Sending Teacher Value-Added into Tailspin: A Simulation Study of Measurement Error and Nonrandom Sorting

By

Cassandra M. Guarino, Eun Hye Ham, Mark D. Reckase, Brian Stacy, Jeffrey M. Wooldridge

Version 12/31/13

Abstract: As a result of federal efforts to hold individual teachers accountable for student learning based on the standardized test results of their students (Race to the Top, U.S. Department of Education, 2009), the use of value-added models of teacher performance as a component of teacher evaluation has been spreading rapidly. Often researchers and policymakers alike feel comfortable with imprecision in the models because they are usually employed primarily to identify teachers in the *high* and *low* ends of the distribution in effectiveness. Using these models to reward teachers at the top and sanction those at the bottom seems a relatively attainable goal.

However, the question of whether we can "get the tails of the distribution right" has not been adequately resolved in the research literature. Related to this is the issue of measurement error in student test scores. This paper addresses the issue of measurement error and its potential to introduce bias in the tails of the effectiveness distribution and sound a warning note that identifying teachers at the top and bottom of the scale may be

less straightforward than previously thought. We produce simulation evidence on the performance of various value-added estimators in the presence and absence of measurement error and show that under certain conditions test measurement error can induce a visible bias in estimation. Moreover, we find that estimators found in the literature that attempt to correct for measurement error are shown to be largely unbiased when assignment is based on true scores (which contain no measurement error) but biased when assignment is based on observed scores (which contain measurement error).

Section 1: Introduction

As a result of federal efforts to hold individual teachers accountable for student learning based on the standardized test results of their students (Race to the Top, U.S. Department of Education, 2009), the use of value-added models of teacher performance as a component of teacher evaluation has been spreading rapidly. McGuinn (2012) reported that at least 43 states required annual teacher evaluations and 32 incorporated student performance measures. That number has likely risen since then.

Value-added models have come under strong criticism for their imprecision and potential for bias (e.g., Rothstein, 2008; Baker et al., 2010). However, a working paper by Kane and Staiger (2008) and a recent one by Chetty, Friedman, and Rockoff (2011) have largely assuaged fears about bias in these models. Both use clever strategies to elicit comparisons of value-added within and across teachers. Kane and Staiger (2008) compare experimental VAM estimates for a subset of Los Angeles teachers with earlier

non-experimental estimates for those same teachers and find that they are similar, suggesting that they are valid. Chetty et al. (2011) find that student achievement responds in expected ways to the entry and exit of teachers with differing value-added to a school.

Given these findings and the pressures to incorporate test-based performance measures into teacher evaluation, policymakers have forged ahead in their efforts to implement and attach stakes to these measures. Often, researchers and policymakers alike accept the imprecision in the models because they are usually employed primarily to identify teachers in the *high* and *low* ends of the distribution in effectiveness. Using these models to reward teachers at the top and sanction those at the bottom seems a relatively attainable goal.

However, the question of whether we can "get the tails of the distribution right" has not been adequately resolved in the research literature. Studies of the intertemporal correlations of value-added for individual teachers suggest a fair amount of fluctuation from year to year with regard to who falls in the tails—e.g., the bottom or top ten or twenty percent (e.g., Newton et al., 2010; Aaronson, Barrow & Sander, 2007).

Moreover, up till now, a side issue in the debate over the proper use and construction of value-added models has been the issue of measurement error in student test scores. Most models either do not address the possibility that measurement error could be affecting value-added estimates in important ways or employ measurement error correction techniques derived from other applications to deal with the problem without sufficiently exploring how well they work.

In this paper, we address the issue of measurement error and its potential to introduce bias in the tails of the effectiveness distribution and sound a warning note that

identifying teachers at the top and bottom of the scale may be less straightforward than previously thought.

We produce simulation evidence on the performance of various value-added estimators in the presence and absence of measurement error and show that under certain conditions measurement error can induce a visible bias in estimation. Included in the set of estimators that we consider are estimators that explicitly "correct" for this problem.

Two issues have the potential to make value-added estimation in the presence of measurement error problematic. The first concerns non-uniform measurement error, which make test score measurement error mechanically correlated with the true achievement level of a student. The second is an issue of non-random assignment of teachers to students when assignment is based on the observed test score. In this case, the measurement error is correlated not only with the prior year test score, but also the teacher assignment indicator.

We find that estimators found in the literature that attempt to correct for measurement error are shown to be largely unbiased when assignment is based on true scores (which contain no measurement error) but biased when assignment is based on observed scores (which contain measurement error). Moreover, these correction techniques may exacerbate rather than mitigate the problem of diagnosing teachers in the tails of the distribution.

This paper is organized as follows. Section 2 describes the sources of measurement error in student test scores. Section 3 describes our simulation methodology. Section 4 describes our findings. Section 5 concludes.

Section 2: Sources of Measurement Error in Assessing Student Achievement

A typical representation of measurement error is as follows, in which student *i*'s measured achievement A^*_{ig} on the test for grade *g* is a linear function of true achievement and a measurement error e_{ig}.

$$A^*_{ig} = A_i + e_{ig} \quad (1)$$

There are multiple possible sources of the errors, e, generally classified in three categories: examinee-related factors, administration-related factors, and test design factors (Boyd et al., 2012; Crocker & Algina, 2008; Raykov & Marcoulides, 2011; Thorndike, 2005). *Examinee-related factors* are due to student attributes or conditions that are irrelevant to their true achievement level. For example, students differ in their familiarity with test-taking as well as in test anxiety, which might depend on grade levels as well as test experience. Their performance might also be affected by individual students' temporary moods or physical conditions during testing. The perceived pressure from families, teachers and peers to receive high scores varies among students. *Administration-related factors* also influence individual test scores. For example, there could be auditory distracters depending on the environment. Those sources of error are likely unpredictable and randomly occurring. This study, however, focuses on more predictable errors that stem from the intrinsic characteristics of a test itself, the *test-design factors* – i.e., the limited number of items and characteristics of the items. If an unlimited number of items could be administered to students and the items covered all levels of difficulty, we could get more precise approximation of student learning levels. In practice, we select a feasible number of items, and a test is likely to have more items

targeting a certain range of achievement level, depending on the purpose of the test. As a result, the combination of items chosen alters the relationship between observed scores, A^*, and errors, e, in a test. In this paper, we focus on the effect of these systematic measurement errors, as a function of test items and student true achievement, on teacher value-added measures.

