratio of .6. The lognormal distribution in row (7) is such that $log(\omega_i)$ has standard normal distribution. Standard errors for the mean across the 1000 reps in parenthesis.

Table 2.2: Simulation Results for OLS/Quantile Regression Estimates with Classical Measurement Error in Dependent Variable.

	OLS	.10	.25	.50	.75	.90
More Heterogeneous Effects: Normal Regression Error						
True Parameter	.077	-.424	-.187	.077	.341	.576
(5) Estimator w/ Rel .8	.076 (.001)	-.366 (.00164)	-.156 (.00132)	.077 (.00125)	.31 (.00136)	.516 (.00169)
Baseline Spec: Student's t w/ 3 d.f. Regression Error						
True Parameter	.075	.055	.062	.075	.088	.094
(6) Estimator w/ Rel .8	.074 (.00037)	.06 (.00059)	.066 (.00041)	.075 (.00035)	.084 (.00042)	.09 (.0006)
Baseline Spec: Lognormal Regression Error						
True Parameter	.075	.044	.067	.087	.091	.086
(7) Estimator w/ Rel .8	.075 (.00045)	.058 (.00043)	.068 (.00035)	.078 (.00038)	.086 (.00053)	.087 (.00106)

The first row of each panel reports the true value of the parameter. The second row reports the estimated coefficient. Rel .8 refers classical measurement error with a reliability ratio of .8. Rel .6 refers classical measurement error with a reliability ratio of .6. The lognormal distribution in row (7) is such that $log(\omega_i)$ has standard normal distribution. Standard errors for the mean across the 1000 reps in parenthesis.

Table 2.3: Simulation Results for OLS/Quantile Regression Estimates with Mean-Reverting Measurement Error in Dependent Variable.

	OLS	.10	.25	.50	.75	.90
Baseline Spec: Normally Distributed Regression Error						
True Parameter	.075	.053	.063	.075	.087	.097
(1) Estimator w/ Mean Reverting	.075 (.00018)	.051 (.00031)	.062 (.00025)	.075 (.00022)	.087 (.00026)	.098 (.00031)
(2) Estimator w/ Larger Mean Reverting	.075 (.00016)	.049 (.00028)	.061 (.00022)	.075 (.0002)	.089 (.00023)	.101 (.00029)
Negative Effect: Normal Regression Error						
True Parameter	-.075	-.097	-.087	-.075	-.064	-.053
(3) Estimator w/ Mean Reverting	-.075 (.00018)	-.099 (.00031)	-.088 (.00025)	-.075 (.00022)	-.063 (.00026)	-.052 (.00031)
Largest Effect at .10 Quantile: Normal Regression Error						
True Parameter	.075	.088	.082	.075	.068	.061
(4) Estimator w/ Mean Reverting	.075 (.00017)	.09 (.0003)	.083 (.00023)	.075 (.00022)	.067 (.00024)	.06 (.0003)

The first row of each panel reports the true value of the parameter. The second row reports the estimated coefficient. Mean-Reverting refers to mean-reverting measurement error with the following form: $E(e_i|\omega_i) = -.3\omega_i$. Larger Mean-Reverting refers to measurement error with the following form: $E(e_i|\omega_i) = -.45\omega_i$. The lognormal distribution in row (7) is such that $log(\omega_i)$ has standard normal distribution. Standard errors for the mean across the 1000 reps in parenthesis.

Table 2.4: Simulation Results for OLS/Quantile Regression Estimates with Mean-Reverting Measurement Error in Dependent Variable.

	OLS	.10	.25	.50	.75	.90
More Heterogeneous Effects: Normal Regression Error						
True Parameter	.077	-.424	-.187	.077	.341	.576
(5) Estimator w/ Mean Reverting	.077 (.00098)	-.368 (.00162)	-.159 (.00126)	.078 (.00119)	.312 (.00133)	.521 (.0016)
Baseline Spec: Student's t w/ 3 d.f. Regression Error						
True Parameter	.075	.055	.062	.075	.088	.094
(6) Estimator w/ Mean Reverting	.075 (.00029)	.056 (.00046)	.065 (.00033)	.075 (.0003)	.085 (.00035)	.093 (.00045)
Baseline Spec: Lognormal Regression Error						
True Parameter	.075	.044	.067	.087	.091	.086
(7) Estimator w/ Mean Reverting	.075 (.00034)	.058 (.00042)	.067 (.00034)	.077 (.00034)	.086 (.00042)	.091 (.00072)

The first row of each panel reports the true value of the parameter. The second row reports the estimated coefficient. Mean-Reverting refers to mean-reverting measurement error with the following form: $E(e_i|\omega_i) = -.3\omega_i$. Larger Mean-Reverting refers to measurement error with the following form: $E(e_i|\omega_i) = -.45\omega_i$. The lognormal distribution in row (7) is such that $log(\omega_i)$ has standard normal distribution. Standard errors for the mean across the 1000 reps in parenthesis.

Table 2.5: Simulation Results for OLS/Quantile Regression Estimates with Heteroskedastic Measurement Error in Dependent Variable.

	OLS	.10	.25	.50	.75	.90
Baseline Spec: Normally Distributed Regression Error						
True Parameter	.075	.053	.063	.075	.087	.097
(1) Estimator w/ Heteroskedastic	.075 (.00023)	.034 (.00038)	.053 (.00031)	.075 (.00028)	.097 (.00033)	.116 (.0004)
(2) Estimator w/ Larger Heteroskedastic	.075 (.00034)	-.133 (.00049)	-.035 (.00041)	.075 (.00039)	.184 (.00042)	.282 (.00052)
Negative Effect: Normal Regression Error						
True Parameter	-.075	-.097	-.087	-.075	-.064	-.053
(3) Estimator w/ Heteroskedastic	-.075 (.00023)	-.116 (.00038)	-.097 (.00031)	-.075 (.00028)	-.053 (.00033)	-.034 (.0004)
Largest Effect at .10 Quantile: Normal Regression Error						
True Parameter	.075	.088	.082	.075	.068	.061
(4) Estimator w/ Heteroskedastic	.075 (.00022)	.067 (.00038)	.07 (.0003)	.075 (.00028)	.079 (.00031)	.084 (.00038)

The first row of each panel reports the true value of the parameter. The second row reports the estimated coefficient. Heteroskedastic refers to heteroskedastic measurement error with the following form: $Var(e_i|x_i) = .25exp(-.1x_i + .01x_i^2)$. Larger Heteroskedastic refers to measurement error with the following form: $Var(e_i|x_i) = .25exp(-.1x_i + .02x_i^2)$. The lognormal distribution in row (7) is such that $log(\omega_i)$ has standard normal distribution. Standard errors for the mean across the 1000 reps in parenthesis.

Table 2.6: Simulation Results for OLS/Quantile Regression Estimates with Heteroskedastic Measurement Error in Dependent Variable.

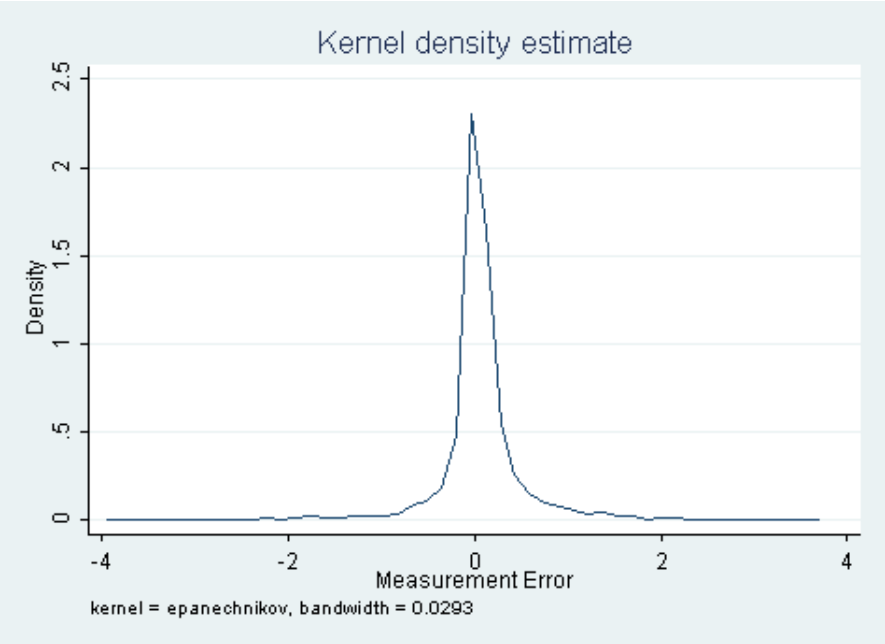
	OLS	.10	.25	.50	.75	.90
More Heterogeneous Effects: Normal Regression Error						
True Parameter	.077	-.424	-.187	.077	.341	.576
(5) Estimator w/ Heteroskedastic	.077 (.001)	-.461 (.00159)	-.205 (.00125)	.078 (.00119)	.36 (.00128)	.614 (.00162)
Baseline Spec: Student's t w/ 3 d.f. Regression Error						
True Parameter	.075	.055	.062	.075	.088	.094
(6) Estimator w/ Heteroskedastic	.075 (.00037)	.014 (.0006)	.04 (.00041)	.075 (.00036)	.11 (.00042)	.137 (.00059)
Baseline Spec: Lognormal Regression Error						
True Parameter	.075	.044	.067	.087	.091	.086
(7) Estimator w/ Heteroskedastic	.075 (.00045)	-.011 (.00044)	.037 (.00038)	.089 (.00039)	.134 (.00053)	.145 (.00106)

The first row of each panel reports the true value of the parameter. The second row reports the estimated coefficient. Heteroskedastic refers to heteroskedastic measurement error with the following form: $Var(e_i|x_i) = .25exp(-.1x_i + .01x_i^2)$. Larger Heteroskedastic refers to measurement error with the following form: $Var(e_i|x_i) = .25exp(-.1x_i + .02x_i^2)$. The lognormal distribution in row (7) is such that $log(\omega_i)$ has standard normal distribution. Standard errors for the mean across the 1000 reps in parenthesis.

