An Analysis of Different Methods for Incorporating Co-Teachers into Value-Added Models

Jenny Gnagey, Ph. D. Candidate

Department of Agricultural, Environmental, and Development Economics

Ohio State University

Working Paper, November 2013

Abstract

Several methods for incorporating co-teachers into value-added analyses are used both in research and in practice. These different techniques rely on widely varying assumptions, yet these assumptions have not been well documented, and their validity has not been well established. As a result the properties of value-added performance metrics for co-teachers are unknown. This study examines the assumptions underlying four different value-added performance metrics for co-teachers and uses simulation analysis to empirically evaluate their properties. Due to lack of knowledge about the true mechanisms underlying the co-teaching process, I do not attempt to verify the validity of each metric's assumptions. Rather I evaluate each metric under three different data generating processes (DGPs) each of which upholds a different set of assumptions. I find that violations of assumptions can lead to substantial biases in co-teacher contribution estimates, but the degree to which these biases distort relative teacher performance rankings varies among metrics. When the percentage of co-taught students is small to moderate some metrics are more robust than others.

1. Introduction

Education data clearly show that, due to various team teaching arrangements and student mobility, it is common for students to receive instruction from more than one teacher within a subject during a school year. Recent studies consistently find that about 20% of elementary and middle school students receive team instruction, and about 20% of teachers participate in team teaching arrangements (Hock and Isenberg 2012; Ruhil, Lewis, and Yandell, 2012; Watson and Thorn 2012; Watson et. al 2011). Additionally, student mobility rates average around 10%, but rates as high as 50% are observed in some urban districts (Been et. al. 2011; Hanushek, Kain, and Rivkin 2004, Kerbow, Azcoitia, and Buell 2003). Throughout this analysis I collectively refer to all situations in which students receive instruction from multiple teachers in the same subject, whether due to team teaching or mobility, as co-teaching. Clearly co-teaching is a significant phenomenon in education systems today.

As value-added analysis is increasingly implemented on larger scales and under higher stakes, the issue of whether and how co-teachers should be incorporated into value-added analyses poses a significant modeling challenge for researchers and policy makers (Baker et. al. 2010; Corcoran 2010; Goe 2008; Steele, Hamilton, and Stetcher 2010). Conflicting findings on the nature of collaborative work has precluded a well-accepted theoretical model (Mas and Moretti 2009, Goldhaber et. al. 2011, Jackson and Brugemann 2009, Koedel 2009). Because it is common for teachers to teach few students outside a team-taught group or share only small groups of students, statistical limitations may render certain modeling approaches inappropriate in these situations (Hock and Isenberg, 2012). As a result, there is no general consensus about how value-added models (VAMs) should be used to evaluate the performance of co-teachers.

Despite these challenges, a variety of methods for incorporating co-teachers into VAMs, and subsequently constructing individual performance metrics for use in performance evaluations, have been implemented both in research and in practice (Hock and Isenberg 2012, Value-Added Research Center 2010, Wright et. al. 2010). These methods rely on widely varying assumptions, yet these assumptions have not been well documented, and their validity has not been well established. As a result the properties of these proposed performance metrics are unknown. This paper explores the impact of using different methods to incorporate co-teachers into value-added analyses in a context that acknowledges both our very limited understanding of co-teaching processes and the limits of statistical analysis. I focus on metrics that have been implemented in research and/or in practice.

First, I articulate the assumptions under which four common value-added co-teacher performance metrics can be expected to accurately estimate the contributions of co-teachers. Second, I use simulation analysis to empirically examine the ability of the four metrics to recover co-teacher contributions. I do not attempt to draw conclusions about the true nature of collaborative work. Rather I examine the estimates produced by these metrics under three data generating processes (DGPs) each of which reflects a possible mechanism underlying the co-teaching process. I discuss and quantify the existence, magnitude, and direction of biases induced when assumptions are violated, and I evaluate the robustness of the metrics to such violations.

Finally, I consider the feasibility of implementing these performance metrics in practice. I focus on the issue of multicollinearities induced when teachers teach many students as a team but few students outside the team. Such multicollinearities undermine the ability of two of the performance metrics considered here to produce reliable performance measures. In light of the simulation results, I evaluate the option of using alternative performance metrics when multicollinearities render these two metrics undesirable.

I draw two main conclusions. First, applying a model that is inconsistent with the DGP can lead to substantial biases in co-teacher contribution estimates. Second, the degree to which these biases distort relative teacher performance rankings varies among metrics, and when the percentage of shared students is small to moderate some metrics are more robust than others.

2. Literature Review

A large literature exists on the assumptions underlying VAMs, the validity of these assumptions, and the robustness of estimates to violations of these assumptions (Ballou, Sanders and Wright 2004; Kane and Staiger 2008; Rothstein 2009, 2010; Todd and Wolpin 2003). A smaller literature examines the issue of incorporating co-teachers into VAMs (Hock and Isenberg 2012, Ruhil, Lewis, and Yandell 2012; Watson and Thorn 2012, Watson et. al. 2011). Among these branches of literature, several studies are particularly relevant to this analysis.

A number of recent studies use simulation analysis to examine the robustness of various VAM specifications to violations of their underlying assumptions. The main advantage of simulation analysis is that key underlying parameters are known and estimated parameters can be compared to the known underlying parameters. Applying these techniques Ponisciak et. al. (2012) find that student-teacher linkage errors can substantially bias teacher effect estimates at error rates as low as 10%. Guarino, Reckase, and Wooldridge (2012) find that non-random student teacher assignments can significantly distort teacher effect estimates, but they find differences in robustness across specifications. The analysis by Guarino, Reckase, and Wooldridge (2012) is particularly relevant because they examine a variety of plausible data generating processes in addition to a variety of possible model specifications. Because of this similarity, this paper borrows heavily from both their framework and their methods.

The current literature on incorporating co-teachers into VAMs is small and largely qualitative. Several studies examine the prevalence and nature of team teaching arrangements (Ruhil, Lewis, and Yandell 2012; Watson and Thorn 2012; Watson et. al. 2011). As noted previously, these studies show that about 20% of both teachers and students in elementary and middle school participate in team teaching arrangements. They also show that team teaching arrangements tend to follow either a partner model or an intervention model. Partner models involve two teachers sharing a group of students with neither teacher team-teaching outside this partnership. Intervention models include arrangements such as pull-out or push-in programs in which an “intervention” teacher shares students with a variety of “mainstream” teachers who each in turn share a handful of students with the intervention teacher (Watson et. al. 2011).

A recent paper by Hock and Isenberg (2012) is the only other study of which I am aware that undertakes a quantitative analysis of methods for incorporating co-teachers in value-added analyses. While this study shares many similarities with theirs in that a similar set performance metrics are analyzed, it differs in several important ways. First, I include an additional performance metric in the comparative analysis. This metric reflects current practices in the academic literature on teacher co-production. Second, I deliberately articulate the assumptions required for each performance metric to produce accurate teacher performance measures. Finally, while Hock and Isenberg use school district data to compare their chosen performance metrics with each other, I use simulation analysis to compare these performance metrics with known underlying parameters that are the product of varying DGPs. This allows me to analyze the biases inherent in various performance metrics in a way that was not possible for Hock and Isenberg. Because of these differences, this study makes an important contribution to the literature on the inclusion of co-teachers in VAMs.

3. Co-teacher Performance Metrics and their Underlying Assumptions

In this section I consider four performance metrics for co-teachers: the three examined in Hock and Isenberg (2012) and one additional metric. I first discuss the concept of dosage because although it is applied across a variety of metrics, the literature is ambiguous on its interpretation. I then describe the modeling strategy and calculation for each performance metric and articulate the assumptions under which each metric produces accurate estimates of the contributions of co-teachers.