In the item response theory (IRT) framework, which has been widely applied in most state testing programs, each test item's response is modeled as a function of item characteristics and a person's ability (Lord, 1980; Hambleton, et al., 1991; Reckase, 2009). There is some variation among item response models depending on the number of identified item parameters or the number of response categories. We applied the 3-parameter logistic (3PL) model[1], which specifies the three parameters for each item *k* in which the response is dichotomous: a_k, a discrimination parameter; b_k, a difficulty parameter; and c_k, a guessing parameter. The 3PL model in equation (2) shows[2], the probability that person *i* with a "true" achievement of A_i will give a correct answer to item *k*, $\mathrm{P}(u_k = 1 | A_i)$, can be modeled as a function of the item's characteristics as well

[1] Different state testing programs apply different models for IRT calibration, depending on the number of item parameters to be estimated - the 1PL, 2PL or 3PL. We chose the 3PL model for this study because it is the most sophisticated among the typically used models, and it was also used by the state from which we obtained a set of item-parameters.

[2] In the equation, 1.7 is a scale constant typically used to minimize the difference between the logistic function and the normal cumulative function.

as the person's true achievement. Students' true achievement cannot be directly observed in practice and can only be estimated given a set of item responses. It is also assumed in this model that a set of items assesses only one common construct of interest.[3]

$$\mathrm{P}(u_k = 1 | A_i) = c_k + (1 - c_k)\frac{e^{1.7a_k(b_k - A_i)}}{1 + e^{1.7a_k(b_k - A_i)}} \quad (2)$$

Uncertainty about student achievement estimates based on a given set of items can be quantified as the variability among students at the same level of achievement in the probability of a correct answer to each item. The variability is heteroskedastic across the levels of estimated achievement and is called the conditional standard error of measurement (CSEM). As shown in Equation (3), a test's CSEM is the sum of each item's CSEM; each item's CSEM is determined by the expected probability of a correct answer; the expected probability of a correct answer is determined by the estimated person achievement level as well as item parameter.

$$\sigma_e^2(A|A^*) = \sum_k^n P_k(A^*) \cdot Q_k(A^*)$$
$$where\ P_k(A^*) \cdot Q_k(A^*) = \frac{P_k(A^*)(1 - c_k)^2}{\alpha_k^2\big(1 - P_k(A^*)\big)(P_k(A^*) - c_k)^2} \quad (3)$$

where A^* is the estimated student ability; $P_k(A^*)$ is the probability of correct answer to item k conditional on A^*; $Q_k(A^*)$ is

[3] Test multi-dimensionality (Reckase & Li, 2007) or construct-shift over grade levels (Martineau, 2006) is an emerging issue in VAM but is outside the scope of this study.

the probability of an incorrect answer to item k conditional on A^*; n is the number of items

Figure 1 shows an example of conditional standard errors of estimate computed by the authors using testing data from a large, diverse anonymous southern state for which we have item-level testing data.

The degree of the uncertainty of estimated student achievement is smaller in the middle of the achievement level, specifically between -1 and 0. It is much larger at the extremely low or high scores. The U-shape is not symmetric. This is distinct from a universal measurement error, which is usually defined in classical test theory as a single value for each test. Classical measurement error is the average, across persons, of the variance of test scores as shown in Equation (4).

$$\sigma_e^2 = \sigma_{A^*}^2 \sqrt{1 - \sigma_A^2 / \sigma_{A^*}^2} \qquad (4)$$

Section 3: Simulation Methodology

To investigate the impact of test-design induced measurement error on teacher value-added estimates, we conduct simulated experiments. For these experiments, we generate "observed" student test scores based on the students' true achievement level and the characteristics of testing items we create. The students' true achievement levels are a function of their baseline achievement, their heterogeneous individual time-constant ability, their teachers' effectiveness levels, and a random error. The test characteristics are modeled on those of standardized tests in the state mentioned above.

We consider two types of teacher effects—null effects and varying effects. That is, in a first set of experiments, we set teacher effects to zero, so that all teachers have equal null effects. In a second set of experiments, we allow for varying teacher effects.

We also allow for different ways of assigning students to teachers. Across all of these different teacher effectiveness and assignment scenarios, we assess the ability of different types of value-added estimators to accurately capture the true teacher effects that were embedded in the data generating process. Below, we describe our methodology in greater detail.

3.1. Score Generating Processes

We simulated observed test-scores for cohorts of 320 students each that were created based on hypothetical student true scores and hypothetical achievement test items. The process was carried out in five steps.

First, we generated each student's true achievement scores for three consecutive years using the following simple value-added model for achievement, where achievement is a cumulative process. We assume that the value-added model holds for the true achievement level of students.

$$\begin{aligned} A_{i4} &= \lambda A_{i3} + E_{i4}\beta + c_i + \epsilon_{i4} \\ A_{i5} &= \lambda A_{i4} + E_{i5}\beta + c_i + \epsilon_{i4} \end{aligned} \tag{5}$$

We chose to call the baseline achievement level grade 3 achievement, A_{i3}, because grade 3 is the earliest grade required to take a state test in most states. The baseline achievement score is drawn from a normal distribution with mean 0 and standard

deviation 1 and is also correlated to c_i, which was normally distributed with a mean of .5 and a standard deviation of .5. The correlation between A_{i3} and c_i was .5. The true achievement level for student *i* in grades A_{i4} or A_{i5}, is a function of the student's prior year true achievement level, a vector of educational inputs, E_{ig} (in this case teacher dummy variables), a student fixed effect, c_i, and an idiosyncratic error term, ϵ_{ig}. We assume a persistence parameter, λ, of .5, corresponding to the level of decay found in prior studies (e.g., Andrabi et al., 2009).[4]

As mentioned above, we explore two types of teacher effects. Our first thought experiment involves setting all teacher effects to 0—in other words, we wanted to create a situation in which all of the true underlying teacher effects were identical. In this way, we could clearly see whether and how the estimated effects might be biased in the presence of measurement error under various student-teacher assignment scenarios. Thus in these simulations, both the mean and the variance of the teacher effects was set to 0. In the second thought experiment, the teacher effects vary and are drawn from a normal distribution with a mean 0 and a standard deviation of .25.