Table 2.7: Summary Statistics, Wave 1 (1992) Male Workers with Positive Earnings

Variable	**Mean**	**Std. Dev.**	**Min.**	**Max.**
Total Reported Annual Earnings	36052.17	28173.51	2800	410000
Total Annual Earnings W-2	33157.61	25492.41	2600	245000
Hours Worked/Week Main Job	43.69	10.52	1	95
Weeks Worked/Year Main Job	50.43	5.47	1	52
Hourly Wage Rate	27.74	486.15	0.96	24000
Years of Tenure Current Job	15.46	11.78	0	55.8
Total Years Worked	37.46	5.92	3	65
Total Years of Education	12.7	3.29	0	17
Age	55.87	4.61	23	77
Black	0.129	0.335	0	1
Hispanic	.091	0.288	0	1
Number of Observations				2975

Figure 2.1: Kernel Estimate of the Density of Measurement Error in Log Earnings



Measurement error defined as difference between log reported earnings and log W-2 earnings.

Table 2.8: Measurement Error Descriptive Statistics

			Quantiles				
Variable	**Mean**	**Std. Dev.**	**.10**	**.25**	**.50**	**.75**	**.90**
Measurement Error	.060	.486	-.268	-.051	.032	.166	.443
Number of Observations						2975	

Measurement error defined as difference between log reported earnings and log W-2 earnings.

Table 2.9: Estimates of Conditional Distribution of Measurement Error

	Mean		**Variance**
VARIABLES	(1)	(2)	(3)
Log W-2 Earnings	-.234***		
	(.018)		
Education		.001	-.103
		(.004)	(.078)
Education Squared			.008**
			(.004)
Experience		-.019	-.079
		(.014)	(.076)
Experience Squared		.0002	.001
		(.0002)	(.001)
Black		-.019	.069
		(.026)	(.190)
Observations	2,975	2,975	2,975

Estimates in column (1) come from OLS regression of measurement error on log true earnings. Estimates in column (2) come from an OLS regression of the measurement error on the covariates. Estimates in column (3) come from an NLS regression of the squared residuals from the OLS regression in column (3) on the covariates. Robust standard errors in parentheses. *** p<0.01, ** p<0.05, * p<0.1

Table 2.10: Estimates of Mincer Equation: Male Workers with Positive Earnings

	OLS	.10	.25	.50	.75	.90
Returns to Education						
Using (True) W-2 Earnings	.073	.054	.079	.075	.074	.083
	(.005)	(.014)	(.007)	(.005)	(.005)	(.006)
Using Reported Earnings	.075	.046	.078	.086**	.083**	.090
	(.005)	(.014)	(.006)	(.006)	(.005)	(.007)
Returns to Experience at 10 Years						
Using (True) W-2 Earnings	.016	.065	.016	.002	.003	.001
	(.011)	(.041)	(.016)	(.013)	(.012)	(.019)
Using Reported Earnings	.002*	.032	.012	-.003	-.007	.001
	(.011)	(.023)	(.021)	(.013)	(.010)	(.026)
Returns to Experience at 25 Years						
Using (True) W-2 Earnings	.004	.020	.002	-.002	-.000	.002
	(.005)	(.022)	(.008)	(.006)	(.001)	(.001)
Using Reported Earnings	-.005**	-.000	-.002	-.005	-.004	.001
	(.006)	(.011)	(.010)	(.006)	(.005)	(.012)
Number of Observations				2975		

All regressions include years of education, experience, experience squared, and an indicator for whether black. All workers have at least $2500 in reported and W-2 earnings in 1991 dollars. Bootstrapped standard errors in parenthesis. 1000 bootstrap replications performed. *** Difference between estimates using W-2 and reported earnings statistically significant at 1% level. ** Difference statistically significant at 5% level. * Difference statistically significant at 10% level

CHAPTER 3

DOES THE PRECISION AND STABILITY OF VALUE-ADDED ESTIMATES OF TEACHER PERFORMANCE DEPEND ON THE TYPES OF STUDENTS THEY SERVE?

This work is coauthored with Cassandra Guarino, Mark Reckase, and Jeff Wooldridge.

3.1 Introduction

Teacher value-added estimates are increasingly being used in high stakes decisions. Many districts are implementing merit pay programs or moving toward making tenure decisions based at least partly on these measures. It is important to understand the chances that a teacher will be misclassified in a way that may lead to undeserved sanctions.

Misclassification rates depend on the precision of teacher effect estimates, which is related to a number of factors. The first is the number of students a teachers is paired with in the data. Teachers that can be matched with more student observations will tend to have more precise teacher effect estimates.

Another factor that can affect the precision of a teacher effect estimate is the error variance associated with students in the teacher's classroom. If the error variance is large, perhaps because the model poorly explains the variation in achievement or because the achievement measures themselves poorly estimate the true ability level of a student, then the precision of a teacher effect estimate will be low.

A question that seems to have lacked much attention is whether the precision varies by the characteristics of the students a teacher faces. Tracking of students into classrooms and sorting of students across schools means that different teachers may face classrooms that are quite different from one another. If it is found that teachers serving certain groups of students have less reliable estimates of value-added than other teachers serving other students, then all else the same, the probability that a teacher is rated above or below a certain threshold will be larger for teachers

serving these groups. High stakes policies that reward or penalize teachers above or below a certain threshold will then, again all else the same, impose sanctions or rewards on teachers serving these groups with a higher likelihood.

There are some reasons for suspecting that the characteristics of students in a classroom relates to the precision of teacher effect estimate. First, there could be a relationship between the characteristics of a classroom and the number of students linked to a teacher. This could be true because of a relationship between class size and student characteristics, because of poor data management for schools serving certain groups, or because of low experience levels for teachers serving certain groups, which limit the number of years that can be used to estimate the teacher's value-added.

Also, heteroskedastic student level error can imply that teachers paired with those students with large error variances may have less reliable teacher effect estimates. There is strong theoretical reason for supposing that the student level error is heteroskedastic. Item response theory suggests that because test items are typically targeted towards students in the center of the achievement distribution, achievement tends to be measured less precisely for students in the tails. The heteroskedasticity is also quite substantial, and suggests that teachers paired with particularly high achieving or low achieving students may have less reliable teacher effect estimates. In addition to heteroskedasticity caused by poor measurement, it is also conceivable that the error variance for true achievement is different for different students.

In the remainder of the paper, we test for heteroskedasticity in the student level error term. In addition, year-to-year stability coefficients, which are very similar to year-to-year correlations, using a variety of commonly used value added estimators are computed for teachers serving different groups of students. Year to year stability coefficients for teachers with students in the bottom quartile, top quartile, and middle two quartiles in classroom level prior achievement are compared to one another.

A test of the homoskedasticity assumption easily rejects. Also, large and statistically significant differences in the stability coefficients among sub groups of teachers are detected, and the differences persist even after the number of student observations for all teachers is artificially created

to be the same and when two years of data are used to compute value added. In many cases, the year-to-year stability coefficients are 25 to more than 50% larger in size for teachers serving initially higher achieving students compared to teachers serving lesser achieving and disadvantaged students.

This finding has several implications. For practitioners implementing high stakes accountability policies, teachers serving certain groups of students may be unfairly targeted for positive or negative sanctions simply because of the composition of their classroom and the variability this creates for their estimates. In this paper, we produce simulation evidence that bears this out. In addition, the heteroskedasticity makes it important for researchers and practitioners to make standard errors heteroskedasticity robust. Also, heteroskedasticity is a potential source of bias for those using empirical Bayes value-added estimates, which assume homoskedasticity.

3.2 Previous Literature

A few studies have examined the stability and precision of teacher effect estimates. Aaronson et al. (2007) examined the stability of teacher effect estimates using three years of data from the Chicago public school system. They find that there is considerable inter-year movement of teachers into different quintiles of the estimated teacher quality distribution, suggesting that teacher effect estimates are somewhat unstable over time. They also find that teachers associated with smaller number of student observations are more likely to be found in the extremes of the estimated teacher quality distribution.

Koedel and Betts (2007) perform a similar analysis as Aaronson et al. (2007) using two years of data from the San Diego public school system and also find that there is considerable movement of teachers across quintiles.

McCaffrey et al. (2009) found year-to-year correlations in teacher value added to be .2 to .5 for elementary school teachers and .3 to .7 for middle school teachers using data from 5 county level school districts from the state of Florida from the years 2000-2005. They find that averaging teacher effect estimates over multiple years of data improves the year-to-year stability of the value-

added measures.

This paper adds to the previous literature by specifically looking at whether the stability of teacher effect estimates is related to the characteristics of the students assigned to the teacher.

3.3 Data

The data come from an administrative data set in large and diverse anonymous state. It consists of 2,985,208 student year observations from years 2001-2007 and grades 4-6. Student-teacher links are available for value-added estimation. Also, basic student information, such as demographic, socio-economic, and special education status, are available. Teacher information on experience is also available. The data include vertically scaled achievement scores in reading and math on a state criterion referenced test. The analysis will focus on value-added for mathematics teachers.

We imposed some restrictions on the data in order to accurately identify the parameters of interest. Students who cannot be linked with a teacher are dropped, as are students linked to more than one teacher in a school year in the same subject. Students in schools with fewer than 20 students are dropped, and students in classrooms with fewer than 12 students are dropped. Districts with fewer than 1000 students are dropped to avoid the inclusion of charter schools in the analysis, which may employ a set of teachers who are somewhat different from those typically found in public schools. Characteristics of the final data set are reported in Table 3.1.[1]

The analysis presented later is done separately for 4th grade and 6th grade. This is done because the degree of tracking may be different in 6th grade from 4th grade, which may cause differences in the year-to-year stability of value-added estimates.