3.1. Interpreting Dosage Parameters

Many techniques for modeling co-teaching situations make use of the concept of dosage. In very general terms, dosage refers to the reconceptualization of binary teacher indicator variables as continuous variables that measure the "treatment dosage" a student receives from a particular teacher. Yet the literature provides nebulous interpretations of the resulting dosage variables. Some interpret dosage variables as normalized measures of absolute instructional time. Others interpret them as the fraction of total instructional time a student receives from a particular teacher in a given subject during a school year. These interpretations are quite different as demonstrated by the following example.

Suppose a student with limited English proficiency (LEP) takes both a mainstream English class and an English Language Learners class in the same school year. This student's English achievement can be described with the following very general formula:

$$y_{it} = f(\boldsymbol{k_{it}}, \boldsymbol{\tau_{it}}, \boldsymbol{z_{it}}, \epsilon_{it}) \qquad (1)$$

where y_{it} is a metric of English performance for student i in year t, $\boldsymbol{k_{it}}$ represents the set of teachers from whom the student received English instruction in year t, $\boldsymbol{\tau_{it}}$ is the set of associated teacher dosage variables, $\boldsymbol{z_{it}}$ represents all other factors that affect student performance, and ϵ_{it} is a random error term.

If one interprets the dosage variables as normalized units of absolute instructional time where one English course corresponds with one unit of dosage, each τ_{itk} for the student would take on the value of 1. On the other hand if one interprets the dosage variables as the percentage of the student's total English instruction during year t, then each τ_{itk} would take on the value of 0.5. The different dosage values will produce different teacher effect estimates, so it is important to examine the theoretical underpinnings of each of these interpretations.

The interpretations are closely tied to an important concept underlying all VAMs: the idea of a "growth standard." A growth standard is a threshold quantity of academic progress which may vary for different student groups. Teachers whose students on average surpass this threshold provide positive value-added and vice versa. Understanding the growth standard is important for understanding the interpretation of dosage variables because, when dosage is used, a teacher's value-added depends on the growth standard *per dosage unit*.

Growth standards can be defined in different ways depending on the nature of the student achievement metric. When a test is vertically scaled, academic progress can be expressed in raw test score units resulting in two possible ways of defining the growth standard. First, one could imagine an instruction-based growth standard specifying a threshold quantity of growth per English course. In the case of the LEP student each teacher would be held accountable for advancing the student by one growth standard, and letting each dosage variable take on the value of one would yield appropriately interpretable and comparable teacher effect estimates. One could also imagine a periodic growth standard specifying a threshold quantity of growth per school year. Here each teacher of the LEP student would be held accountable only for advancing the student by half a growth standard, and letting each dosage variable take on the value of 0.5 would yield appropriately interpretable and comparable teacher effect estimates.

When a test is not vertically scaled test scores are typically normalized to share a common mean and standard deviation within each grade and subject each year. Here the growth standard is typically expressed in terms of a student's ability to maintain his position in the distribution of test scores relative to his comparable academic peers. Because test scores are normalized every year (rather than after the completion of each individual course) this context automatically defines a periodic growth standard. Therefore, when tests are not vertically scaled and test scores are normalized each year dosage variables should be interpreted as the percentage of instruction a student received from a particular teacher in a given subject during a school year. Thus in the context of the LEP student letting both dosage variables take on the value of 0.5 yields appropriately interpretable and comparable teacher effect estimates.

The models below are described in the context of a periodic growth standard, but they would be equally applicable in the context of an instruction-based growth standard.

3.2. The Partial Credit Method

Maintaining consistency with the model names and notation used in Hock and Isenberg (2012) the most straightforward method of incorporating co-teachers into VAMs is the "Partial Credit Method." This is the method used by the SAS EVAAS model (Wright et. al 2010). The method modifies the traditional VAM by replacing binary teacher indicator variables with continuous dosage variables that represent the percentage of instruction a student receives from a particular teacher in a given subject during a school year. The estimating equation for this approach is

$$y_{it} = \boldsymbol{\pi}'\boldsymbol{z_{it}} + \boldsymbol{\psi}'\boldsymbol{w_{it}} + \varepsilon_{it} \qquad (2)$$

where y_{it} is a measure of academic performance for student i in year t, $\boldsymbol{z_{it}}$ represents all factors that impact student achievement other than teachers, and $\boldsymbol{\pi}$ represents the parameters associated with $\boldsymbol{z_{it}}$. The elements of the $\boldsymbol{w_{it}}$ vector are the dosage variables reflecting the percentage of instruction student i received from each teacher. The vector $\boldsymbol{\psi}$ represents the teacher effect parameters and ε_{it} is the error term. With the Partial Credit Method, the estimated parameters, $\widehat{\boldsymbol{\psi}}$, are used directly as the individual performance measure for both solo- and co-teachers.

The key underlying assumption of the Partial Credit Method is that each percentage of instruction received from a particular teacher has a constant effect on student achievement; a teacher's effectiveness is constant regardless of the teaching arrangement and there are no interaction effects. This assumption implies that teacher effects are additively linear as reflected in equation (2).

Although this assumption is intuitive, it has little empirical support. Goldhaber et. al. (2011) use this method to model cross-subject teacher impacts on student learning and find significant effects, but they do not explicitly test the hypothesis of additive linearity.

3.3. The Teacher Team Method

Another method for incorporating co-teachers in VAMs is what Hock and Isenberg (2012) term the "Teacher Team Method." Variations of this method have been used in the Washington D.C. and New York City value-added programs (Isenberg and Hock 2010, 2011; Value-Added Research Center 2010). The unifying characteristic of these methods is the technique for incorporating co-teaching situations into VAMs; however, subsequent methods for constructing individual performance measures differ. The estimating equation for this method is

$$y_{it} = \boldsymbol{\pi}'\boldsymbol{z_{it}} + \boldsymbol{\gamma}'\boldsymbol{c_{it}} + \xi_{it} \qquad (3)$$

Whereas the $\boldsymbol{w_{it}}$ matrix in (2) consisted of one column per teacher, the $\boldsymbol{c_{it}}$ matrix includes a column for each teacher and a column for each group of co-teachers. Each student is modeled as receiving instruction from either a single teacher or a particular group of co-teachers. Therefore, this model does not use dosage parameters and the $\boldsymbol{c_{it}}$ matrix contains only ones and zeros.

Because a teacher may both solo- and co-teach or participate in multiple co-teaching arrangements the parameter estimates, $\hat{\boldsymbol{\gamma}}$, will not always yield a unique performance measure for each teacher. A composite performance metric is needed. I analyze the metric used during the first year of Washington D.C.'s value-added program (Hock and Isenberg 2012, Isenberg and Hock 2010). This composite metric, $\widehat{\omega}_k$, is calculated as follows

$$\widehat{\omega}_k = \sum_{m \in M_k} p_{km} \times \hat{\gamma}_m \qquad (4a)$$

where M_k represents the set of co-teaching groups to which teacher k belongs (including the "solo team" of just teacher k), and $\hat{\gamma}_m$ is the estimated effect of co-teaching group m. Although the Teacher Team regression model does not incorporate dosage proportions, these proportions are used to determine the "weight" given to each $\hat{\gamma}_m$. Specifically p_{km} is the fraction of teacher k's total dosage attributions that come from co-teaching group m.

This is a reasonable metric if the teacher's true weighted average contribution to student learning is

$$\omega_k = \sum_{m \in M_k} p_{km} \times \gamma_m \qquad (4b)$$

where the difference between $(4a)$ and $(4b)$ is the presence and absence respectively of hats on the gamma and omega parameters differentiating between estimated and true underlying parameter values.