In the second step, we determined our hypothetical test characteristics by specifying a set of test item parameters for each grade: each item is characterized by its own difficulty, discrimination, and guessing parameters according to the 3PL model shown in Equation (2). We created three item parameters per item (difficulty, discrimination, and guessing parameters) modeled on those of the aforementioned

[4] We test the sensitivity of this choice by setting lambda to .75 and 1. For brevity, these results are not reported in the paper.

southern state's tests were used to create approximately 50 dichotomous items each for grade 3 (the base year), grade 4 and grade 5 in our simulation, respectively.[5]

In the third step, the probability that each simulated student with his or her given level of true ability gave a correct answer to each item in the test was computed based on the 3PL model in equation (2) with the student's true abilities and the item's parameters plugged into the model.

Fourth, given the computed probabilities, individual student responses to the set of items were predicted introducing a random error in each student's item response level in the following manner. The individual's computed probability of a correct answer to an item, $\mathrm{P}(u_k = 1 | A_i)$, was compared with a random number drawn from a uniform distribution bounded between 0 and 1. If $\mathrm{P}(u_k = 1 | A_i)$ was larger than the random number, then '1' (for a correct answer) was assigned; otherwise, '0' (for an incorrect answer) was assigned.

Finally, in the fifth step, each student's *observed* achievement score was estimated from the generated item responses based on the IRT model in equation (2) assuming their true scores were unknown[6].

---

[5] These item parameters were obtained by the authors by calibrating the actual state test's student item response data in these three grades – we only include dichotomous items – using PARSCALE.

[6] We estimate each student's A^*_{ig} by maximum likelihood estimation of equation (2) using that student's set of item responses for all items on the test for grade *g*.

3.2. Grouping and Assignment Scenarios

In order to investigate bias in teacher value-added estimates under different contextual conditions, we simulated different ways of assigning teachers to students. In the first step, students were grouped into 16 classrooms of 20 students in one of two ways. Students were either randomly grouped into classrooms (RG) or dynamically grouped (DG) into classrooms, meaning that students were grouped in classrooms based on their prior achievement level. Some randomness was included in the dynamic grouping process, so that classrooms were not formed of, say, the 20 absolute worst students. In each grade, 16 classrooms were formed where the first classroom tended to contain the students with the worst prior year score and the 16th classroom tended to have the best.

We conduct the simulated experiments twice, once using the observed test scores and once using the true student achievement scores to highlight the impact of test-induced measurement error on teacher value added estimates.

3.3. Estimators

We consider five commonly used estimators, including two measurement error correction estimators. For the estimator that we term *Dynamic Ordinary Least Squares (DOLS)*, we regresses current scores – either true scores, A_{ig}, or our "observed" IRT scale scores, A^*_{ig} depending on the simulation– on the prior year's score, either A_{ig-1} or A^*_{ig}, and teacher indicators. Guarino, Reckase, and Wooldridge (forthcoming) found this

estimator to be the most robust estimator across a variety of simulated student grouping and assignment conditions because it essentially controls for or closely proxies the assignment mechanism under nonrandom assignment. However, that study considered only true scores and scores with the addition of classical measurement error. How DOLS performs under test-induced measurement error was not investigated.

For the estimator that we term *Pooled OLS (POLS),* we regress the gain score in achievement (the current minus the prior score) on teacher indicators. POLS implies a no decay assumption on prior achievement. For both estimators, the regression coefficients associated with each teacher indicator are regarded as the teacher value-added estimates.

A well known result is that under classical measurement error assumptions, inclusion of a variable measured with error can bias OLS estimates of not only the coefficient on the mismeasured variable but also coefficients on other regressors. When measurement error, e_{ig}, is added to the achievement measures, the general value-added model shown in Equation (5) will be changed as follows:

$$\mathrm{A}_{ig} = \lambda \mathrm{A}_{i,g-1} + \mathrm{E}_{ig}\beta + \mathrm{c}_i + \epsilon_{ig} + e_{ig} - \lambda e_{i,g-1} \qquad (6)$$

The estimate of the coefficient on the prior year test score is attenuated towards zero under classical measurement error—i.e., when e_{ig} is randomly distributed across prior achievement levels. The bias for other covariates is difficult to characterize generally, but depends partially on the correlation between the covariate and the mismeasured variable. In the case of no correlation between the covariate and the mismeasured variable, the estimates of the coefficient for that covariate are not affected by the measurement error in the mismeasured variable.

In general it is even more difficult to characterize the size or direction of the bias caused by measurement error that does not conform to the classical case (such as the type of measurement error generated through the IRT process in our simulated observed scores). It is no longer necessarily the case that the estimates are attenuated. In general, even the direction of the bias for the mismeasured variable is unclear. For covariates, correlated with mismeasured independent variables, it is also unclear.

In the case of test score floors and ceilings, the true achievement level of the student and the test score are mechanically correlated with one another. In this case, then measurement error in the dependent variable can also cause bias in the estimates.

In our simulations, we consider two estimators that attempt to "correct" estimates for measurement error. The first is generally termed the *Errors in Variables Regression (EIVReg)* estimator. This estimator is in frequent use in policy (see, for example, the NYC 2010 technical report (Value-Added Research Center 2010) and in the value-added literature (e.g., Walsh & Isenberg, 2013). The estimator is similar to OLS except that an estimate of the measurement error variance is subtracted from the diagonals of the X'X matrix to adjust the estimates for measurement error. In our case, the estimator is:

$$\begin{pmatrix} \hat{\lambda}_{EIVReg} \\ \hat{\beta}_{EIVReg} \end{pmatrix} = \begin{pmatrix} A'_{g-1}A_{g-1} - \sum_{i=1}^{N} var(e_{i,g-1}) & A'_{g-1}X_g \\ X'_g A_{g-1} & X'_g X_g \end{pmatrix}^{-1} \begin{pmatrix} A'_{g-1}A_g \\ X'_g A_g \end{pmatrix} \qquad (9)$$

It can be shown that under classical measurement error assumptions and under the assumption that the measurement error in $A_{i,g-1}$ is uncorrelated with $X_{i,g}$, then this estimator is consistent (Fuller 1987).

Another approach potential way to adjust for measurement error is regression calibration, which involves forming the best linear predictor of the true prior year achievement level given the IRT score and the other covariates. We use the Heteroskedastic Errors in Variables (HEIV) version of this estimator, which takes into account the heteroskedastic nature of the measurement error when forming the best linear predictor. A full explanation is given in Sullivan (2001) and Andrabi et al. (2011).