3.4 Model

The model of student achievement will be based on the education production function, which is laid out in , Todd and Wolpin (2003), Harris et al. (2011), and Guarino et al. (2012), among other

[1]These restrictions eliminated about 31.2% of observations in 4th grade and 19% in 6th grade

places.[2] Student achievement is a function of past achievement, current student and class inputs, along with a teacher effect.

$$\begin{aligned} A_{igt} &= \tau_t + \lambda_1 A_{ig-1t} + \lambda_2 A^{alt}_{ig-1t} + X_{igt}\gamma_1 \\ &\quad + \bar{X}_{igt}\gamma_2 + T_{igt}\beta + v_{igt} \end{aligned} \tag{3.1}$$

with

$$v_{igt} = c_i + \varepsilon_{igt} + e_{igt} - \lambda_1 e_{ig-1t} - \lambda_2 e^{alt}_{ig-1t}$$

where A_{igt} is student i's test score in grade g and year t. τ_t is a year specific intercept. A^{alt}_{ig-1t} is the test score in the alternate subject, which in the analysis presented below is the reading score. X_{igt} is a vector of student level covariates including free and reduced price lunch and limited English proficiency status, gender, and race. $\bar{X}_{igt}$ consists of class level covariates, including lagged achievement scores, class size, and demographic composition. T_{igt} is a vector of teacher indicators. The teacher effects are represented in the β vector. c_i represents a student fixed effect. ε_{igt} represents an idiosyncratic error term affecting achievement. e_{igt} is measurement error in the test scores with e^{alt}_{igt} representing the measurement error in the alternate subject score.

3.4.1 Estimation Methods

Teacher effects were estimated using two commonly used value-added estimators.[3]

The first is a dynamic OLS estimator (DOLS), which includes teacher indicators in an OLS regression based on equation (1).[4] The estimator is referred to as dynamic because prior year

[2]The model shown includes a lagged score of the alternate subject, which isn't necessary under the assumptions typically made in deriving the regression model based on the education production function. However, including this variable is common in practice, so we chose to include it as well.

[3]We have studied two more estimators based on a gain score equation. One estimator based on teacher fixed effects, and another based on empirical Bayes. The patterns for these two other estimators are similar to those reported for DOLS and EB Lag.

[4]This estimator was found to be the most robust of all the estimators evaluated in Guarino et al. (2012)

achievement is controlled for on the right hand side. The coefficients on the teacher indicator variables are interpreted as the teacher effects. We run our models using one year of data and again using two years of data. Because the effects of class average covariates are not properly identified in a teacher fixed effects regression with only one year of data, these variables are dropped from the DOLS regressions.[5] Additionally, when one year of data is used to estimate value-added, the year specific intercepts are dropped.

The second is an empirical Bayes estimator (EB Lag) which treats teacher effects as random. The estimator follows closely the approach laid out in Kane and Staiger (2008). The parameters of the control variables are estimated in a first stage using OLS, then unshrunken teacher effect estimates are formed by averaging the residuals from the first stage among the students within a teacher's class. The shrinkage term is the ratio of the variance of persistent teacher effects to the sum of the variances of persistent teacher effects, idiosyncratic classroom shocks, and average of the individual student shocks.[6] Teacher effects are interpreted as the shrunken averaged residuals for each teacher.

[5]We have tried a two step method that can identify the effect of class average covariates in a teacher fixed effects regression as a sensitivity check, and the results are similar. First, using the pooled data with multiple years, equation (1) is estimated using OLS with teacher fixed effects included. Then, a residual is formed.

$$\begin{aligned} w_{igt} &= A_{igt} - \hat{\tau}_t - \hat{\lambda}_1 A_{ig-1t} - \hat{\lambda}_2 A^{alt}_{ig-1t} - X_{igt}\hat{\gamma}_1 - \bar{X}_{igt}\hat{\gamma}_2 - \hat{f}(exper_{igt}) \\ &= T_{igt}\beta + \hat{v}_{igt} \end{aligned}$$

which is then used in a second stage regression to form teacher effects using a sample based on 1 year of data.

[6]It is common to treat the variance of the individual student shocks as uniform across the population of students. In an effort to evaluate commonly used estimators, we also computed the shrinkage term by using the same variance term for the student level shocks for all teachers. Under heteroskedasticity, this shrinkage term would not be the shrinkage term used by the BLUP.

3.5 Heteroskedastic Error

There is good reason to suspect that the error in the student achievement model is heteroskedastic. We will present some basic theory suggesting that measurement error in test scores is heteroskedastic. Also, we will offer some possible reasons why the error variance of actual achievement may be heteroskedastic.

3.5.1 Heteroskedastic Measurement Error

Item response theory is typically the foundation for estimating student achievement. A state achievement test is typically composed of 40-50 multiple choice questions, or items. Each student can either answer a question correctly or incorrectly, and the probability of answering any individual question is assumed to be a function of the item characteristics and the achievement level of the student. The typical model of a correct response to an item assumes (See Reckase (2009) for more details):

$$Prob(u_{ij} = 1|a_i, b_i, c_i, \theta_j) = c_i + (1 - c_i)G(a_i(\theta_j - b_i)$$

where u_{ij} represents an incorrect or correct response to item i by student j. a_i is a discrimination parameter, b_i is a difficulty parameter, and c_i is a guessing parameter for item i. θ_j is the achievement level of student j. Often, a logit functional form is assumed for $G(\cdot)$, although the probit functional form is also used. In the case of the logit form we have:

$$Prob(u_{ij} = 1|a_i, b_i, c_i, \theta_j) = c_i + (1 - c_i)\frac{e^{(a_i(\theta_j - b_i))}}{1 + e^{(a_i(\theta_j - b_i))}}$$

Parameters can then be estimated using maximum likelihood or alternatively using a Bayesian estimation approach. To illustrate why heteroskedasticity exists, we will focus on maximum likelihood estimation. Lord (1980), under the assumption that the answer to each test item by each respondent is independent conditional on θ, showed that the maximum likelihood estimate of θ

has a variance of:

$$\sigma^2(\hat{\theta}|\theta) = \left(\sum_{i=1}^{n} (c_i a_i)^2 \frac{e^{(a_i(\theta_j - b_i))}}{(1+e^{(a_i(\theta_j - b_i))})^2} \right)^{-1}$$

where n is the number of items. As can be seen, the variance would be minimized with respect to θ if $\theta_j - b_i = 0$ for all items, and as $\theta_j - b_i$ approaches $\pm\infty$, the variance grows large.

Since test items are often targeted toward students near the proficient level, in the sense that $\theta_j - b_i$ is near 0 for these students, students in the lower and upper tail often have noisy estimates of their ability. The intuition is that the test is aimed at distinguishing between students near the proficiency cutoff, and so the test offers little information for students near the top or bottom of the distribution.

Plots of the estimated standard deviation of the measurement error (SEM) on the student's test score level are shown below in Figure 3.1. The SEMs are on the vertical axis and the student's test score are on the horizontal axis for grades 3 through 6 for mathematics. The plots are from the 2006 State X Technical Report on Test Characteristics. The measurement error variance is a function of the test score level. Students in the extreme ranges of the test score distribution have a measurement error variance that is substantially larger than in the center.

Also, it may be the case that some groups of students may be less likely to answer all questions on the exam. As described in State X technical reports, test scores are computed for all students who answer at least 6 questions in each of 2 sessions. Students who answer only a fraction of the total number of questions on the exam will tend to have less precisely estimated test scores.

A prediction of the theory presented above is that the error variance will be related to all variables that predict current achievement. This is because the variance of the measurement error is directly related to the current achievement of the student, so all variables that influence the current achievement level of the student should also be related to the measurement error variance. In the test of heteroskedasticity that follow, this is the pattern that emerges.

3.5.2 Other Possible Causes of Heteroskedastic Student Level Error

In addition to heteroskedasticity generated from measurement, it is possible that other sources of heteroskedasticity exist. Little literature exists on this topic, but there are many potential causes, and we can only speculate on what they may be. Some groups of students, such as those with low prior year achievement, may have more variation in unobserved factors such as motivation, classroom disruptions, neighborhood effects, family effects, or learning disabilities. In addition, students who perform poorly on tests may tend to leave many questions blank or guess at answers, and thus their scores from test to test may be more variable.

In the following sections, we test for heteroskedasticity empirically, and look for possible differences in the error variance among groups. This serves to demonstrate that the theoretical worries are justified and can motivate some predictions about how the precision of teacher effect estimates may depend on certain characteristics of the their students.

3.6 Testing for Heteroskedasticity

Under homoskedasticity:

$$E(v_{ig}^2|Z_{ig}) = \sigma_v^2$$

where Z_{ig} are the covariates in the regression model. We implemented a simple test of the homoskedasticity assumption examining whether squared residuals are related to student characteristics.

The first test simply grouped students into three groups: those with prior year test scores in the bottom 25%, the middle 50%, and the top 25%. We then calculated the average squared residuals for each group of students. We used the residuals from the DOLS regressions, which made use of teacher indicators. Results are included in Table 3.2. One thing to note is that the average squared residuals for the group of students in the bottom 25% in terms of prior year achievement are much larger than those for the group of students in the top 25%. The average squared residuals are around

45% larger for the bottom 25% compared with the top 25% for 4th grade and more than twice as large for 6th grade, even though under homoskedasticity, we would expect them to be similar. This is suggestive that more unexplained variation exists for the group of students in the bottom 25% of the prior year achievement score.

Next we regressed the squared residuals on the covariates as well as on their squares and cubes. Results for grades 4 and 6 are reported in Table 3.3. We found that several of the variables including the lagged test scores, as well as the indicators for the student being African-American, free and reduced priced lunch, and limited English proficiency status were statistically significant predictors at the 10% level.

Since the precision and stability of a teacher's value-added measure depends in part on how much unexplained variation there is in the student's test scores, as will be explained below, this suggests that teachers paired with large numbers of disadvantaged or low achieving students may have less precise teacher value-added estimates. In the following sections, we will present evidence of this. Specifically, we will show that teachers of these types of students tend to have less stable teacher effect estimates over time.

In addition to the regressions presented in Table 3.3, we performed the traditional Breusch-Pagan test, using fitted values, for heteroskedasticity separately for grade 4 and 6 and using the DOLS estimators. The test easily rejects the null hypothesis that the error is homoskedastic, with p-values for all grades and estimators less than .0001.