As a result this metric relies on several assumptions to produce accurate individual performance measures. First, this model assumes that a teacher's effectiveness can change from one teaching arrangement to another. Second, it makes the assumption that when teacher effectiveness changes between teaching arrangements, the teacher effectiveness parameter of interest is not the teacher's solo-teaching effect but rather the dosage-weighted average effectiveness across the relevant teaching contexts. Third, it assumes that effectiveness changes between teaching arrangements happen in such a way that a teacher's effectiveness in a particular co-teaching arrangement is precisely equal to the effectiveness of the other teacher(s) in the context of that arrangement. In other words within a particular co-teaching arrangement all teachers contribute equally to the team. Finally, Hock and Isenberg make an important distinction between "fully-interacted" teams and "aggregate" teams. If a model specifies fully-interacted teams, each unique combination of teachers *and* associated dosage parameters is considered a separate team. If a model specifies aggregate teams, each unique combination of teachers is considered a separate team regardless of associated dosage parameters. A model that specifies aggregate teams relies on the additional assumptions that a fixed group of teachers has a constant effect on student learning regardless of the division of teaching responsibilities.

Although these assumptions have not been tested in the field of teaching, a recent study of berry pickers found that the productivity of an individual berry picker converges to the average productivity of his friends when his friends are present (Mas and Moretti 2009). These findings are consistent with the assumption of equal contributions to the team.

3.4. The Teacher Interaction Method

I add this method to the suite of methods examined by Hock and Isenberg (2012). Although it is not applied in any value-added evaluation program of which I am aware, the interaction model has support from several recent academic studies of teacher co-productivity (Koedel and Betts 2009, Jackson and Brugman 2009). In contrast to the Teacher Team Method, instead of *replacing* individual dosage variables with co-teaching group variables, this technique *adds* co-teaching group variables while retaining the full set of individual dosage variables from equation (2). For this reason I call it the "Teacher Interaction Method." The estimating equation for this method is

$$y_{it} = \boldsymbol{\pi}'\boldsymbol{z_{it}} + \boldsymbol{\varphi}'\boldsymbol{v_{it}} + \mu_{it} \qquad (5)$$

As indicated above $\boldsymbol{v_{it}}$ includes a column for each teacher and a column for each group of co-teachers. The individual columns are exactly as they were in $\boldsymbol{w_{it}}$ and the group columns are exactly as they were in $\boldsymbol{c_{it}}$. This can be interpreted as the student receiving a certain dosage of each teacher in the co-teaching arrangement in addition to a full year's dosage of the interaction effect of the co-teaching group as a whole.

Like the Teacher Team Method, the Teacher Interaction Method does not always yield a unique individual performance measure. I construct a composite performance metric, $\hat{\theta}_k$, that parallels the technique used to calculate the Teacher Team composite performance metric.

$$\hat{\theta}_k = \sum_{m \in M_k^*} p_{km} \times (\hat{\varphi}_k + \hat{\varphi}_m) = \hat{\varphi}_k + \sum_{m \in M_k^*} p_{km} \times \hat{\varphi}_m \qquad (6a)$$

Here $\hat{\varphi}_k$ indicates the parameter associated with teacher k's dosage variables which can be interpreted as the solo-team effect. M_k^* denotes the set of co-teaching arrangements to which teacher k belongs excluding the solo team.

Similar to the Teacher Team metric, $\hat{\theta}_k$ is a reasonable metric of teacher performance if the true weighted average contribution of a co-teacher to student learning is

$$\theta_k = \sum_{m \in M_k^*} p_{km} \times (\varphi_k + \varphi_m) = \varphi_k + \sum_{m \in M_k^*} p_{km} \times \varphi_m \qquad (6b)$$

where again the difference between $(6a)$ and $(6b)$ is the presence and absence respectively of hats on the phi and theta parameters.

The Teacher Interaction metric, $\hat{\theta}_k$ (like the Teacher Team performance metric $\widehat{\omega}_k$) assumes that a teacher's effectiveness can change from one teaching arrangement to another. It also assumes that when teacher effectiveness changes between teaching arrangements, the teacher effectiveness parameter of interest is not the teacher's solo-teaching effect but rather the dosage-weighted average effectiveness across the relevant teaching contexts. The Teacher Interaction metric assumes a specific underlying structure governing changes in effectiveness across teaching arrangements. It assumes that a teacher's effectiveness in a co-teaching situation can be decomposed into two distinct components: the teacher's baseline level of individual effectiveness, $\hat{\varphi}_k$, and an interaction term, $\hat{\varphi}_m$, specific to the co-teaching arrangement m. Fourth, it assumes that each teacher in the co-teaching arrangement makes an equal contribution to the interaction effect. Finally, as with the Teacher Team method, one can distinguish between fully-interacted and aggregate teams. Specifying aggregate teams requires the additional assumption that a fixed group of teachers has a constant interaction effect on student achievement regardless of the division of teaching responsibilities.

The assumptions underlying the interaction model have some empirical support in the teaching field. In their study of cross-subject teacher co-production, Koedel and Betts (2009) find evidence that reading achievement gains are co-produced by reading and math teachers through a mechanism where both reading and math teachers each exert individual direct effects and each combination of reading and math teachers also exerts an interaction effect. Both the direct and interaction effects are found to be significant. Although not a direct application of the interaction model, in their study of teacher peer learning Jackson and Brugman (2009) also find evidence consistent with significant interaction effects.

3.5. The Full Roster Method

The Full Roster Method accommodates co-teaching by replicating each student observation by the number of teachers in the co-teaching group from whom he received instruction in a given subject and year. Each teacher in the co-teaching group is subsequently linked to one of these duplicate observations. Thus each duplicate observation is linked to exactly one teacher and each teacher is linked to all of the students she taught during the year regardless of whether or not they were co-taught with other teachers. This is the method currently used by the Washington D.C. value-added program (Isenberg and Hock 2012). The estimating equation for the Full Roster Method is:

$$y_{itk} = \boldsymbol{\eta}'\boldsymbol{z_{itk}} + \boldsymbol{\beta}'\boldsymbol{t_{itk}} + \zeta_{itk} \qquad (7)$$

Here k denotes a teacher to whom student i is linked. The $\boldsymbol{t_{itk}}$ matrix contains only one column per teacher. Each student-teacher link variable is set to one, so there are no dosage parameters in $\boldsymbol{t_{itk}}$. The dosage parameters reappear when the model is estimated with weighted regression techniques in which the weights are equal to the corresponding dosage variables. The Full Roster Method produces one estimate for each teacher, so there is no need for a composite performance metric. The estimates, $\widehat{\boldsymbol{\beta}}$, are used as the individual performance measures.

The Full Roster Method cannot be theoretically reconciled because it posits multiple education production functions for a single student simultaneously. Hock and Isenberg (2012) show that when the Full Roster method specifies fully-interacted teams and includes only teacher- and team-indicator variables as covariates, the resulting contribution point estimates are identical to those produced by the Teacher Team method, that is $\widehat{\boldsymbol{\beta}} = \widehat{\boldsymbol{\omega}}$. They also show that this property does not generally hold when either aggregate teams are specified or additional student covariates are included in the model. They present evidence that when aggregate teams are specified and student covariates are included, the Teacher Team and Full Roster methods produce very similar contribution estimates ($corr = 0.995$ in math and 0.994 in reading.) It is important to note that these correlations include estimates for both those who co-taught (21%)

who would be directly affected and those who did not co-teach (79%) who would only be indirectly affected. The models used in this study specify fully-interacted teams but include a lagged test score and a student heterogeneity term. Thus they do not meet the criteria for exact replication of the Teacher Team contribution estimates.

4. Method of Analysis

I subject the four performance metrics discussed in Section 3 ($\boldsymbol{\widehat{\psi}}$, $\boldsymbol{\widehat{\theta}}$, $\boldsymbol{\widehat{\omega}}$, and $\boldsymbol{\widehat{\beta}}$) to the task of recovering the actual contributions of co-teachers to student achievement growth under three possible DGPs. The three DGPs correspond to the assumptions underlying each of the three theoretically reconcilable performance metrics: the Partial Credit Method, the Teacher Interaction Method, and the Teacher Team Method. While I expect each metric to perform well when its own assumptions are upheld, I am interested in examining the robustness across DGPs and the biases produced when a metric's assumptions are violated.