Generally speaking, it can be shown that by regressing the IRT achievement score on the best linear predictor of the prior year achievement score and the other covariates using OLS, one can consistently estimate the teacher effects in equation (6). In the typical formulas used to form the best linear predictor, it is assumed that any measurement error is uncorrelated with the true prior year achievement level as well as the other regressors. When this assumption is violated, the estimator does not generally give consistent estimates.

An additional issue that warrants attention is the case in which principals group and assign students to teachers based on their observed scores. In this case the teacher assignment variable is correlated with the measurement error as well as the underlying true score. In this case, it is theoretically unclear whether using one of the measurement error corrected estimators is superior to using a method that ignores measurement error.

In the case of the DOLS estimator, the moment conditions required to consistently estimate the parameters are violated in two ways under nonrandom assignment of this nature. The measurement error is correlated with both the lagged achievement score and the teacher assignment indicators. Whether the bias is worse or better for teacher value-

added estimators than in the case where grouping and assignment is based on true scores is difficult to say in general.

In the case of the EIVReg estimator, the moment condition is corrected in the case of the lagged test score, but not for the teacher assignment indicators, so some bias may emerge when grouping is done in this way.

The regression calibration estimators also fail, because in these methods the measurement error is assumed to be uncorrelated with not only the true score but also all other covariates in the regressions. When this assumption is not met, the steps used to create the best linear predictor of the lagged true scores are not generally valid.

A final estimation approach that we consider is the growth percentile method, such as that used in the Colorado Growth Model (CGM) as described in Betebenner (2011, 2012). Since this estimation process is based on multiple steps involving predicted quantiles, it does not conform to the discussion above. The literature on how measurement error affects quantile regression is less developed (one example of a study on this topic is Hausman 2001). However, given that the CGM is commonly used in practice, it is unclear how the estimator will perform empirically under measurement error, and we thus investigate its performance in our simulation.

Section 4: Results

Results from Experiment with Null Teacher Effects

We first report on the simulations containing null teacher effects. We present the teacher value added estimates from the 100 simulation replications per scenario in the form of box and whisker plots, which allow us to examine both the precision and bias of value-added estimators over the simulation repetitions. In each of the 100 replications per

scenario, the teachers are consistently assigned to classrooms in rank order of prior scores. In other words, teacher 1 always gets the lowest scoring classroom and teacher 16 always gets the highest ranked classroom. Thus the box and whisker plots depict a probability distribution of the value-added estimates for each teacher under these classroom assignment conditions over 100 sample draws. Another way of thinking about this experiment is to consider how the value-added estimates for a single teacher might differ under 16 different types of classroom environments.

In the plots, the line near the center of each box represents the median teacher value-added estimate. The bottom of the box represents the 25th percentile, and the top of the box represents the 75th percentile from the 100 simulation repetitions. Since all the teacher effects are zero in the simulation, a scenario in which the majority of a teacher's VAM estimates over the 100 repetitions differ from zero indicates some bias.

The first set of boxplots, shown in Figures 2 and 3, show the value-added distributions for our set of five estimators computed in the most basic manner (i.e., with only one lag of test scores and with no nonlinear functions of prior test scores, such as polynomials or b-splines, on the right-hand-side). For each simple estimator, we show two plots. In the first plot, we see value-added distributions when test scores are true scores—i.e., we use the underlying true score for each student without computing observed scores from them. Thus in these plots, the scores have no measurement error. The second plot for each estimator depicts the distributions when we use observed scores in our regressions and as a basis for classroom assignment.

For DOLS in Figure 2, we see no bias in the case in which true scores are used but a notable tendency for bias in teacher value-added when we go from true scores to the

observed IRT scores, with the teacher of classroom 16—i.e., the classroom with the best students—falsely appearing, by and large, to be more effective.

For POLS in Figure 2, we see bias in both the true score case and the observed score case. POLS, which uses a gain score as the dependent variable and contains no lagged test score on the right-hand-side, has no way of controlling for the grouping of students into classrooms. Thus it is biased even when true scores are used. The transition to observed scores only serves to accentuate the bias. It is interesting to note that the bias is in the opposite direction of that seen for DOLS under test measurement error.

For the CGM also shown in Figure 2, we find evidence of bias when true scores are used. In the observed score scenario, that bias is somehow compensated for, and the picture shows only a slight nonlinear pattern in bias. It should be noted here that this estimator does not correspond to the actual CGM in policy use because it does not contain multiple lags of test scores or a B-spline function of these lags. It is rarely applied in the simple manner shown here. The CGM with its usual features appears in a later figure.

Figure 3 shows the results for the two estimators that attempt to "adjust" for measurement error. Both these estimators are equivalent to DOLS in the case in which true scores are used (this is because the computed reliabilities reduce to an adjustment factor of 1). Thus both estimators will be unbiased under true scores. When observed scores are used, EIV does not inspire confidence, particularly on the lower tail, and HEIV suffers from severe bias.

Since the IRT process is a nonlinear transformation of true scores, we also investigate what happens when we include nonlinear functions of the lagged test score in DOLS and the CGM. With regard to the latter approach, the lagged test score is generally entered into the quantile regressions using a B-spline function, thus this version corresponds more accurately to the version used in practice.

Figure 4 shows the results when these more flexible estimators are applied to the observed scores. The simple inclusion of multiple lags of test scores in DOLS in panel A (in our simulation there are three lags available for the "5th" grade reported here) does nothing to mitigate bias. The application of polynomials to DOLS—we include squares and cubes of all prior test scores—changes the picture to restore a median of around zero. However, there is a visible loss of precision in the tails. Not surprisingly, the same pattern emerges with the use of a B-spline function in prior test scores rather than a polynomial. The CGM with the B-spline looks best but the tails appear slightly biased downward.

In summary, this simple thought experiment in which teacher effects do not vary but classroom achievement does has demonstrated that bias may creep into teacher value-added estimation due to test-based measurement error. The direction of bias may vary with the estimation approach used. In particular, the tails of the teacher effectiveness distribution may not be trustworthy when simple estimators are employed.

Results from Experiment with Varying Teacher Effects

For the experiment in which we vary teacher effects, we report our results in the form of tables displaying rank correlations between the true teacher effects and the estimated ones. The tables show these correlations for different estimators (columns) and

different teacher-student grouping scenarios (rows). In these simulations, we create a more complete set of scenarios for grouping students into classrooms and assigning the to teachers. We group students in classrooms both randomly and dynamically (based on prior test scores). Also we can assign these classrooms to teachers in different ways depending on whether teachers are randomly chosen to teach certain classrooms or nonrandomly chosen. In the latter case, the teacher to classroom assignment can be "positive," meaning that the better teachers are assigned to the classrooms with higher prior test scores, or "negative," meaning that the better teachers are assigned to the classrooms with lower prior test scores.