3.7 Evidence of Differences in Classroom Compositions

For there to be differences in the stability or the precision of teacher effect estimates due to student level heteroskedastic error, it is necessary for variation in classroom compositions to exist. For particular districts or states with little variation in classroom composition, it is unlikely that there will be large differences in the stability and precision of estimates due to heteroskedasticity. Also, there are some variables, such as gender, in which there may be a relationship with the error variance, but don't impact the precision and stability of teacher effect estimates, since there is little

variation across classrooms with respect to the variables.

To show that there is variation in classroom composition with respect to certain variables across the state, we included a set of summary statistics in the middle panels of Table 3.1 on classroom characteristics, which show that classrooms vary in their characteristics along a number of dimensions. The average past year math score of students in a class ranges from a score of 686.75 to 2066.737 for grade 4 and 866 to 2097 for grade 6. The interval between classrooms 2 standard deviations above the mean and 2 standard deviations below the mean is [1128.797,1697.353] for grade 4 and [1383.791,1911.623] for grade 6. Additionally, the proportion free and reduced priced lunch, limited English proficiency status, Hispanic, and African-American variables all range from 0 to 1.

3.8 Effects of Heteroskedastic Student Level Error on Precision of Teacher Value-Added Estimates

3.8.1 Simple Model of Heteroskedasticity

This model is designed to show, in the simplest case, how heteroskedasticity in the student level error can produce heteroskedasticity in teacher effect estimates. In the model there are two types of students and two teachers that students can be assigned to. The student types differ in the size of the student's error variance.

The achievement equation model is:

$$A_i = T_{0i}\beta_0 + T_{1i}\beta_1 + \varepsilon_i$$

where A_i is the achievement level of student i, T_0 and T_1 are teacher assignment indicator variables for the two teachers, teacher 0 and teacher 1, β_0 and β_1 are teacher effects for teacher 0 and teacher 1, and ε_i is an error term assumed to be independent of teacher assignment.

Let the variable S_i indicate which of the two student types the student belongs to and $v_0 < v_1$.

$$\begin{aligned} Var(\varepsilon_i) &= v_0 \qquad if \quad S_i = 0 \\ Var(\varepsilon_i) &= v_1 \qquad if \quad S_i = 1 \end{aligned}$$

$$v_0 < v_1$$

In this simple case, an OLS estimate of the teacher effect for teacher k produces:

$$\begin{aligned} \hat{\beta}_k - \beta_k &= \left(\sum_{i=1}^{N} T_{ki}^2 \right)^{-1} \left(\sum_{i=1}^{N} T_{ki}\varepsilon_i \right) \\ &= \frac{\sum_{i=1}^{N} T_{ki}\varepsilon_i}{N_k} \\ &= \bar{\varepsilon}_k \end{aligned}$$

where $\bar{\varepsilon}_k$ is the average error for the students that teacher k receives and N_k is the number of student observations for teacher k.

Let's suppose that each teacher has some students from S=0 and some from S=1. And also that teacher 0 tends to get more students from group 0, and teacher 1 tends to get more students from group 1.

We can use the Central Limit Theorem for inference. According to Greene (2008) (pg 1051, Lindeberg-Feller Central Limit Theorem with Unequal Variances) a central limit theorem result is possible as long as the random variables are independent with finite means and finite positive variances. Also, the average variance, $\frac{1}{N_k}(\sum_{i=1}^{N_k} \sigma_{\varepsilon_{ik}}^2)$, where N_k is the number of students for teacher k, must not be dominated by any single term in the sum and this average variance must converge to a finite constant, $\bar{\sigma}_{\varepsilon_k}^2$ as the number of students per teacher goes to infinity.

$$\bar{\sigma}_{\varepsilon_k}^2 = \lim_{N_k \to \infty} \frac{1}{N_k} (\sum_{i=1}^{N_k} \sigma_{\varepsilon_{ik}}^2)$$

Assume that all of those conditions hold. In that case,

$$\sqrt{N_k}(\hat{\beta}_k - \beta_k) \xrightarrow{d} Normal(0, \bar{\sigma}_{\varepsilon_k}^2)$$

and

$$Avar(\hat{\beta}_k) \approx \frac{\bar{\sigma}^2_{\varepsilon_k}}{N_k}$$

In this simple example the average variance, $\bar{\sigma}^2_{\varepsilon_k}$, for teacher 1 will tend to be larger than teacher 0, since they have more students from S=1. Therefore the asymptotic variance of the teacher effect estimate for teacher 1 will tend to be larger.

3.8.2 Including other Covariates in Achievement Model

Adding in covariates along with the teacher indicator variables complicates the result. In this case the achievement model is:

$$A_i = T_{0i}\beta_0 + T_{1i}\beta_1 + X_i\gamma + \varepsilon_i$$

where X_i is a vector of covariates.

A well known result (see Wooldridge (2010)), is that the OLS estimate of the teacher fixed effect for teacher k is:

$$\begin{aligned} \hat{\beta}_k - \beta_k &= \bar{A}_k - \bar{X}_k\hat{\gamma}_{FE} - \beta_k \\ &= \bar{\varepsilon}_k - \bar{X}_k(\hat{\gamma}_{FE} - \gamma) \end{aligned}$$

where $\bar{A}_k$ and $\bar{X}_k$ are the class averages of achievement and the covariates, and $\hat{\gamma}_{FE}$ is the fixed effects estimator of γ. It's straight forward to show that

$$Avar(\hat{\beta}_k) \approx \frac{\bar{\sigma}^2_{\varepsilon_k}}{N_k} + \bar{X}_k Avar(\hat{\gamma}_{FE})\bar{X}'_k$$

$\frac{\bar{\sigma}^2_{\varepsilon_k}}{N_k}$ will tend to be larger for teacher 1 than teacher 0. However, because of the additional terms in the $Avar(\hat{\beta}_k)$, it is not theoretically clear which teacher will have the less precise teacher

effect estimate when the relationships between the covariates and the student types are unknown. Ultimately, whether teacher effect estimates are less precise for some teachers is an empirical question. The important point is that it is possible for some teachers to have less precise estimates due to student characteristics, so it is worthwhile to check whether that is the case.

3.9 Inter-year Stability of Teacher Effect Estimates by Class Characteristics

Imprecision of teacher effect estimates has some important implications, especially for policies that use teacher value-added estimates to make inferences about teacher quality.

The precision of a teacher effect estimate will affect how well that estimate can predict the true teacher effect. If the estimated teacher effect is quite noisy, then the estimate will tend to poorly predict the true teacher effect. This section explains how examining the year to year stability of value-added estimates can reveal important information about the measures for those intending to use them for high stakes policies. The year to year stability is calculated by regressing the value-added measure in year t on a value-added measure in a previous year. We calculate separate stability coefficients for teachers with classrooms in the bottom 25%, middle 50%, and top 25% in terms of their students incoming average achievement. Those wishing to skip the technical details may move on to the next section.

Following McCaffrey et al. (2009), we can model a teacher effect estimate for teacher j in year t as:

$$\hat{\beta}_{jt} = \beta_j + \theta_{jt} + v_{jt}$$

where $\hat{\beta}_{jt}$ is the teacher effect estimate, β_j is the persistent component of the teacher effect, θ_{jt} is a transitory teacher effect that may have to do with a special relationship a teacher has with a class or some temporary change in a teacher's ability to teach, and v_{jt} is an error term due to sampling variation. The variance of v_{jt} will be related to the number of student observations used to estimate a teacher effect and the error variance associated with the students in the particular

teacher's class.

An important coefficient for predicting the persistent component of the teacher effect using an estimated teacher effect, which is essentially what a policy to deny tenure to teachers based on value added scores would be doing, is the stability coefficient, as termed by McCaffrey et al. (2009). The stability coefficient for teacher j is:

$$S_j = \frac{\sigma^2_\beta}{\sigma^2_\beta + \sigma^2_{\theta_{jt}} + \sigma^2_{v_{jt}}}$$

Note that the stability depends on the variance of the error term v_{jt}.

Assuming that the expectation of β_j conditional on $\hat{\beta}_{jt}$ is linear and that β_j, θ_{jt}, and v_{jt} are uncorrelated, then:

$$E(\beta_j|\hat{\beta_{jt}}) = \alpha + \frac{Cov(\hat{\beta_{jt}}, \beta_j)}{Var(\hat{\beta_{jt}})}\hat{\beta}_{jt} = \alpha + \frac{\sigma^2_\beta}{\sigma^2_\beta + \sigma^2_{\theta_{jt}} + \sigma^2_{v_{jt}}}\hat{\beta}_{jt} = \alpha + S_j\hat{\beta}_{jt}$$

and then also assuming that θ_{jt} and v_{jt} are mean zero, we get:

$$E(\beta_j|\hat{\beta_{jt}}) = (1 - S_j)\mu_{\beta_j} + S_j\hat{\beta_{jt}}$$

where μ_{β_j} is the mean of β_j.[7] So the weight that $\hat{\beta_{jt}}$ receives in predicting β_j is related to the stability coefficient. If the stability coefficient is small, then the estimated teacher effect receives little weight in the conditional expectation function and is of little use in predicting β_j.

[7] If the conditional expectation function isn't linear, then the algebra shown works for the linear projection, which is the minimum mean squared error predictor among linear functions of the estimated teacher effect.

The assumption that β_j, θ_{jt}, and v_{jt} are uncorrelated essentially implies that the teacher effect estimates are unbiased. There is some empirical support for this assumption at least for the DOLS and EB Lag estimators. Kane and Staiger (2008), Kane et al. (2013), and Chetty et al. (2011) both find that similar value-added estimators are relatively unbiased. If the estimates are biased, then we are effectively evaluating the stability of reduced form coefficients and not the causal effects of teachers on achievement. The estimators evaluated are commonly used in practice and conceivably will be used as the basis for high stakes policies, so it still may be of interest to know how they vary from year-to-year.