The analysis is carried out as three separate simulations each corresponding to one of three DGPs. While the three simulations are similar, there are important differences among them. This section describes the simulations. The discussion below describes a single repetition, and each simulation consists of 2,000 repetitions of this scenario. Most components discussed below apply to all three simulations. Exceptions are discussed in the text.

4.1. Scenario

The simulation is set in a hypothetical school district in which there is a single school and in which value-added data have been collected. The value-added dataset consists of three cohorts of students who were each observed in grades four through six, and consequently each fourth through sixth grade teacher was observed teaching three separate classes of students.

There are 50 teachers in each grade four through six. Within each grade, 10 teachers (20%) are involved in co-teaching arrangements which is consistent with empirical studies (Hock and Isenberg 2012; Ruhil, Lewis, and Yandell, 2012; Watson and Thorn 2012; Watson et. al 2011). Also consistent with empirical studies (Watson et. al. 2011) the simulation distinguishes between four types of teachers:

1. **Regular teachers:** Each regular teacher teaches a self-contained class of 24 students. Regular teachers do not co-teach. There are 40 regular teachers per grade.
2. **Partner co-teachers:** Each partner co-teacher teaches one group of students individually and co-teaches another group of students with one other partner co-teacher. There are six partner co-teachers per grade yielding three partner co-teaching pairs. For one pair, each teacher teaches 17 students individually and 7 students together. In the second pair, each teacher teaches 12 students individually and 12 students are shared. In the third pair, each teacher teaches 7 students individually and 17 students are shared. Seven was chosen as the minimum number of students per individual or co-teaching arrangement because this is the cutoff the Washington D.C. value-added program uses as the minimum number of students per individual or team estimate (Hock and Isenberg, 2012).
3. **Mainstream co-teachers:** Each mainstream co-teacher teaches 17 students individually and shares 7 students with the intervention co-teacher (see below.) There are three mainstream co-teachers per grade.
4. **Intervention co-teachers:** Each intervention co-teacher teaches 7 students individually and shares seven students with each of the three mainstream teachers. There is one intervention co-teacher per grade. Each mainstream-intervention teacher combination is considered a separate team therefore each intervention teacher is part of three teams.

This pattern of teaching arrangements is the same for each grade four through six. With 50 teachers per grade, there are 150 teachers total. There are 1,147 students per cohort totaling 3,441 students across all three cohorts.

4.2. Data Generating Processes

The data generating process is a simplified cumulative education production function with geometric decay of time-varying inputs and persistent baseline ability.

$$y_{ipre} = \alpha_i + v_{ipre} \tag{8a}$$

$$y_{i4} = \lambda\left(\varepsilon_{ipre}\right) + \boldsymbol{\delta_4}'\boldsymbol{q_{i4}} + \alpha_i + v_{i4} \tag{8b}$$

$$y_{i5} = \lambda^2\left(\varepsilon_{ipre}\right) + \lambda(\boldsymbol{\delta_4}'\boldsymbol{q_{i4}} + \varepsilon_{i4}) + \boldsymbol{\delta_5}'\boldsymbol{q_{i5}} + \alpha_i + v_{i5} \tag{8c}$$

$$y_{i6} = \lambda^3\left(\varepsilon_{ipre}\right) + \lambda^2(\boldsymbol{\delta_4}'\boldsymbol{q_{i4}} + \varepsilon_{i4}) + \lambda(\boldsymbol{\delta_4}'\boldsymbol{q_{i5}} + \varepsilon_{i5}) + \boldsymbol{\delta_6}'\boldsymbol{q_{i6}} + \alpha_i + v_{i6} \tag{8d}$$

Here, y_{it} represents student i's test score in grade t, and y_{ipre} can be thought of as a pretest score measured in the year prior to fourth grade. The term α_i represents time-constant ability, and v_{it} could be idiosyncratic factors that affect learning or a combination of idiosyncratic factors and measurement error (Jacob, Lefgren, and Sims 2010). The parameter λ reflects the decay rate of time-varying inputs, which, in this context, include teacher effects and idiosyncratic factors. The term $\boldsymbol{\delta_t}'\boldsymbol{q_{it}}$ represents the student teacher linkage matrix and associated parameter vector for teachers and/or teams of grade t. In the first simulation this will be replaced by the Partial Credit linkage matrix and parameter vector $(\boldsymbol{\psi}'\boldsymbol{w_{it}})$. In the second simulation it will be replaced by the Teacher Interaction linkage matrix and parameter vector $(\boldsymbol{\varphi}'\boldsymbol{v_{it}})$, and in the third, it will be replaced by the Teacher Team linkage matrix and parameter vector $(\boldsymbol{\gamma}'\boldsymbol{c_{it}})$.

For simplicity, teachers and idiosyncratic factors are the only time-varying inputs. While this is a gross simplification, this DGP is consistent with major empirical realities. First, with geometric decay of teacher effects and perfect persistence of baseline ability, α_i, it is consistent with recent research demonstrating both the fade out of teacher impacts over time and nearly perfect persistence of certain long-term knowledge (Andrabi et. al. 2011; Jacob, Lefgren, and Sims 2010; Kane and Staiger 2008; Rothstein 2010). Second, the DGP reflects heterogeneity in achievement levels but no heterogeneity in the *rate* of learning. This is consistent with research that finds significant explanatory power of student fixed effects in models of student achievement but insignificant explanatory power of student fixed effects in models of test score gains (Kane and Staiger 2008, McCaffrey et. al 2009, Rothstein 2010).

I set the decay parameter to 0.4 which is consistent with the range (0.2 - 0.5) of recent empirical estimates of the one-year persistence rate of teacher effects. (Andrabi et. al. 2011; Jacob, Lefgren, and Sims 2010, Kane and Staiger 2008, Rothstein 2010). The individual teacher effects are drawn from a normal distribution of mean zero and a standard deviation of 0.1; the α_i's are drawn from a normal distribution of mean zero and a standard deviation of 0.8; the υ_{it}'s are drawn from a normal distribution of mean of zero and a standard deviation of 0.5. These standard deviations were chosen for several reasons. First, these values give y_{it} a mean of zero and a standard deviation of approximately one to reflect the practice of normalizing student test scores to have a mean of zero and a standard deviation of one. This makes the teacher effect standard deviation (0.1) one tenth of the student achievement standard deviation which is a robust empirical finding (Aaronson, Barrow and Sanders 2007; Jacob and Lefgren 2008; Nye, Konstantopoulos, and Hedges 2004; Rockoff 2004). I chose the standard deviation of the error term so that when a teacher solo-teaches 24 students, the precision of the teacher effect estimates produced by my simulation is consistent with the level of precision typically found in empirical analyses (Ballou, Sanders, and Wright 2004; McCaffrey et. al. 2009).

Dosage allocations between teachers are constant across students within a team[1]. Dosage parameters are random draws from 0.1 to 0.9 at intervals of 0.1. They are drawn at the beginning of each repetition and remain fixed within teams over the three hypothetical observation years. Intervention teachers have the same dosage allocation across all teams to which they belong.

In the second simulation the interaction effects are drawn from a normal distribution with a mean of zero and a standard deviation of 0.05. This is toward the high end of the empirical effect sizes estimated by Koedel and Betts (2009) and Jackson and Brugman (2009) because in most co-teaching arrangements teachers interact directly whereas their estimates come from environments where teachers interacted indirectly.

Simulating team effects for the third simulation poses several challenges. First, there is relatively little empirical evidence to guide decision making. Second, the Teacher Team Method assumes equal contributions to the team effect but it provides no insight about how the overall team effect is generated. For example the model says nothing about whether two teachers with above average individual effects are or are not likely to have a higher team effect than that of two teachers with below average individual effects. Mas and Moretti's (2009) study of berry pickers suggests that on average, individual productivity converges to average group productivity. Building on this finding, I draw the team effect from a normal distribution with a mean of zero and a standard deviation of 0.1 for which the correlation coefficient between the team effect and the average within-team individual effects is 0.5.