Table 1 shows results for the simple estimators and the two measurement error correction estimators discussed in Figures 2 and 3 of the previous experiment. Whereas the simple DOLS estimator does best in this group under true scores, it struggles somewhat with negative assignment under observed scores. POLS, EIV, and HEIV all show severe bias in one of the two dynamic grouping/nonrandom assignment scenarios in the directions expected as a result of the prior discussion. The simpleCGM performed more evenly under negative and positive assignment but still underperformed in both scenarios relative to DOLS.

Table 2 shows results for the estimators that include multiple lags, polynomials, and splines. Here we see that DOLS with two lags performs substantially better than the CGM and surpasses any of the estimators in the previous table under observed score scenarios.

These findings are, for the most part, consistent with those in the experiment with null teacher effects. Measurement error correction techniques appear to be of limited use,

and, although flexible specifications of DOLS appear to labor less in recovering teacher effects than other approaches, neither set of simulation experiments permits us to feel comfortable about effect estimates in the tails of the teacher effectiveness distribution.

Section 5: Conclusion

This paper has shown that under certain conditions—namely, when teachers are assigned high or low performing students and measurement error is present in student test scores—value added models may mischaracterize teachers as highly effective or ineffective. Moreover, models may disagree as to who is thus characterized. Surprisingly, models that attempt to correct for measurement error appear to fail in this effort.

These findings, although preliminary, should give pause to the research and policy community in attaching stakes to value-added measures of teacher performance. More research is needed to investigate the extent and impact of measurement error and its interaction with classroom assignment in actual data, and we are currently in the process of conducting such investigations. It may be that the conditions for producing bias are present only to a negligible degree, and there is little need to worry about this problem. Unless this is established, however, we should be circumspect before we assert that we have correctly estimated who belongs in the "tails" of the distribution of effective teachers.

References

Aaronson, D., Barrow, L., Sander, W. (2007). Teachers and Student Achievement in the Chicago Public High Schools. *Journal of Labor Economics*, 25(1), 95-135.

Andrabi, T., Das, J., Khwaja, A., and Zajonc, T. (2009). Do Value-Added Estimates Add Value? Accounting for Learning Dynamics, HKS Faculty Research Working Paper Series RWP09-034, John F. Kennedy School of Government, Harvard University, http://dash.harvard.edu/handle/1/4435671, accessed on 5/15/12.

Baker, E., Barton, P. Darling-Hammond, L., Haertel, E., Ladd, H. Linn, R., Ravitch, D., Rothstein, R., Shavelson, R., Shepard, L. (2010) Problems with the Use of Student Test Scores to Evaluate Teachers. *EPI Briefing Paper 278.*

Betebenner, Damian W, "A technical overview of the student growth percentile methodology: Student growth percentiles and percentile growth projections/ trajectories. The National Center for the Improvement of Educational Assessment," 2011.

, "Growth, standards, and accountability," GJ Cizek, Setting Performance Standards: Foundations, Methods & Innovations, 2012, pp. 439–450.

Boyd, D., Lankford, H., Loeb, S., & Wyckoff, J. (2012) Measuring Test Measurement: A General Approach, *NBER Working Paper 18010*.

Chetty, R., Freidman, J., & Rockoff, J. (2011). "The Long-Term Impacts of Teachers: Teacher Value-Added and Student Outcomes in Adulthood." NBER Working Paper 17699.

Crocker, L., & Algina, J. (2008). *Introduction to Classical and Modern Test Theory*. Mason, OH: Cengage Learning.

Dieterle, S., Guarino, C., Reckase, M., & Wooldridge, J. (unpublished draft) *How do Principals Group and Asign Students to Teachers? Finding Evidence in Administrative Data and the Implications for Value-added.*

Fuller, W.A. (1987). *Measurement Error Models*. New Yerk, NY: John Wiley & Sons, Inc.

Guarino, C., Reckase, M., & Wooldridge, J. (forthcoming) Can Value-Added Measures of Teacher Performance Be Trusted? *Education Finance and Policy*.

Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of Item Response Theory*. Newbury Park, CA: Sage Press

Hausman, J., "Mismeasured Variables in Econometric Analysis: Problems from the Right and Problems from the Left," *The Journal of Economic Perspectives*, 2001

Kane, T. & Staiger, D. (2008) Estimating Teacher Impacts on Student Achievement: An Experimental Evaluation, Working Paper 14607, National Bureau of Economic Research.

Koedel, C. & Betts, J. (2009) Value-added to what? How a ceiling in the testing instrument influences value-added estimation. Working Paper 14778, National Bureau of Economic Research.

Martineau, J. A. (2006). Distorting value added: the use of longitudinal, vertically scaled student achievement eata for growth-based, value-added accountability. *Journal of Educational and Behavioral Statistics*, *31*(1), 35–62.

McGuinn, P. (2012) The State of Teacher Evaluation Reform: State Education Agency Capacity and the Implementation of New Teacher-Evaluation Systems, Washington: Center for American Progress.

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

Lockwood, J & McCaffrey, D. (2013). Should non-linear functions of test scores be used as covariates in a regression model? R. Lissitz (ed.), *Value-added Modeling and Growth Modeling with Particular Application to Teacher and School Effectivenes.* Charlotte, NC: Information Age Publishing.

Raykov, T., & Marcoulides G.A. (2011). *Introduction to Psychometric Theory*. New York, NY: Routledge.

Reckase, M. D. (2009). *Multidimensional Item Response Theory*. New York, NY: Springer.

Reckase, M. D., & Li, T. (2007). Estimating gain in achievement when content specifications change: a multidimensional item response theory approach. In R. W. Lissitz (Ed.), *Assessing and modeling cognitive development in school*. Maple Grove, MN: JAM Press.

Rothstein, J. (2008) Teacher Quality in Educational Production: Tracking, Decay, and Student Achievement. *NBER Working Paper Series,* Working Paper 14442, http://www.nber.org/papers/w14442.

Sanders, W. & Horn, S. (1994) The Tennessee Value-Added Assessment System (TVAAS): Mixed-Model Methodology in Educational Assessment. *Journal of Personnel in Education*, 8, 299-311.