The stability coefficient can be estimated by an OLS regression of current year teacher value-added estimates on past year estimates of teacher value-added and a constant. This does impose the additional assumption that the variances of θ_{jt} and v_{jt} are constant over time and that the transitory teacher effect and error terms are uncorrelated over time. In that case the OLS estimates are estimating the population parameter:

$$\frac{Cov(\hat{\beta}_{jt-1}, \hat{\beta}_{jt})}{Var(\hat{\beta}_{jt-1})} = \frac{\sigma^2_{\beta_j}}{\sigma^2_{\beta_j} + \sigma^2_{\theta_{jt-1}} + \sigma^2_{v_{jt-1}}} = S_j$$

Since the variance of the teacher effect estimates tends to be constant over time, the regression coefficient is nearly identical to the inter-year correlation coefficient.

The stability coefficient will be estimated for different subgroups of teachers based on the characteristics of the students a teacher receives. Specifically, the stability will be computed for teachers that received classes in the bottom 25%, middle 50% and top 25% of classroom average prior test score in both years t and $t-1$. If the variance of v_{jt} differs across subgroups of teachers, then the stability and the degree to which the estimate predicts the true teacher effect will also differ.

Another ratio may be of interest. Following McCaffrey et al. (2009) once again, the reliability of a teacher effect estimate, denoted as R_{jt}, is:

$$R_{jt} = \frac{\sigma^2_{\beta} + \sigma^2_{\theta_{jt}}}{\sigma^2_{\beta} + \sigma^2_{\theta_{jt}} + \sigma^2_{v_{jt}}}$$

It may be of interest to know how much a teacher affected student learning in a given year. This may be the case in a merit pay system, for instance. In this case, we would be interested in the expected value of $\beta_j + \theta_{jt}$ conditional on the estimated teacher effect in year t. Using similar assumptions as before:

$$E(\beta_j + \theta_{jt} | \hat{\beta_{jt}}) = (1 - R_{jt})\mu_{\beta} + R_{jt}\hat{\beta_{jt}}$$

Under an additional assumption that variance of β_j and θ_{jt} do not vary across subgroups, then the stability of teacher value added estimates will be proportional to the reliability. This is simply because:

$$R_{jt} = \frac{\sigma_\beta^2 + \sigma_{\theta_{jt}}^2}{\sigma_\beta^2} S_j$$

3.9.1 Brief Overview of the Analysis

Given that there may be differences in the degree of tracking or sorting in elementary and middle schools, the analysis is done separately by grade. Additionally, since it may be that teachers of certain types of classrooms are less experienced, and this may affect the year-to-year stability of the teacher's value-added estimate, the teacher's level of experience is controlled for in the regressions by creating separate dummy variable for each possible year of experience and including each of those variables in the regressions.

The estimates for the different subgroups were computed by an OLS regression of the current year value-added estimate on the lagged teacher value-added estimate interacted with a subgroup indicator variable, a subgroup specific intercept, and an indicator for the teacher's level of experience. The regression equation is:

$$\begin{aligned}\hat{\beta}_{jt} &= \sum_{g=1}^{3} \alpha_g 1\{subgroup_{jt} = g\} + \sum_{g=1}^{3} \gamma_g \hat{\beta}_{jt-1} 1\{subgroup_{jt} = g\} \\ &\quad + \sum_{\tau=1}^{M} \zeta_\tau 1\{exper_{jt} = \tau\} + \phi_{jt}\end{aligned}$$

where $\hat{\beta}_{jt}$ is teacher j's value added estimate in year t, $subgroup_{jt}$ is a variable indicating the teacher's subgroup, and $exper_{jt}$ is the teacher's experience level. The γ_g parameters are the parameters of interest in the analysis. One way to think about them is as a group specific autoregressive coefficient for a teacher's value-added score, and they are quite similar to group specific year-to-year correlations in value-added.

The advantage of the regression based approach over calculating year-to-year correlations is that it is much easier to calculate test statistics using conventional regression software. In the following sections, we will test whether the year-to-year stability of teacher value-added estimates for different subgroups are statistically different from one another.

The analysis is also repeated for each grade with the number of student observations artificially set to be equal. Since the precision of estimates for a teacher depends on both the number of student observations and the degree of variation in the student level error, it is of interest to identify the separate effects of these two sources of variability in teacher effect estimates. In order to make the number of student observations equal for all teachers, first all teachers with less than 12 student observations were dropped. Then for those teachers with more than 12 student observations, students are randomly dropped from the classroom until the the number of student observations is 12 for all teachers.[8] To give an example, suppose a teacher has 20 students in a class, then 8 of the students are randomly dropped, so that the teacher's value-added estimate is based on the scores of only 12 students.

First, results will be reported in which all teacher effects are estimated using only one year of data. Then, the analysis will be reported using two years of data for each teacher. When two years of data are used to compute value-added the groupings into bottom 25%, middle 50%, and top 25% are based on the two year average of prior year test score within the teacher's classrooms. This then averages over the same sample of students used to compute the two year value-added measures.

In the case of the estimates based on two years of data, the teacher effect estimate for year t will be estimated using years t and $t-1$. The stabilities are computed by regressing the value-added estimate for year t on year $t-2$. This is done so that the years in which teacher effects are estimated do not overlap, which will avoid sampling variation or class level shocks affecting both estimates.

[8] We have also done the analysis where the number of observations is set to 15 and 20, and the general patterns reported are the same.

3.10 Results on the Stability of Teacher Effect Estimates by Subgroup

The inter-year stabilities for subgroups of teachers based on the average past year score of the students in the class are reported below.[9] We perform separate tests for whether the estimates for the middle 50% and top 25% statistically differ from the bottom 25%. Also, a joint test that the estimates for the middle 50% and top 25% are both statistically different from the bottom 25% is reported.

Although there is variation in what is statistically significant across grades and estimators, a few patterns do emerge. The stability ratio tends to be highest for teachers facing classrooms in the middle 50% and top 25% in average lagged score compared to teachers in the bottom 25%. The stability ratio is typically 25 to over 50% larger for teachers with classrooms in the middle 50 and top 25%. This pattern is true even after the number of student observations is fixed at 12 and in some cases when 2 years of data are used to compute value-added.

3.10.1 DOLS Stabilities

Table 3.4 shows the results for the DOLS estimator. Results for 4th grade and 6th grade are shown separately. The left panels show the DOLS teacher value-added estimates when the data is based on only one year of data. The right panel are based on estimates with two years of data. Within each panel, results labeled "Unrestricted Obs" are based on teacher value-added estimates that use all the available student observations in a year. Results labeled "12 Student Obs" are based on only 12 randomly chosen student observations in each year. For the two year results, the results reported under the "12 Student Obs" column are based on 12*2=24 student observations. Standard errors are clustered at the school level.[10] A "+" symbol indicates that the middle 50% (or top 25%

[9] We have also examined whether the inter-year stability differs when classrooms are grouped according to proportion free-and-reduced price lunch, proportion Hispanic, and proportion African-American. We found that teachers in classrooms with high proportions of minority and low-income students also have lower inter-year stabilities. Results are available upon request.

[10] We have also tried clustering at the teacher level, but the school level standard errors were more conservative, so we chose to report those.

as the case may be) coefficient is statistically different from the bottom 25% at the 5% level.

3.10.1.1 4th Grade Results

In 4th grade, the stability for teachers with classes in the bottom 25% of prior year achievement is .359, and the stabilities for the middle 50% and top 25% are .483 and .555 respectively when the number of student observations is unrestricted. The coefficients for the middle 50% and top 25% statistically differ from the coefficient for the bottom 25% at the 5% level. The patterns are quite similar once the number of student observations is fixed at 12, although predictably the estimates are somewhat smaller, since in the unrestricted case each teacher's value-added estimate is based on at least 12 observations. The stability for the bottom 25% is .308 while the stabilities for middle and top are .392 and .471 respectively and are statistically different from the bottom. Additionally, in both the unrestricted and restricted to 12 observations cases, the joint test that both the middle 50% and top 25% coefficients differ from the bottom rejects comfortably at the 5% level.

For the cases in which two years of data are used, the stability is calculated using four years of data. The teacher effect estimate in year t, which uses data from year t and $t-1$, is regressed on the teacher effect estimate from year $t-2$, which uses years $t-2$ and $t-3$. For a teacher to be included in one of the quartile groupings, the teacher had to have a two-year average prior year achievement score in that quartile range for years t and $t-2$. This dramatically reduced the sample of teachers available to compare.

When two years of data are used to estimate teacher value-added in 4th grade the stability for teachers with classes in the bottom 25% increase to .551 and to .646 and .730 for the middle and top, respectively, in the unrestricted observations case. The difference between the coefficients for the top and bottom is statistically significant at the 5% level. The point estimate for the middle 50% is larger than the bottom 25%, but the difference between the two is not statistically significant at the 5% level. The joint test that top or the middle coefficient differs from the bottom is significant at the 5% level. When the number of student observations per year is fixed at 12, the point estimates in the case of the middle and top are larger than the bottom, and both are statistically different from

the bottom. The joint test that either the middle or top is different from the bottom also rejects.

3.10.1.2 6th Grade Results

The results for 6th grade are broadly similar to 4th grade using one year of data. With one year of data and unrestricted observations the stabilities tend to be higher than in 4th grade. This is likely due to 6th grade teachers having more student observations per year. In this case, the stabilities are .534, .619, and .665 for the bottom, middle, and top respectively. The tests for whether the top stabilities are different from the bottom rejects, while the test for the middle 50% does not. The joint test also rejects. When 12 student observations are used, the stabilities are .356, .401, and .479, respectively, for the bottom, middle, and top. Once again the test that the top and bottom differ and the joint test rejects, while the test that the middle differs from the bottom does not.

In the case of two years of data, none of the estimates statistically differ from one another in either the case of unrestricted observations or the case restricted to 12 student observations.