4.3. Estimation

Lag score value-added models are used to estimate teacher effects. When a lag score model with the correct student-teacher linkage matrix *and* with the student ability parameter, α_i,

[1] Therefore, my simulation does not distinguish between fully-interacted and aggregate teams.

included as a covariate is overlaid on the above DGP, the resulting estimation equation for (8b), (8c), and (8d) above can be written as follows:

$$y_{it} = \lambda y_{it-1} + \boldsymbol{\delta}'\boldsymbol{q_{it}} + (1-\lambda)\alpha_i + \varepsilon_{it} \qquad (9)$$

Several observations are worth noting. First, even though the student heterogeneity term in the DGP only influences a student's achievement level and does not influence their learning *rate*, the student heterogeneity term still appears with a non-zero coefficient in the above lag score estimation equation. This is driven by the different persistence rates for ability and for all other time-varying impacts inherent in the DGP. Specifically, ability is perfectly persistent while time varying impacts decay geometrically at the constant rate, λ. This demonstrates that without specific knowledge of the underlying DGP, a constant student term in a lag score model cannot necessarily be interpreted as heterogeneity in the learning *rate*.

Second, when α_i is included as a covariate, student achievement heterogeneity is completely controlled for, and one can expect the estimated coefficient on the lagged test score to recover the structural decay parameter. In reality it is often impossible to completely control for ability. The extreme case is when no student covariates above and beyond the lagged test score are included. When no attempt is made to control for ability in the context of the above DGP, the correlation between ability and the lagged test score will cause the estimate of the coefficient on the lagged test score to be biased upward from its underlying structural value. This poses a problem if the goal is to estimate the persistence parameter itself, but is less of a problem if the goal is to estimate the immediate impact of contemporaneous inputs such as teachers. In fact, a major rationale for the lag score model is that the lag score proxies for unobserved, time-constant, student heterogeneity.

If contemporaneous teacher assignments are not correlated with unobserved ability, unbiased estimates of teacher effects can be obtained from a lag score model with no additional controls for student ability, and the fact that the lagged test score acts as a proxy for ability will improve the precision of the estimates. The danger of imposing a lag score model in the context of a DGP with persistent student ability but not controlling for ability above and beyond the lag-score proxy is if teacher assignment is correlated with time-constant heterogeneity in achievement levels. Because this simulation is a controlled experiment, I impose random assignment of students to teachers. However, this is a serious concern for applications of the lag score model in non-experimental settings considering that time-constant student heterogeneity in achievement *levels* is a well-established empirical reality.

Equation (9) represents four different sets of estimation equations as $\boldsymbol{q_{it}}$ can be replaced with $\boldsymbol{w_{it}}$, $\boldsymbol{v_{it}}$, $\boldsymbol{c_{it}}$, and $\boldsymbol{t_{itk}}$ respectively. All four sets of estimation equations are applied separately to each of the three DGPs. Let the superscript on an estimator denote the process that generated the data used for estimation. For example $\widehat{\boldsymbol{\psi}}^{PC}$ indicates the Partial Credit contribution estimates when the DGP was the Partial Credit Method, and $\widehat{\boldsymbol{\psi}}^{TI}$ indicates the Partial Credit contribution estimates when the DGP was the Teacher Interaction Method.

In an attempt to maintain consistency with practical applications of value-added models, each model estimates separate individual*year, interaction* year, and/ or team*year effects. Where necessary, these single-year estimates are used to calculate the appropriate single-year individual performance metric. The single-year individual performance metrics are then averaged across the three years. These three-year averages are the estimates I compare with the true underlying teacher contributions[2].

[2] This is roughly the method proposed the Ohio Teacher Evaluation System (Ohio Department of Education 2012).

4.4. Evaluating Statistical Significance

The main output from each simulation is a dataset in which the unit of observation is a simulated teacher. For each teacher, I observe the true individual effect, the true contribution, and the estimated contributions produced by each of the four methods. Each simulation consists of 2,000 repetitions, and each repetition simulates150 teachers (3 grades*50 teachers per grade.) The final teacher dataset from each simulation has 300,000 teacher-level observations.

I use the teacher datasets to conduct the analysis of bias, and I analyze the data from each simulation separately. In order to assess the significance of the results, I treat each teacher dataset as a representative sample of teachers who have received value-added estimates, and I compute bootstrapped standard errors for all reported statistics by redrawing with replacement from the teacher dataset. Each standard error is based on 5,000 bootstrap samples. Boot strapped test statistics are computed using the bootstrapped standard error. In some cases bootstrapped confidence intervals are also reported.

4.5. Summary of Methods of Analysis

Each simulation generates data according to one of three DGPs: the Partial Credit Method, the Teacher Interaction Method, or the Teacher Team Method. Within each simulation, teacher contributions are estimated via the Partial Credit Method, the Teacher Interaction Method, the Teacher Team Method, and the Full Roster Method. Each simulation consists of 2,000 repetitions of the above scenario. *All* parameters, including the $\boldsymbol{\psi}$, $\boldsymbol{\varphi}$, and $\boldsymbol{\gamma}$ vectors, are redrawn in each repetition. In the first simulation the DGP is the Partial Credit Method, and the performance estimates $\widehat{\boldsymbol{\psi}}^{PC}$, $\widehat{\boldsymbol{\theta}}^{PC}$, $\widehat{\boldsymbol{\omega}}^{PC}$, and $\widehat{\boldsymbol{\beta}}^{PC}$ are calculated for 300,000 teachers. In the second simulation the DGP is the Teacher Interaction Method, and the performance estimates $\widehat{\boldsymbol{\psi}}^{TI}$, $\widehat{\boldsymbol{\theta}}^{TI}$, $\widehat{\boldsymbol{\omega}}^{TI}$, and $\widehat{\boldsymbol{\beta}}^{TI}$ are calculated for 300,000 teachers. In the third simulation the DGP is the Teacher Team Method and the performance estimates $\widehat{\boldsymbol{\psi}}^{TT}$, $\widehat{\boldsymbol{\theta}}^{TT}$, $\widehat{\boldsymbol{\omega}}^{TT}$, and $\widehat{\boldsymbol{\beta}}^{TT}$ are calculated for 300,000 teachers.

The main results presented in this paper come from estimates of equation (9) in which the student ability parameter, α_i, is included as a covariate. I do this to isolate the effects of differences between co-teacher incorporation methods from the effects of other misspecifications. Appendix B presents results from models estimated without including the student ability parameter[3]. Appendix C contains results from estimates where student ability is included but estimates are shrunk using an Empirical Bayes technique. These adjustments do not change the qualitative nature of the main results presented here. Appendices are available from the author upon request.

5. Results

The simulation results reveal several main findings. Under all DGPs the Partial Credit and Teacher Interaction metrics perform similarly to each other, and the Teacher Team and Full Roster metrics perform similarly to each other. Between these two groups of metrics, there are substantial differences. In many cases applying a model that is inconsistent with the DGP leads to substantial biases in co-teacher contribution estimates, but the degree to which these biases distort relative teacher performance rankings varies. When the percentage of shared student is small to moderate, the Partial Credit and Teacher Interaction metrics preserve teacher rankings more robustly across DGPs than the Teacher Team and Full Roster metrics.