Sullivan, D. (2001) A Note on the Estimation of Linear Regression Models with Heteroskedastic Measurement Errors. Working Paper.

Thorndike, R. M. (2005). *Measurement and Evaluation in Psychology and Education* (7th ed.). Upper Saddle River, NJ: Pearson Education.

US Department of Education (2009) Race to the Top Program: Executive Summary, http://www2.ed.gov/programs/racetothetop/executive-summary.pdf, accessed on 9/8/10.

Value-Added Research Center (2010) NYC Teacher Data Initiative: Technical Report on the NYC Value-Added Model. http://schools.nyc.gov/NR/rdonlyres/A62750A4-B5F5-43C7-B9A3-F2B55CDF8949/87046/TDINYCTechnicalReportFinal072010.pdf, accessed on 5/15/12.

Walsh, E. & Isenberg, E. (2013) How does a value-added model compare to the Colorado Growth Model. Mathematica Policy Research, Working Paper.

Tables and Figures

Table 1. Rank Correlation between True Teacher Effects and Value-added Estimates under True and Observed Score Scenarios

Estimator	Score type	Grouping and Assignment Scenario			
		Random-Random	Dynamic-Random	Dynamic-Positive	Dynamic-Negative
DOLS	TRUE	0.84	0.84	0.86	0.83
	Observed	0.81	0.81	0.85	0.57
POLS	TRUE	0.83	0.81	0.61	0.88
	Observed	0.77	0.66	0.11	0.86
CGM	TRUE	0.74	0.70	0.14	0.77
	Observed	0.76	0.75	0.53	0.50
EIV	Observed	0.80	0.79	0.87	0.21
HEIV	Observed	0.81	0.80	0.86	0.24

Table 2. Rank Correlation between True Teacher Effects and Value-added Estimates based on Multiple Lags, Polynomials, and Splines under Observed Score Scenarios

Estimator	Score type	Grouping and Assignment Scenario			
		Random-Random	Dynamic-Random	Dynamic-Positive	Dynamic-Negative
DOLS-Two Lags	Observed	0.80	0.81	0.81	0.80
DOLS-Polynomial	Observed	0.81	0.81	0.76	0.63
DOLS-B-spline	Observed	0.81	0.81	0.76	0.63
CGM-B-spline	Observed	0.76	0.74	0.52	0.50

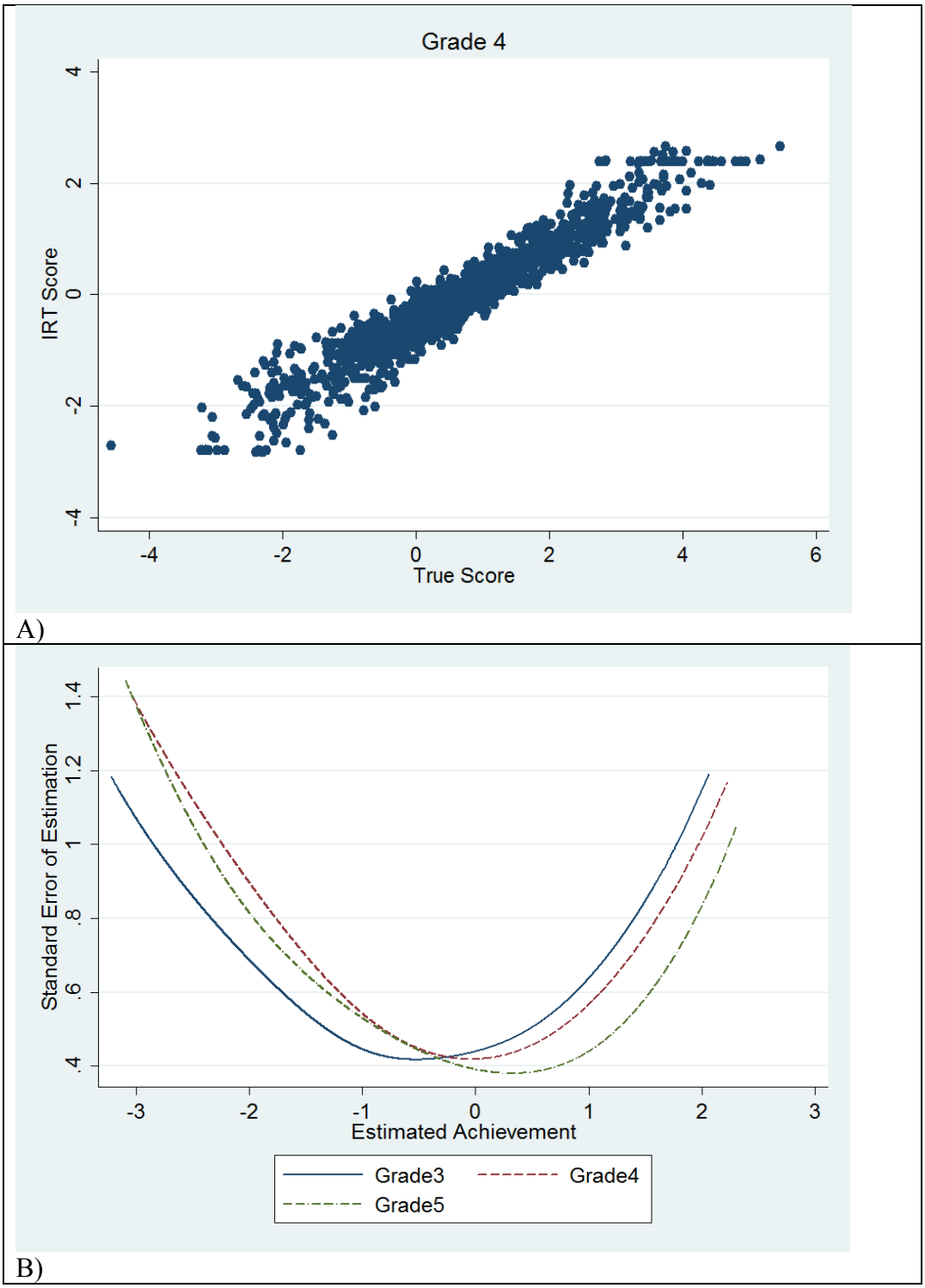


Figure 1. Examples of the Simulated Student True and Observed Score Distribution and the Conditional Standard Error of Measurement. A) Scatter plot of the simulated true and observed scores. B) Conditional standard error of measurement of the simulated tests