3.10.2 EB Lag Stabilities

The results for the empirical Bayes estimates can be found in Table 3.5 and are quite similar to those for the DOLS estimates. One difference between the empirical Bayes and DOLS specifications is that the regressions corresponding to the empirical Bayes estimates include classroom aggregates of the individual covariates, since this is often one of the justifications for using this approach over DOLS.[11]

In the case of one year of data and 4th grade, the stability estimates are .361, .483, and .551 for the bottom 25%, middle 50%, and top 25%, respectively, in the unrestricted observations case. In the case where the number of student observations is set to 12, the stability estimates are .309, .391, and .461 respectively. In both cases, the middle 50% and top 25% estimates are statistically significantly different from the bottom 25%. The estimates are very similar to the DOLS case.

[11] We have also included class aggregates in the DOLS regressions, and the results do not change much. Estimates of the class level aggregates were identified for DOLS using the two step approach described previously.

In the two year case in 4th grade, the pattern is again fairly similar to the DOLS results. When the number of observations is unrestricted, only the top and bottom 25% stabilities are statistically from one another. The p-value of the joint test is .0511, however. When the number of observations is restricted to 12, the estimates are .476, .584, and .657, respectively. The difference between the top 25% and bottom 25% coefficients is statistically at the 5% level. The joint test rejects at the 5% level as well.

In 6th grade with one year of data, the only statistically significant difference at the 5% level is between the top 25% and bottom 25% in the unrestricted case, with point estimates of .650 for the top and .548 for the bottom. In the case of 2 years of data, no statistically significant differences are detected.

3.11 Sensitivity Checks

We performed a number of sensitivity checks. All of them support the conclusion that differences exist in the inter-year stabilities across sub-groups.

We performed the analysis using English language arts scores and found similar patterns as mathematics. The teachers assigned to students in the bottom 25% tended to have less stable value-added scores from year to year. One thing interesting to note is that English language arts value-added scores tended to be less stable from year-to-year overall compared to mathematics. This finding is consistent with the findings reported in the MET project reports.

Since it conceivable that teachers of students with low average prior achievement scores are inexperienced and inexperienced teachers also have lower inter-year stabilities, the analysis was repeated dropping all teachers with less than 5 years of experience. However, the teacher's experience was controlled for in the regression of the teacher's current value-added score on their prior value-added score specifically to account for this issue, and the patterns described above were very similar to those seen in this sensitivity check as expected.

As an additional sensitivity check, we repeated the analysis with school dummies. We were still able to detect statistically significant differences in inter-year stabilities across sub-groups.

We tried estimating the empirical Bayes estimates using an alternate estimator. In the alternate estimator, we estimated the model parameters using a mixed effects estimator that treated the teacher effects as random. These results were very similar to the empirical Bayes approach outlined above that was based on the approach taken in Kane and Staiger (2008).

Also, we used twice lagged reading and math scores as instruments for the once lagged reading and math scores to help account for measurement error in these variables as another sensitivity check. Again, statistically significant differences were found in the stabilities across sub-groups.

Finally, we performed the analysis separately for the six largest school districts in the state. The general patterns held. In a majority of the cases, the stability coefficient was estimated to be the smallest in the case of the bottom 25%. In no case was the stability coefficient of the middle 50% or top 25% statistically significantly smaller than the bottom 25%. In some districts, the teachers with classrooms in the middle 50% had the largest year-to-year stability, while in others the top 25% had the largest year-to-year stability. In one case the year to year stability of the bottom 25% was the largest, but it wasn't statistically significantly so. The estimates were quite noisy when the sample was separated in this way, so it is not clear whether this reflected real differences across districts or not. It seems possible that in different context the group of teachers that has the largest year-to-year stability could differ. However, our main takeaway is that some groups of teachers have less stable value added estimates from year-to-year.

Tables for all of these sensitivity checks are available upon request.

3.12 High Stakes Policy Simulation

There is an increasing push to use value-added estimates for high stakes decisions such as tenure or merit pay bonuses. Since the precision and stability of a teacher's value-added estimate is related to the makeup of the teacher's class, it may be the case that the teachers serving certain groups of students may be more likely receive a sanction or bonus.

In order to examine this, we produced a simulation in which high stakes decisions are made based upon value-added scores, and teachers differ in the stability of their value-added estimates.

We base the stability level of the measure of value-added on the results we found in the previous sections. Each teacher is ranked and flagged if they are in the bottom or top 10% according to their teacher value-added score. We then calculate the proportion of teachers associated with each stability level that are labeled as either in the bottom or top 10%.

The simulation consists of 300 teachers and 3 stability levels. 100 teachers are assigned to each stability level. The true teacher effects are normally distributed and have a mean of 0 and a variance of 1. The "estimated" teacher effects have estimation error added that is normally distributed with mean 0, and the variance depends on the stability level of the teacher.

Two sets of stability levels were chosen. The first corresponds to the DOLS estimates in 4th grade with 12 student observations and one year of data, with stabilities of .308, .392, and .471. The second corresponds to the DOLS estimates in 4th grade with 12 student observations and 2 years of data, with stability levels of .465, .578, and .660.

We calculate the average proportion of teachers associated with each stability level over the 5000 reps. Results are included in Table 3.6. The results from the simulation using the DOLS estimates in 4th grade with 12 student observations and one year of data can be found in the upper panel. For teachers associated with the stability of .308, which was the stability associated with teachers of classrooms in the bottom 25% in the analysis above, the proportion found in the bottom or top 10% was .249. When the stability level was .392 the proportion dropped to .195, and when the stability went to .471, the proportion fell to .156. This last drop was nearly a 10 percentage point change from the lowest stability. The results using two years of data show a similar pattern and can be found in the bottom panel. Teachers associated with the lowest stability have a proportion of .243. Teachers associated with stabilities of .578 and .660, which were associated with students in the middle 50% and top 25%, respectively, were found in the bottom or top 10% of the estimated teacher quality distribution at a proportion of .193 and .164 respectively. This represents an almost 8 percentage point drop for the latter.

The simulation results indicate that the differences in stability levels found in this analysis can have a large impact on the likelihood that a teacher finds his or herself in the top or bottom of the

estimated teacher quality distribution.

3.13 Conclusion

This paper provides evidence that the variability and stability of teacher effect estimates depends on the characteristics of a teacher's class. Policies to deny tenure to teachers and policies designed to reward teacher performance in a given year, which are based on teacher value-added estimates, may differentially impact teachers with certain types of students.

The relationship between the stability of estimates and the classroom characteristics of students extends beyond the number of student observations. There is a strong theoretical reason for suspecting that a student's error term is heteroskedastic and statistical tests bear this out. As a consequence of this and student tracking and sorting into schools, teachers will serve different groups of students and have differences in the precision of their teacher effect estimates as a result. The differences in the stability ratios are large in magnitude and statistically significant even after fixing the number of student observations to a constant.

Also, some evidence is presented that the relationships remain even as more observations are added. When two years of data are used, there still exist statistically significant and large differences for different subgroups of teachers.

The heteroskedasticity is likely due in part to heteroskedastic measurement error variance. Assuming the item response model is correct, heteroskedastic measurement error is a direct result of the maximum likelihood estimation procedure which produces estimates of the achievement level of each student. The patterns that teachers of students with lagged achievement scores in the middle of the achievement distribution tend to have the highest inter-year stabilities is consistent with heteroskedasticity caused by the measurement error, although teachers with students in the top 25% also tend to have more stable estimates. One reason the top and bottom may be different is that there may be greater potential for guessing or item non-response for students at the bottom of the distribution. It may be possible to reduce the heteroskedasticity by improving measurement. Future work will hopefully explore how much of the heteroskedasticity is attributable to measurement.

Heteroskedastic student level error also has other implications for researchers and policymakers. Empirical Bayes estimators are commonly computed assuming homoskedastic student level error. This assumption does not seem to be true, and since there are large differences in stability ratios that appear to be driven by heteroskedasticity, the violation of this assumption may impact the teacher rankings that are created using the empirical Bayes estimators. Allowing heteroskedasticity in the student level error should be done if possible.

Additionally, it is quite common for standard errors and the corresponding confidence intervals to be based on a homoskedasticity assumptions.[12] It is important that the confidence intervals accurately reflect imprecision caused by all sources of variability, not just the number of student observations, so standard errors should at least be made heteroskedasticity robust. This is particularly important since the teacher value-added estimates are being made publicly available in some school districts.

It is important to understand the limitations of any measure of performance. The analysis presented here does suggest that for all subgroups value-added measures do have positive inter-year stabilities, so information can be gathered for all subgroups of teachers. However, teachers of certain groups of students will tend to have less precise and less stable teacher value-added estimates. As a result of this, it is the opinion of the authors that care should be used in evaluating teachers using value-added estimators and value-added estimates should not be used as the sole basis of any high stakes policy involving teachers.