Table 1 presents several descriptive statistics to gauge the general performance of the four estimation methods across the three DGPs. From this broad-brush overview, all four

[3] Regarding the choice of error term standard deviation (0.5) discussed in Section 4.2, DGPs including or excluding the student heterogeneity term makes only a small difference in estimate precision. For example, when the Partial Credit estimation method is used on data generated by the Partial Credit method, the average standard error for a regular teacher estimate is 0.103 when the ability term is included and 0.114 when it is excluded. These both fall in the empirical ranges discussed in Ballou, Sanders, and Wright (2004) and McCaffrey et. al.(2009).

estimation techniques appear to perform very similarly. This is not surprising considering that a majority of the teachers in the sample are not co-teachers and are not directly affected by the different estimation strategies. As discussed later in this section, there are significant differences between estimation methods conditional on DGP and teacher characteristics, but a broad overview helps to put these differences in context.

Several summary statistics are used to evaluate the general performance of the four estimation methods under each DGP. First is the average within grade rank correlation between a teacher's true contribution and their estimated contribution. Here, the estimated contribution always refers to the three year average estimated contribution discussed in the methods section. The true contribution refers to the underlying parameter values $\boldsymbol{\psi}$, $\boldsymbol{\theta}$, and $\boldsymbol{\omega}$ where $\boldsymbol{\psi}$ is a vector of random draws as discussed in the methods section and $\boldsymbol{\theta}$, and $\boldsymbol{\omega}$ are calculated from vectors of random draws according to $(6b)$ and $(4b)$ respectively. Rank always refers to within-grade rank. With three grades and 2,000 repetitions the average correlations represent the average of 6,000 rank correlations. The value is nearly identical across all estimation methods and DGPs with values ranging from 0.824-0.831.

The second summary statistic is the absolute value of the median change in within grade rank. Specifically this is the absolute value of the difference between the within grade rank of the estimated contribution and the within grade rank of the true contribution. This measure indicates the precision with which contribution estimates preserve rankings. Recall that there are 50 teachers per grade thus the maximum value is 49. The median change is 5 across all estimation methods and DGPs.

The third summary statistic is the median *net* change in within grade rank. This gives a sense of the extent to which any changes in rankings are likely to be in one direction or the other, indicating bias. Overall (i.e. without conditioning on teacher characteristics) the median change is zero across all estimation methods and DGPs.

The fourth summary statistic is the percentage of teachers whose true within grade contribution rank is in the bottom 50% but whose contribution estimate incorrectly ranks them in the top 50%. Again, unconditional on teacher characteristics, this percentage is nearly identical across estimation methods and DGPs with a value of about 18%.

The final summary statistic is the percentage of teachers whose true within grade contribution rank is in the bottom 20% but whose estimated within grade contribution rank is outside the bottom 20% (i.e. in the top 80%.) Again, without conditioning on teacher characteristics, this percentage is nearly constant across estimation methods and DGPs with a value of about 31.5%.

5.1. Results from the Partial Credit DGP

In order to investigate the ability of the four performance metrics to accurately estimate co-teacher contributions under the Partial Credit DGP, I first focus on the case of the partner co-teachers whose situation is most straightforward.

Table 2 summarizes information on the average net difference between the estimated and the true contributions of partner co-teachers conditional on the within grade quintile of the teacher's individual effect and a variety of other conditioning variables[4]. These additional conditioning variables include the within grade quintile of their partner's individual effect, the number of students co-taught, and the percentage of responsibility the teacher had for the co-

[4] Under the Partial Credit DGP the true contribution and the true individual effect are always exactly the same. This is not generally true for the Teacher Interaction and Teacher Team DGPs. I distinguish between contributions and individual effects here to maintain consistency with discussions of the Teacher Interaction and Teacher Team DGPs later on.

taught students. Quintile 1 contains teachers with individual effects in the lowest 20%, and quintile 5 contains those with individual effects in the highest 20%.

To summarize results efficiently, Table 2 presents information on the direction, magnitude, and statistical significance of cell-specific average differences. Individual point estimates and standard errors corresponding to Table 2, as well as Tables 4 and 6 discussed later, are reported in Appendix A and are available from the author upon request. In Table 2 point estimates are color-coded at intervals of 0.01 (10% of a teacher effect standard deviation.) Statistically significant results are indicated with one or more stars and are inferred from bootstrapped test statistics based on 5,000 resamples. Resampling was conducted separately for each cell. Cell sizes are approximately 7,200 for comparisons based on the quintile of individual effects only, 1,400 for the quintile/partner quintile individual effect comparisons, and 2,400 for all other comparisons.

The Partial Credit and Teacher Interaction methods show nearly no signs of bias with net differences in all cells being less than or equal to 0.005 (5% of a teacher effect standard deviation) and the large majority of average differences not statistically different from zero. Note that with repeated t-tests, false positives are expected 5% of the time.

The results for the Teacher Team and Full Roster methods display evidence of statistically significant bias when teachers with high and low individual effects are paired together. In such cases, the contribution estimates of teachers with low individual effects are biased upward and those of teaches with high individual effects are biased downward. The average bias exceeds 30% of a teacher effect standard deviation when teachers at extreme quintiles are paired together. Table 2 also shows that this bias increases as the number of shared students increases and appears to reach a peak when responsibility is shared relatively equally.

The biases in the Teacher Team metric make sense when one considers that this method assumes equal contributions to the team, and this assumption is violated by the Partial Credit DGP. When the Teacher Team estimation method is imposed over the Partial Credit DGP, teachers with high individual effects who are paired with teachers with low individual effects do not receive credit for their disproportionate contributions to the team, and those with low individual effects receive more credit than they deserve for their less than proportionate contributions. The nearly identical pattern of biases in the Full Roster Metric is not surprising given that this metric is known to behave similarly to the Teacher Team metric. The absence of bias in the Teacher Interaction method also makes sense as one would expect the erroneous interaction parameters to add imprecision to the estimation process, but not bias. Although results are presented only for partner co-teachers, the same mechanisms are at work for all types of co-teachers (partner, mainstream, and interaction) but effect sizes may differ among them.

An important question is whether these biases are large enough to make a difference in relative teacher contribution rankings. Table 3 considers two summary statistics to provide evidence on this question. The first is the median net rank change of teachers with individual effects in the bottom 20% whose co-teachers have individual effects in the top 20%. Because intervention teachers co-teach with multiple teachers, in this case the second conditioning factor is that the median rank among their three partners is in top 50% of individual effect rankings. The second summary statistic is the percentage of teachers with individual effects in the bottom 20% who are misclassified in the top 80% again conditioning on the individual effect of the co-teacher as above. Thus both statistics examine the case in which teachers at opposite ends of the individual effect distribution co-teach together.

95% bootstrapped percentile confidence intervals are given in brackets and 95% bootstrapped bias-corrected and accelerated (BCa) confidence intervals are given in parentheses[5]. A star indicates a confidence interval that does not overlap with the confidence interval of the corresponding statistic associated with the Partial Credit metric whose estimates are unbiased under the Partial Credit DGP.

For both statistics, the results for the Teacher Interaction metric are not statistically different from the results for the Partial Credit metric. However, the results for the Teacher Team and Full Roster metrics show statistically significant positive increases in both net rank change and the misclassification percentage for partner and intervention co-teachers. This reflects the positive bias in the estimated contributions of teachers with low individual effects when they are paired with teachers whose individual effects are high. These observed increases are also of practical significance as the gap between the Partial Credit misclassification confidence band and the Teacher Team and Full Roster misclassification confidence bands is around 10% for both partner and intervention co-teachers. Partner and intervention teachers with individual effects in the bottom quintile who share students with teachers whose individual effects are in the top quintile are at least 10% more likely to be misclassified in the top 80% under the Teacher Team and Full Roster methods than under the Partial Credit or Teacher Interaction Methods. These same teachers are likely to be ranked 4-5positions (out of 50) higher than their true contribution rank under the Teacher Team and Full Roster Methods.