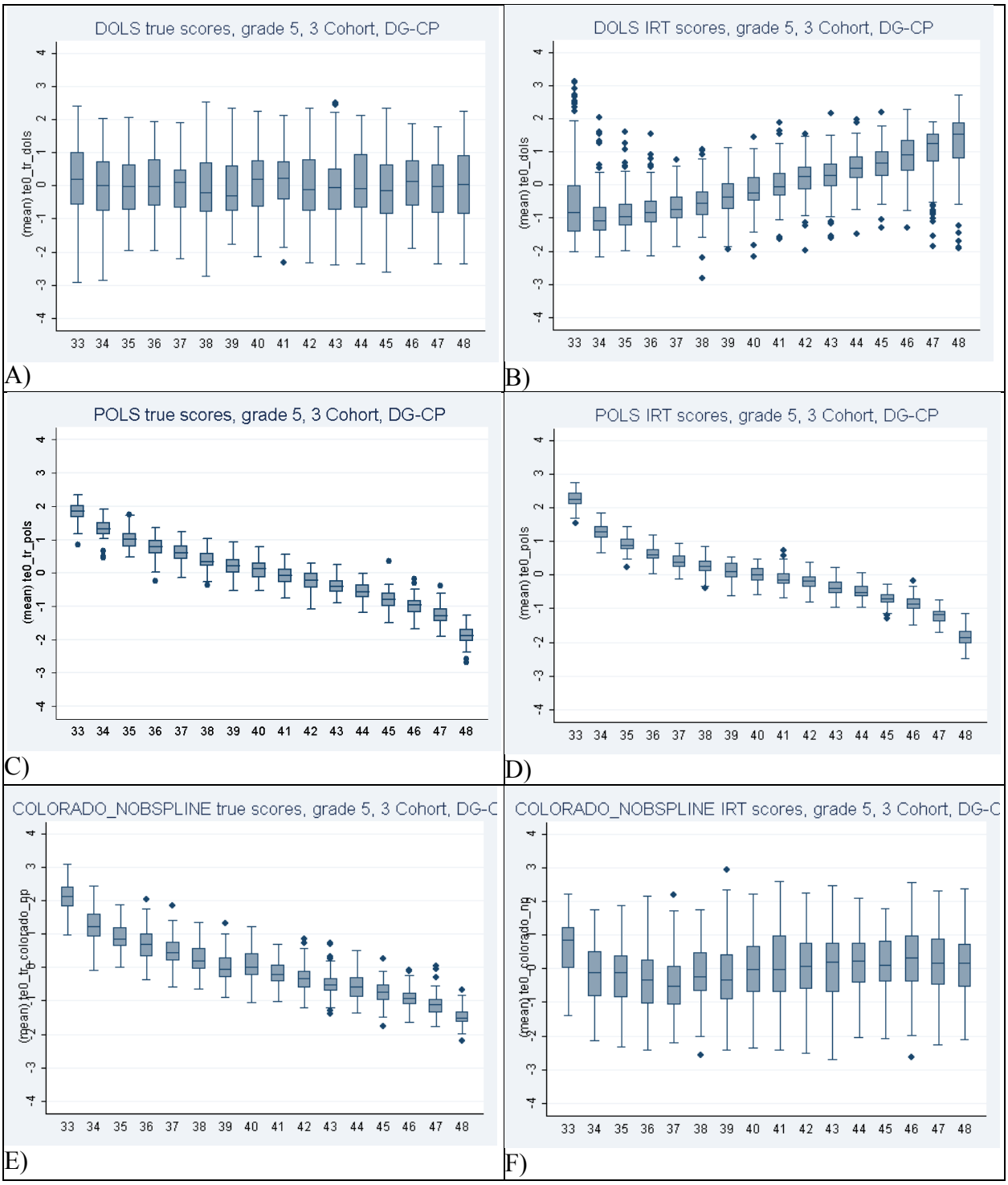
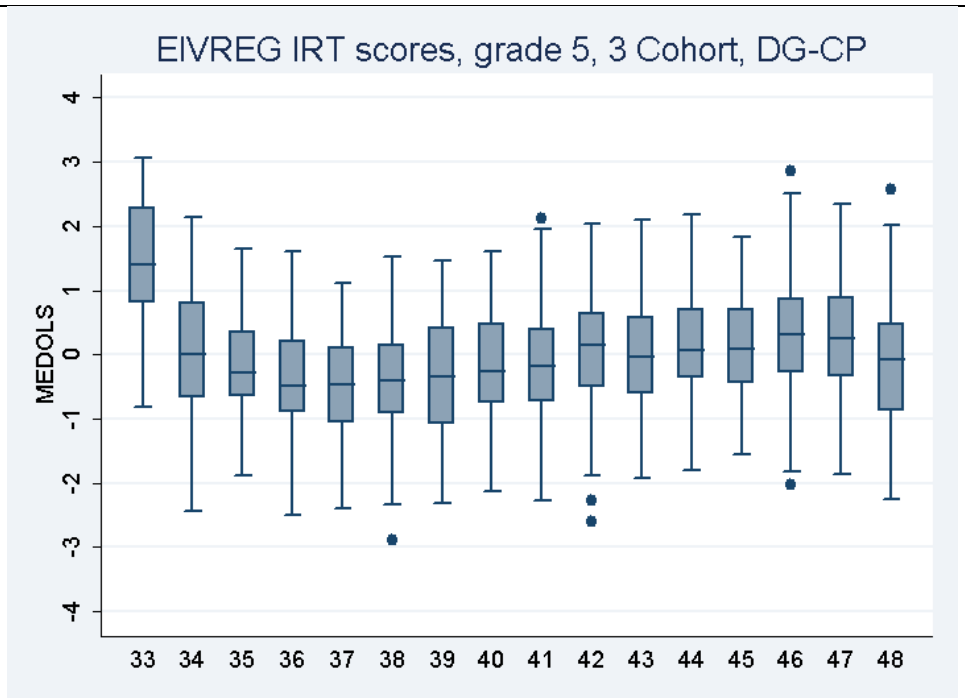
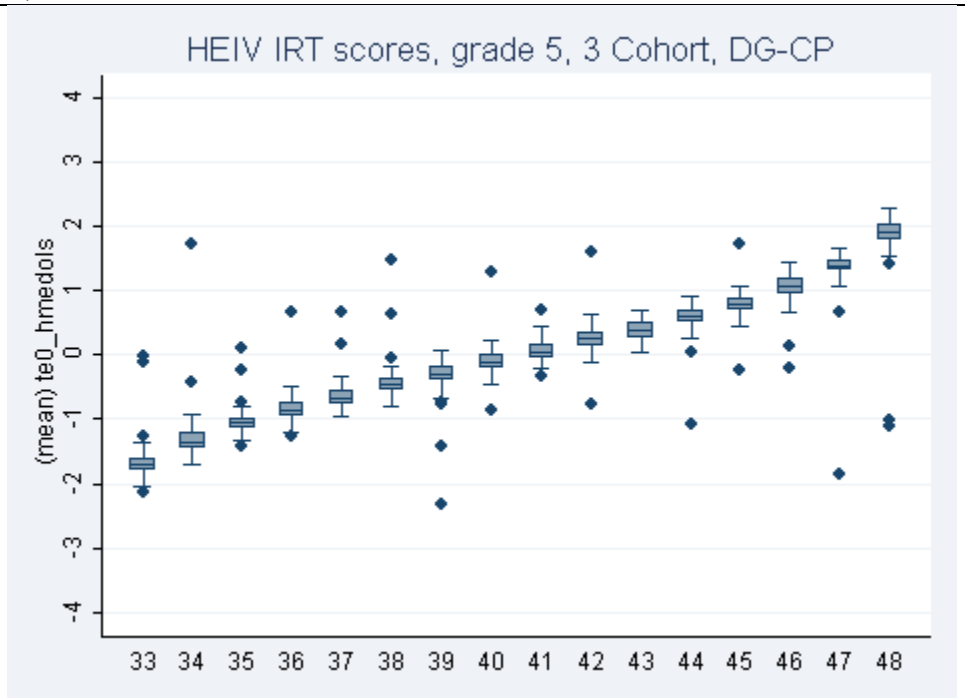