[12] Ballou et al. (2004) assume homoskedasticity in computing standard errors, as does the value-added estimator employed by the NYC school district

APPENDIX

APPENDIX

TABLES AND FIGURES

Figure 3.1: Standard Error of Measure Plots for Mathematics Grades 3- 6

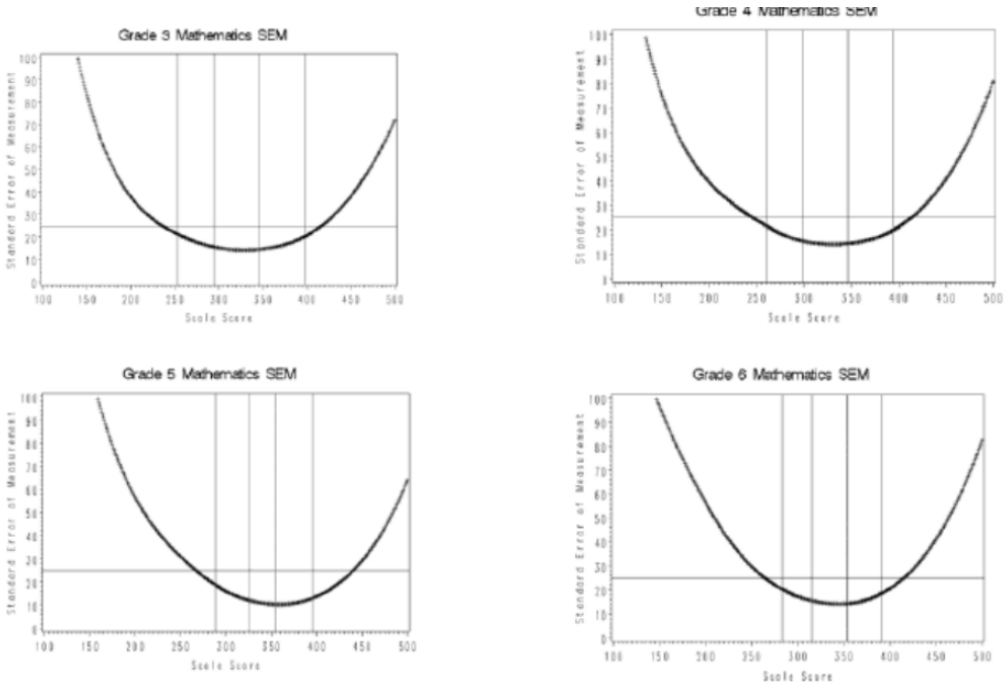


Table 3.1: Summary statistics

4th Grade				
Variable	**Mean**	**Std. Dev.**	**Min.**	**Max.**
Math Scale Score	1543.377	240.699	581	2330
Reading Scale Score	1591.033	291.045	295	2638
Math Standardized Scale Score	0.103	0.947	-3.957	3.409
Reading Standardized Scale Score	0.105	0.928	-4.578	3.753
Black	0.208	0.406	0	1
Hispanic	0.224	0.417	0	1
Free and Reduced Price Lunch	0.486	0.5	0	1
Limited English Proficiency	0.173	0.378	0	1
Avg. Lag Math Score	1413.075	142.139	686.75	2066.737
Prop. FRL	0.496	0.28	0	1
Prop. LEP	0.17	0.213	0	1
Prop. Hispanic	0.218	0.245	0	1
Prop. Black	0.216	0.248	0	1
Students/Teacher	49.008	38.534	12	412
Teacher Years of Experience	8.902	8.887	0	47
# of Teachers	14,820			
# of Schools	1,768			
N		726,299		

6th Grade				
Variable	**Mean**	**Std. Dev.**	**Min.**	**Max.**
Math Scale Score	1701.841	232.71	569	2492
Reading Scale Score	1704.809	294.454	539	2758
Math Standardized Scale Score	0.092	0.913	-4.163	3.354
Reading Standardized Scale Score	0.071	0.928	-4.049	3.526
Black	0.224	0.417	0	1
Hispanic	0.223	0.416	0	1
Free and Reduced Price Lunch	0.476	0.499	0	1
Limited English Proficiency	0.174	0.379	0	1
Avg. Lag Math Score	1647.707	131.958	866	2097
Prop. FRL	0.496	0.259	0	1
Prop. LEP	0.172	0.205	0	1
Prop. Hispanic	0.214	0.234	0	1
Prop. Black	0.24	0.245	0	1
Students/Teacher	145.378	165.685	12	1036
Teacher Years of Experience	9.571	9.362	0	40
# of Teachers	5,323			
# of Schools	796			
N		773,849		

Table 3.2: Average Squared Residuals for DOLS based on Subgroups of Prior Year Class Average Achievement

Grade	Overall	Bottom 25%	Middle 50%	Top 25%
4th Grade	18644.722	28091.514	13665.164	19352.092
N	709302	174780	356821	177701
6th Grade	16395.069	29825.119	11574.907	12670.438
N	723292	179894	357843	185555

All regressions include lagged math and ELA test scores, indicators for Black, Hispanic, free and reduced price lunch, limited English proficiency, female, and year dummies.

Table 3.3: Tests for Heteroskedasticity

VARIABLES	Grade 4 DOLS Squared Residuals	Grade 6 DOLS Squared Residuals
Math Lag Score	-91.60***	-176.5***
	(5.357)	(15.28)
Math Lag Score Squared	0.00705*	0.00709
	(0.00396)	(0.00939)
Math Lag Score Cubed	1.05e-05***	1.57e-05***
	(9.45e-07)	(1.90e-06)
Reading Lag Score	-45.76***	-55.68***
	(2.772)	(5.663)
Reading Lag Score Squared	0.0161***	0.0173***
	(0.00195)	(0.00341)
Reading Lag Score Cubed	0.79e-07	-8.73e-07
	(4.43e-07)	(6.65e-07)
Black	293.8*	473.6**
	(177.3)	(205.2)
Hispanic	-265.9*	-272.0*
	(154.7)	(159.8)
FRL	540.6***	1,104***
	(114.9)	(120.3)
LEP	1,249***	711.8***
	(190.7)	(183.3)
Female	-1,436***	-2,609***
	(97.34)	(114.0)
Constant	134,184***	262,361***
	(2,472)	(8,485)
Observations	709,302	723,292
R^2	0.050	0.079
Joint Test	886.6	862.4
p-value	0	0

All regressions include lagged math and ELA test scores, indicators for Black, Hispanic, free and reduced price lunch, limited English proficiency, female, and year dummies. Standard errors clustered at school level in parentheses. Joint Test refers to F test statistic that all coefficients equal to 0.

*** p<0.01, ** p<0.05, * p<0.1

Table 3.4: Estimates of Year to Year Stability for DOLS by Subgroups of Class Achievement

DOLS 4th grade

	1 Year of Data		2 Years of Data	
	Unrestricted Obs	12 Student Obs	Unrestricted Obs	12 Student Obs
Bottom 25%	0.359***	0.308***	0.551***	0.465***
	(0.0277)	(0.0266)	(0.0437)	(0.0449)
Middle 50%	0.483***+	0.392***+	0.646***	0.578***+
	(0.0181)	(0.0180)	(0.0325)	(0.0315)
Top 25%	0.555***+	0.471***+	0.730***+	0.660***+
	(0.0255)	(0.0246)	(0.0495)	(0.0485)
Observations	8,124	7,650	2,735	2,527
R^2	0.227	0.165	0.357	0.298
Joint Test	14.70	10.14	3.677	4.436
p-value	4.81e-07	4.27e-05	0.0257	0.0121

DOLS 6th grade

	1 Year of Data		2 Years of Data	
	Unrestricted Obs	12 Student Obs	Unrestricted Obs	12 Student Obs
Bottom 25%	0.534***	0.356***	0.812***	0.574***
	(0.0452)	(0.0476)	(0.0588)	(0.0756)
Middle 50%	0.619***	0.401***	0.717***	0.560***
	(0.0209)	(0.0247)	(0.0447)	(0.0485)
Top 25%	0.665***+	0.479***+	0.711***	0.575***
	(0.0263)	(0.0310)	(0.0403)	(0.0508)
Observations	4,290	3,772	1,506	1,359
R^2	0.481	0.288	0.642	0.445
Joint Test	3.684	3.233	1.193	0.0274
p-value	0.0256	0.0401	0.304	0.973

All regressions include lagged math and ELA test scores, indicators for Black, Hispanic, free and reduced price lunch, limited English proficiency, female, and year dummies. Standard errors clustered at school level in parentheses

*** p<0.01, ** p<0.05, * p<0.1

+ Indicates value statistically different from Bottom 25% at 5% level

Joint Test: F-test statistic that Middle 50 % and Top 25 % coefficients different from Bottom 25%

Table 3.5: Estimates of Year to Year Stability for EB Lag by Subgroups of Class Achievement

EB Lag 4th grade

	1 Year of Data		2 Years of Data	
	Unrestricted Obs	12 Student Obs	Unrestricted Obs	12 Student Obs
Bottom 25%	0.361***	0.309***	0.571***	0.476***
	(0.0278)	(0.0269)	(0.0445)	(0.0459)
Middle 50%	0.483***+	0.391***+	0.659***	0.584***
	(0.0183)	(0.0180)	(0.0341)	(0.0318)
Top 25%	0.551***+	0.461***+	0.733***+	0.657***+
	(0.0254)	(0.0246)	(0.0497)	(0.0491)
Observations	8,124	7,650	2,735	2,527
R^2	0.220	0.157	0.352	0.291
Joint Test	13.80	8.813	2.985	3.697
p-value	1.16e-06	0.000158	0.0511	0.0252

EB Lag 6th grade

	1 Year of Data		2 Years of Data	
	Unrestricted Obs	12 Student Obs	Unrestricted Obs	12 Student Obs
Bottom 25%	0.548***	0.354***	0.814***	0.583***
	(0.0433)	(0.0482)	(0.0497)	(0.0702)
Middle 50%	0.614***	0.385***	0.717***	0.551***
	(0.0199)	(0.0247)	(0.0432)	(0.0481)
Top 25%	0.650***+	0.457***	0.714***	0.561***
	(0.0267)	(0.0318)	(0.0405)	(0.0529)
Observations	4,290	3,772	1,506	1,359
R^2	0.437	0.224	0.610	0.387
Joint Test	2.492	2.402	1.558	0.0715
p-value	0.0835	0.0913	0.212	0.931

All regressions include lagged math and ELA test scores, indicators for Black, Hispanic, free and reduced price lunch, limited English proficiency, female, class averages of all preceding variables, class size, a quadratic function of experience, and year dummies. Standard errors clustered at school level in parentheses.

*** p<0.01, ** p<0.05, * p<0.1

+ Indicates value statistically different from Bottom 25% at 5% level

Joint Test: F-test statistic that Middle 50 % and Top 25 % coefficients different from Bottom 25%

Table 3.6: High Stakes Policy Simulation

Simulation 1: DOLS Stability, 4th Grade, 12 Student Obs, 1 year of Data

Stability	Error Variance of VAM Estimate	Proportion Found in Bottom or Top 10%
.308	2.247	.249
.392	1.551	.195
.471	1.123	.156

Simulation 2: DOLS Stability, 4th Grade, 12 Student Obs, 2 years of Data

Stability	Error Variance of VAM Estimate	Proportion Found in Bottom or Top 10%
.465	1.151	.243
.578	.730	.193
.660	.515	.164

Simulations results are based on 5000 Monte Carlo repetitions. There are 100 teachers per type. True teacher effects are distributed Normal(0,1). Error in the value-added measures is normally distributed with mean 0 and a variance listed in the "Error Variance of VAM Estimate" column.