The Teacher Team and Full Roster estimates for the mainstream teachers do not show evidence of statistically significant bias. This does not necessarily mean an absence of bias, but

[5] Future revisions of this paper will use 10,000 bootstrap replicates to compute bootstrapped confidence intervals. BCa confidence intervals could not be calculated for median net rank change due to the small range of bootstrapped median values. Given that percentile confidence intervals are generally biased, I am investigating alternative evaluation methods.

rather that any bias that may be present is relatively small and is not detectable at current levels of statistical power. The absence or low level of bias here is likely due to the small number of shared students taught by mainstream co-teachers. Finally, it is worth noting that although these results focus on the bottom of the distribution, results not presented here show generally equal and opposite patterns at the top of the distribution.

5.2. Results from the Teacher Interaction DGP

Results from the Teacher Interaction DGP are very similar to the results for the Partial Credit DGP. Table 4 summarizes information on conditional average differences between estimated and true contributions as was done for the Partial Credit DGP. In the context of the Teacher Interaction DGP it is important to reiterate that the true contribution will, in general, differ from the true individual effect. The information in Table 4 refers to the average differences between estimated and true *contributions* conditional on *individual* effect quintiles. I condition on individual effects because they are exogenous as opposed to contributions which are endogenous. The pattern of biases is very similar to that which was observed under the Partial Credit DGP. In results not presented here, I also calculate average contribution differences conditional on the quintile of the true interaction effect, but this does not appear to be a source of bias in any of the estimation methods. These results are not trivial. They suggest that the simpler Partial Credit estimation method is quite robust to the more complicated and, given the results of recent research, perhaps more realistic Teacher Interaction DGP. Excluding interaction effects from a regression equation will generally bias estimates of *individual* direct effects. However, it appears that the biased estimates of the individual effect are relatively unbiased estimates of the *contribution* parameters of interest.

Although Table 3 provided evidence on the degree to which bias of the observed magnitude can influence teacher effect rankings, Table 5 presents evidence specific to the Teacher Interaction DGP simulation. The pattern of median net rank changes across estimation methods and conditioning factors is very similar to that observed under the Partial Credit DGP. To clarify, here median net rank change is defined exactly as it has been previously: estimated *contribution* rank – true *contribution* rank. The conditioning factors refer to the quintile of the *individual* effect rank. Because under the Teacher Interaction DGP the true contribution is not the same as the true individual effect, I do not compute conditional misclassification percentages.

5.3. Results from the Teacher Team DGP

Table 6 presents information on average differences for the four performance metrics under the Teacher Team DGP. As with Table 4, the differences represent the average difference between estimated and true *contributions* and the quintiles represent the quintile of the *individual* effect. The data show a number of trends opposing those exhibited in Tables 2 and 4: the Teacher Team and Full Roster methods show nearly no evidence of bias and both the Partial Credit and Teacher Interaction methods show evidence of bias. Biases are most severe when teachers with the highest and the lowest individual effects are paired together, but here the bias is downward for teachers with low individual effects and upward for teachers with high individual effects. As with the Partial Credit and Teacher Interaction DGPs these biases are exaggerated as the number of co-taught students increases.

This pattern of biases makes sense. The Teacher Team DGP generates team effects such that team contributions tend toward the average individual effect. Thus teachers who are less effective individually are likely to be more effective in a team and vice versa. Because the Partial Credit and Teacher Interaction methods assume unequal contributions to the team, teachers with low individual effects do not receive full credit for their contributions in the team and teachers with high individual effects receive more credit than they deserve for their contributions in the team. This results in downwardly biased contribution estimates for teachers with low individual effects and upwardly biased estimates for teachers with high individual effects. It appears that the primary driver of bias in all scenarios is the inconsistency between the proportions in which teachers actually contribute to the team and the assumptions regarding contribution proportions imbedded in the modeling method.

Table 7 examines the degree to which these biases change estimated contribution rankings. Because these biases work in the opposite direction of those examined previously, they may not have effects of the same magnitude. Specifically, if the greatest bias falls on those in the tails of the *contribution* distribution, then the bias is reinforcing and ability of the bias to change rankings is limited. Therefore, it depends on the degree to which those in the tails of the individual effect distribution are also those in the tails of the contribution distribution.

Table 7 presents evidence that, in the case of this simulation, the biases generated by the Partial Credit and Teacher Interaction methods under the Teacher Team DGP have more muted effects on teacher contribution rankings than the biases examined previously. The table displays separate median net rank changes for teachers in the first and second quintiles. I include the second quintile because of the downward nature of the bias at the bottom of the distribution. Although there are statistically significant differences in the median net rank changes of partner and intervention teachers between the Teacher Team method and the Partial Credit and Teacher Interaction methods, these biases have smaller impacts on rank movement than those observed previously. Under the Partial Credit and Teacher Interaction methods a net rank movement of -3 is the most extreme end of any of the reported confidence intervals. This is in contrast to confidence intervals with a net rank change lower bound of 4 observed in Tables 3 and 5.

While these results seem to support the relative robustness of the Partial Credit and Teacher Interaction metrics across DGPs, I hesitate to draw this conclusion unconditionally. Biases of this nature could lead to more severe rank changes under different conditions. Specifically, as the percentage of co-taught students approaches 100%, these biases will begin to dominate teacher contribution estimates. Therefore, my results suggest that the Partial Credit and Teacher Interaction methods are moderately robust to the Teacher Team DGP *when the percentage of co-taught students is small to moderate.*

6. Discussion and Conclusions

The simulation results reveal two main findings. First, applying a model that is inconsistent with the DGP can lead to substantial biases in co-teacher contribution estimates. Second, the degree to which these biases distort relative teacher performance rankings varies. Under the Partial Credit and Teacher Interaction DGPs, the Teacher Team and Full Roster metrics induce relatively large distortions in teacher contribution rankings. In contrast, when the percentage of shared students is small to moderate the Partial Credit and Teacher Interaction metrics are moderately robust to alternative DGPs.

Further complicating matters, Hock and Isenberg (2012) note that, in practice, the percentage of shared students often is not small to moderate; rather teachers involved in co-teaching relationships tend to teach many students together but few outside the team. Not only does this increase the severity of bias, but it also induces a high degree of multicollinearity between teacher variables which leads to unstable contribution estimates under the Partial Credit and Teacher Interaction methods. This instability renders these methods undesirable in such situations.

It is important to state and accept the fact that the multicollinearity problem cannot be solved. For teachers in certain co-teaching situations it is statistically impossible to separate the effects of one teacher from another even if these effects are separable in theory. The problem is similar to cases in which a teacher teaches only a small number of students. However, whereas in the case of small sample sizes, such teachers are typically identified and dropped from the model, in the case of multicollinearity researchers and practitioners have developed methods, such as the Teacher Team and the Full Roster methods, to retain these teachers in the model and generate individual value-added performance measures for them. Yet this solution creates its own set of problems. The validity of these individual performance metrics rely on assumptions about the mechanisms underlying the co-teaching process which may not be upheld. As the simulations results show, the Teacher Team method can induce substantial biases if the assumption of equal contributions is violated.

I conclude that the limits of VAMs to produce accurate individual performance measures for co-teachers must be acknowledged and accepted. Given the limited feasibility of the Partial Credit and Teacher Interaction Methods and the lack of knowledge about the true nature of co-teaching processes, the use of value-added performance measures for co-teachers in high stakes settings may not be appropriate.

References

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

Andrabi, Tahir, Jishnu Das, Asim Ijaz Khwaja, and Tristan Zajonc. 2011. Do Value-Added Estimates Add Value? Accounting for Learning Dynamics. *American Economic Journal: Applied Economics* 3(3):29-54.

Baker, Eva L., Paul E. Barton, Linda Darling-Hammond, Edward Haertel, Helen F. Ladd, Robert L. Linn, Diane Ravitch, Richard Rothstein, Richard J. Shavelson, and Lorrie A. Shepard. 2010. *Problems with the Use of Student Test Scores to Evaluate Teachers.* Washington, D.C.: Economic Policy Institute. Available at: www.epi.org/publication/bp278/ (accessed Apr. 21, 2012.)