Figure 2: Simple Estimators under True and Observed Score Scenarios. All Teachers Have Null Effects.



A)



B)

Figure 3: Measurement Error Correction Estimators. All Teachers Have Null Effects. All Scores Used are IRT Scores.

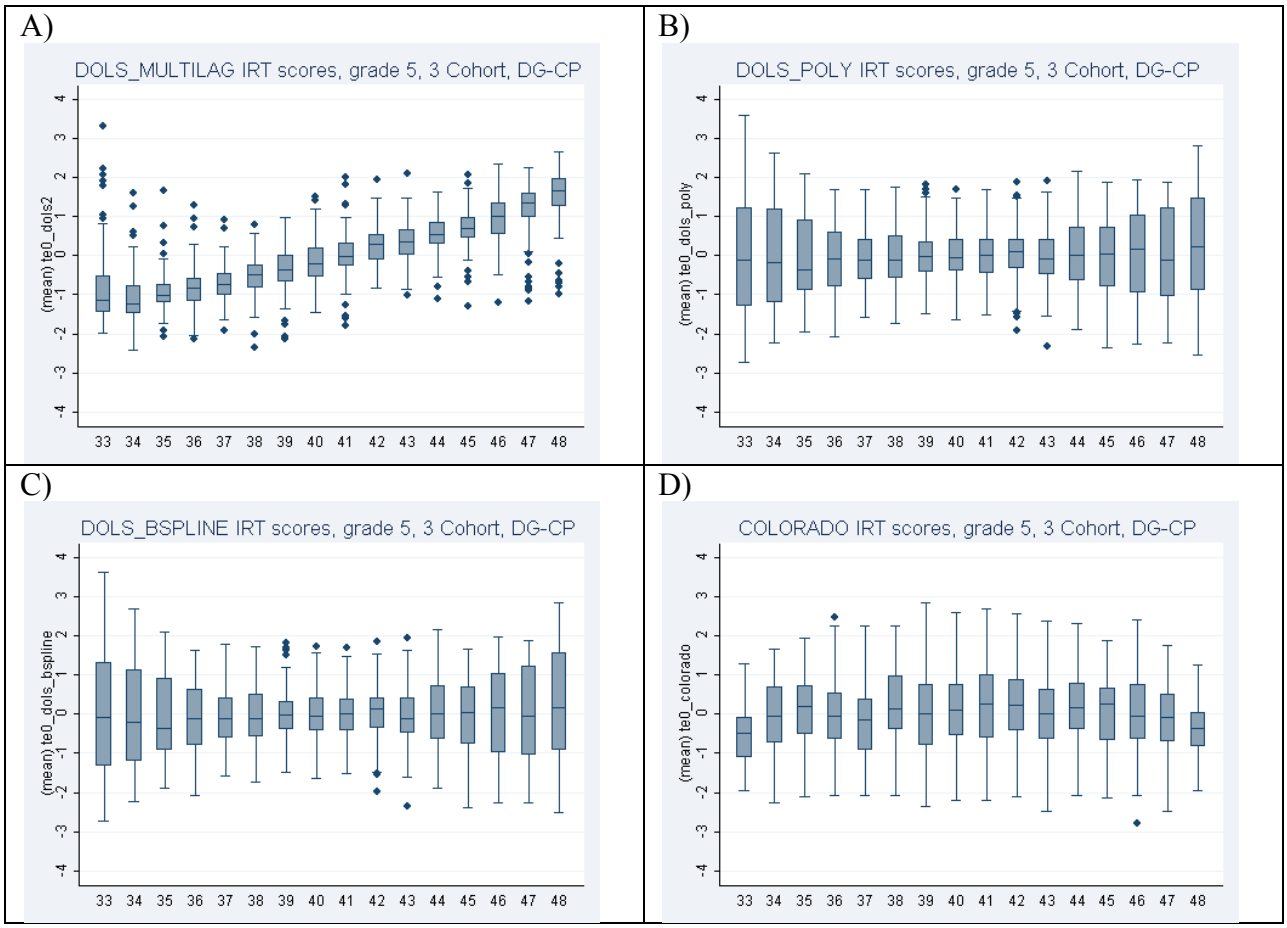


Figure 4: All Estimators Use Multiple Lags. All Teachers Have Null Effects. A) DOLS with multiple lags entered linearly. B) DOLS with multiple lags and a cubic polynomial in all three lags. C) DOLS with multiple lags and B-spline function in all three lags. D) Colorado Growth Model with multiple lags and B-spline function in all three lags.