BIBLIOGRAPHY

BIBLIOGRAPHY

Aaronson, D., Barrow, L., and Sander, W. (2007). Teachers and student achievement in the chicago public high schools. *Journal of Labor Economics*, 25(1):95–135.

Abrevaya, J. and Dahl, C. M. (2008). The effects of birth inputs on birthweight: evidence from quantile estimation on panel data. *Journal of Business & Economic Statistics*, 26(4):379–397.

Angrist, J., Chernozhukov, V., and Fernandez-Val, I. (2006). Quantile regression under mispecification, with an application to the u.s. wage structure. *Econometrica*.

Arias, O., Hallock, K. F., and Sosa-Escudero, W. (2001). Individual heterogeneity in the returns to schooling: Instrumental variables quantile regression using twins data. *Empirical Economics*.

Ballou, D., Sanders, W., and Wright, P. (2004). Controlling for student background in value-added assessment of teachers. *Journal of Educational and Behavioral Statistics*, 29(1):37–65.

Bound, J., Brown, C., Duncan, G. J., and Rodgers, W. L. (1994). Evidence on the validity of cross-sectional and longitudinal labor market data. *Journal of Labor Economics*, pages 345–368.

Bound, J. and Krueger, A. B. (1991). The extent of measurement error in longitudinal earnings data: Do two wrongs make a right? *Journal of Labor Economics*.

Boyd, D., Grossman, P., Lankford, H., Loeb, S., and Wyckoff, J. (2008). Who leaves? teacher attrition and student achievement. Technical report, National Bureau of Economic Research.

Bricker, J. and Engelhardt, G. V. (2008). Measurement error in earnings data in the health and retirement study. *Journal of Economic and Social Measurement*, 33(1):39–61.

Buchinsky, M. (1994). Changes in the u.s. wage structure 1963-1987: Application of quantile regression. *Econometrica*.

Buchinsky, M. (1998). Recent advances in quantile regression models: A practical guideline for empirical research. *The Journal of Human Resources*.

Card, D. (1999). The causal effect of education on earnings. *Handbook of labor economics*, 3:1801–1863.

Center, V.-A. R. (2010). Nyc teacher data initiative: Technical report on the nyc value-added model 2010. Technical report.

Chen, X., Hong, H., and Tamer, E. (2005). Measurement error models with auxiliary data. *The Review of Economic Studies*.

Chernozhukov, V. and Umantsev, L. (2001). Conditional value-at-risk: Aspects of modeling and estimation. *Empirical Economics*, 26(1):271–292.

Chetty, R., Friedman, J. N., and Rockoff, J. E. (2011). The long-term impacts of teachers: Teacher value-added and student outcomes in adulthood. Technical report, National Bureau of Economic Research.

Condie, S., Lefgren, L., and Sims, D. (2014). Teacher heterogeneity, value-added and education policy. *Economics of Education Review*, 40(0):76 – 92.

Dee, T. S. (2004). Teachers, race, and student achievement in a randomized experiment. *Review of Economics and Statistics*, 86(1):195–210.

Eide, E. R. and Showalter, M. H. (1999). Factors affecting the transmission of earnings across generations: A quantile regression approach. *Journal of Human Resources*, pages 253–267.

Goldhaber, D. and Chaplin, D. (2012). Assessing the 'rothstein falsification test': Does it really show teacher value-added models are biased? *Center for Education Data & Research Working Paper*.

Goldhaber, D., Walch, J., and Gabele, B. (2013). Does the model matter? exploring the relationship between different student achievement-based teacher assessments. *Statistics and Public Policy*, 1(1):28–39.

Greene, W. H. (2008). *Econometric Analysis*. Pearson.

Guarino, C., Reckase, M. D., and Wooldridge, J. M. (2012). Can value-added measures of teacher performance be trusted? Technical report, Discussion Paper series, Forschungsinstitut zur Zukunft der Arbeit.

Guarino, C. M., Reckase, M. D., Stacy, B., and Wooldridge, J. M. (2014). Evaluating specification tests in the context of value-added estimation. Technical report, Michigan State Education Policy Center.

Haider, S. and Solon, G. (2000). Non random selection in the hrs social security earnings sample. *RAND, Labor and Population Program Working Paper Series*.

Haider, S. and Solon, G. (2006). Life-cycle variation in the association between current and lifetime earnings. *The American Economic Review*.

Hanushek, E. A. (1979). Conceptual and empirical issues in the estimation of educational production functions. *Journal of human Resources*, pages 351–388.

Harris, D., Sass, T., and Semykina, A. (2011). Value-added models and the measurement of teacher productivity.

Harvey, A. C. (1976). Estimating regression models with multiplicative heteroscedasticity. *Econometrica: Journal of the Econometric Society*, pages 461–465.

Hausman, J. (2001). Mismeasured variables in econometric analysis: Problems from the right and problems from the left. *The Journal of Economic Perspectives*.

Imbens, G. M. and Wooldridge, J. M. (2008). Recent developments in the econometrics of program evaluation. Technical report, National Bureau of Economic Research.

Imbens, G. W. (2000). The role of the propensity score in estimating dose-response functions. *Biometrika*, 87(3):706–710.

Imberman, S. A. and Lovenheim, M. F. (2013). Does the market value value-added? evidence from housing prices after a public release of school and teacher value-added. Technical report, National Bureau of Economic Research.

Jacob, B. A. and Lefgren, L. (2008). Can principals identify effective teachers? evidence on subjective performance evaluation in education. *Journal of Labor Economics*, 26(1):101–136.

Kane, T. J., McCaffrey, D. F., Miller, T., and Staiger, D. O. (2013). Have we identified effective teachers? validating measures of effective teaching using random assignment. research paper. met project. *Bill & Melinda Gates Foundation*.

Kane, T. J. and Staiger, D. O. (2002). The promise and pitfalls of using imprecise school accountability measures. *The Journal of Economic Perspectives*, 16(4):91–114.

Kane, T. J. and Staiger, D. O. (2008). Estimating teacher impacts on student achievement: An experimental evaluation. Technical report, National Bureau of Economic Research.

Kane, T. J. and Staiger, D. O. (2010). Learning about teaching: Initial findings from the measures of effective teaching project. *Bill & Melinda Gates Foundation*.

Kim, B. and Solon, G. (2005). Implications of mean-reverting measurement error for longitudinal studies of wages and employment. *Review of Economics and Statistics*, 87(1):193–196.

Koedel, C. and Betts, J. (2007). Re-examining the role of teacher quality in the educational production function. Technical report, Department of Economics, University of Missouri.

Koenker, R. (2005). *Quantile Regression*. Cambridge University Press.

Koenker, R. and Bassett, G. (1978). Regression quantiles. *Econometrica*.

Koenker, R. and Hallock, K. F. (2001). Quantile regression. *The Journal of Economic Perspectives*.

Lockwood, J. and McCaffrey, D. F. (2009). Exploring student-teacher interactions in longitudinal achievement data. *Education*, 4(4):439–467.

Loeb, S., Soland, J., and Fox, L. (2014). Is a good teacher a good teacher for all? comparing value-added of teachers with their english learners and non-english learners. *Educational Evaluation and Policy Analysis*.

Lord, F. M. (1980). *Applications of Item Response to Theory to Practical Testing Problems*. Lawrence Erlbaum.

Machado, J. A. and Mata, J. (2005). Counterfactual decomposition of changes in wage distributions using quantile regression. *Journal of applied Econometrics*, 20(4):445–465.

McCaffrey, D. F., Lockwood, J., Koretz, D., Louis, T. A., and Hamilton, L. (2004). Models for value-added modeling of teacher effects. *Journal of educational and behavioral statistics*, 29(1):67–101.

McCaffrey, D. F., Sass, T. R., Lockwood, J., and Mihaly, K. (2009). The intertemporal variability of teacher effect estimates. *Education*, 4(4):572–606.

Mincer, J. A. (1974). Schooling and earnings. In *Schooling, experience, and earnings*, pages 41–63. Columbia University Press.

Morris, C. N. (1983). Parametric empirical bayes inference: theory and applications. *Journal of the American Statistical Association*, 78(381):47–55.

Neal, D. and Schanzenbach, D. W. (2010). Left behind by design: Proficiency counts and test-based accountability. *The Review of Economics and Statistics*, 92(2):263–283.

Pischke, J.-S. (1995). Measurement error and earnings dynamics: Some estimates from the psid validation study. *Journal of Business & Economic Statistics*, 13(3):305–314.

Reckase, M. (2009). *Multidimensional Item Response Theory*. Springer.

Rivkin, S. G., Hanushek, E. A., and Kain, J. F. (2005). Teachers, schools, and academic achievement. *Econometrica*, 73(2):417–458.

Rockoff, J. E. (2004). The impact of individual teachers on student achievement: Evidence from panel data. *The American Economic Review*, 94(2):247–252.

Rosenbaum, P. R. and Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70(1):41–55.

Rothstein, J. (2009). Student sorting and bias in value-added estimation: Selection on observables and unobservables. *Education*, 4(4):537–571.

Rothstein, J. (2010). Teacher quality in educational production: Tracking, decay, and student achievement. *The Quarterly Journal of Economics*, 125(1):175–214.

Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of educational Psychology*, 66(5):688.

Rubin, D. B., Stuart, E. A., and Zanutto, E. L. (2004). A potential outcomes view of value-added assessment in education. *Journal of educational and behavioral statistics*, 29(1):103–116.

Todd, P. E. and Wolpin, K. I. (2003). On the specification and estimation of the production function for cognitive achievement*. *The Economic Journal*, 113(485):F3–F33.

Winters, M. A., Dixon, B. L., and Greene, J. P. (2012). Observed characteristics and teacher quality: Impacts of sample selection on a value added model. *Economics of Education Review*, 31(1):19–32.

Wooldridge, J. M. (2010). *Econometric Analysis of Cross Section and Panel Data*. MIT Press.