Ballou, Dale, William Sanders, and Paul Wright. 2004. Controlling for Student Background in Value-Added Assessment of Teachers. *Journal of Educational and Behavioral Statistics* 29(1):37-65.

Been, Vicki, Ingrid Gould Ellen, Amy Ellen Schwartz, Leanna Stiefel, and Meryle Weinstein. 2011. Does Losing Your Home Mean Losing Your School?: Effects of Foreclosures on the School Mobility of Children. *Regional Science and Urban Economics.* 41(4):707-414.

Corcoran, Sean P. 2010. *Can Teachers be Evaluated by their Students' Test Scores? Should They Be? The Use of Value-Added Measures of Teacher Effectiveness in Policy and Practice.* Providence, RI: Annenberg Institute for School Reform. Available at: http://annenberginstitute.org/pdf/valueaddedreport.pdf (accessed Apr. 21, 2012.)

Goe, Laura. 2008. Key Issue: Using Value-Added Models to Identify and Support Highly Effective Teachers. Washington D.C.: National Comprehensive Center for Teacher Quality. Available at: www2.tqsource.org/strategies/het/UsingValueAddedModels.pdf (accessed Sept. 22, 2012.)

Goldhaber, Dan, Pete Goldschmidt, Philip Sylling, Fannie Tseng. 2011. Teacher Value-Added at the High School Level: Different Models, Different Answers? Working Paper, Center for Education Data and Research, Seattle, WA.

Guarino, Cassandra M., Mark. D. Reckase, and Jeffery M. Wooldridge. 2012. Can Value Added Measures of Teacher Performance Be Trusted? Working Paper, Michigan State University, East Lancing, MI.

Hanushek, Eric A., John F. Kain, and Steven G. Rivkin. 2004. Disruption versus Tiebout Improvement: the Costs and Benefits of Switching Schools. *Journal of Public Economics.* 88(9-10):1721-1746.

Hock, Heinrich and Eric Isenberg. 2012. Methods for Accounting for Co-Teaching in Value

Added Models. Working Paper, Mathematica Policy Research, Princeton, NJ.

Isenberg, Eric and Heinrich Hock. 2010. Measuring School and Teacher Value Added for IMPACT and TEAM in DC Public Schools. Washington, D.C.: Mathematica Policy Research. Available at: http://mathematica-mpr.com/publications/redirect_PubsDB.asp?strSite=PDFs/education/valueadded_techrprt.pdf (accessed Sept. 22, 2012.)

Isenberg, Eric and Heinrich Hock. 2011. Design of Value-Added Models for IMPACT and TEAM in DC Public Schools. Washington, D.C.: Mathematica Policy Research. Available at: http://mathematica-mpr.com/publications/redirect_PubsDB.asp?strSite=PDFs/education/valueadded_models.pdf (accessed Sept. 22, 2012.)

Jackson, C. Kirabo and Elias Brugman. 2009. Teaching Students and Teaching Each Other: The Importance of Peer Learning for Teachers. *American Economic Journal: Applied Economics* 1(4):85-108.

Jacob, Brian A., Lars Lefgren, and David P. Sims. 2010. The Persistence of Teacher-Induced Learning. *Journal of Human Resources* 45(4):915-943.

Jacob, Brian A. and Lars Lefgren. 2008. Can Principals Identify Effective Teachers? Evidence on Subjective Performance Evaluation in Education. *Journal of Labor Economics* 26(1):101-136.

Kane, Thomas J., and Douglas O. Staiger. 2008. Estimating Teacher Impacts on Student Achievement: An Experimental Evaluation. Working paper, NBER, no. 14607.

Kerbow, David, Carlos Azcoitia, and Barbara Buell. 2003. Student Mobility and Local School Improvement in Chicago. *Journal of Negro Education* 72(1):158-164.

Koedel, Cory. 2009. An Empirical Analysis of Teacher Spillover Effects in Secondary School. *Economics of Education Review* 28(6):682-692.

Mas, Alexandre and Enricho Moretti. 2009. Peers at Work. *American Economic Review* 99(1):112-145.

Mihaly, Kata, Daniel F. McCaffrey, J.R. Lockwood, and Tim R. Sass. 2010. Centering and Reference Groups for Estimates of Fixed Effects Modeling. Modifications to felsdvreg. http://www.ode.state.oh.us/GD/Templates/Pages/ODE/ODEDetail.aspx?page=3&TopicRelationID=1230&ContentID=125739The Stata Journal 10(1):82-103.

McCaffrey, Daniel F., Tim R. Sass, J.R. Lockwood, and Kata Mihaly. 2009. The Intertemporal Variability of Teacher Effect Estimates. *Education Finance and Policy* 4(4):572-606.

Nye, Barbara, Spyros Konstantopoulos, and Larry V. Hedges. 2004. How Large Are Teacher Effects? *Education Evaluation and Policy Analysis.* 26(3):237-257.

Ohio Department of Education. 2012. Ohio Teacher Evaluation System. Columbus, OH. Available at: http://www.ode.state.oh.us/GD/Templates/Pages/ODE/ODEDetail.aspx?page=3&TopicRelationID=1230&ContentID=125739 (accessed Feb. 26, 2012).

Ponisciak, Stephen, Kaveh Akram, Michael McCants, Frank Erickson, and Robert Meyer. 2012. The Effects of Student-Teacher Linkage Data Errors on Value-Added Results. Working Paper, Value-Added Research Center, University of Wisconsin, Madison, WI.

Rockoff, Jonah E. 2004. The Impact of Individual Teachers in Students Achievement: Evidence from Panel Data. *American Economic Review: Papers and Proceedings* 94(2):247-252.

Rothstein, Jesse. 2009. Student Sorting and Bias in Value-Added Estimation: Selection on Observables and Unobservables. *Education Finance and Policy* 4(4):537-571.

Rothstein, Jesse. 2010. Teacher Quality in Educational Production: Tracking, Decay, and Student Achievement. *Quarterly Journal of Economics.* 125(1):175-214.

Ruhil, Ani, Marsha Lewis, and Nicole Yandell. 2012. Value-Added Assessment and Roster Verification Evaluation. Presentation at Ohio Education Research Center Conference, 2012, Columbus, OH.

Steele, Jennifer L., Laura S. Hamilton, and Brian M. Stecher. 2010. Incorporating Student Performance Measures into Teacher Evaluation Systems. Santa Monica, CA: RAND Corporation. Available at: http://www.rand.org/pubs/technical_reports/TR917.html (accessed April 21, 2012).

Todd, Petra E. and Kenneth I. Wolpin. 2003. On the Specification and Estimation of the Production Function for Cognitive Achievement. *The Economic Journal* 113(485):F3-F33.

Value-Added Research Center. 2010. NYC Teacher Data Initiative: Technical Report of the NYC Value-Added Model. Madison, WI. Available at: schools.nyc.gov/NR/.../TDINYC TechnicalReportFinal072010.pdf (accessed Sept. 22, 2012.)

Watson, Jeff and Chris Thorn. 2012. Student Teacher Linkage Quality: Discrepancy Rates and Team Teaching. Working Paper, Wisconsin Center for Education Research, University of Wisconsin, Madison, WI.

Watson, Jeffery, Christopher Thorn, Steve Ponisciak, and Fred Boehm. 2011. Measuring the Impact of Team Teaching on Student-Teacher Linkage Data. Working Paper, Value-Added Research Center, University of Wisconsin, Madison, WI.

Wright, Paul S., John T. White, William L. Sanders, and June C. Rivers. 2010. SAS EVAAS Statistical Models. Cary, NC: SAS Institute Inc. Available at: www.sas.com/resources/asset/SAS-EVAAS-Statistical-Models.pdf (accessed Sept. 22, 2012.)

Table 1. Model Performance Summary Statistics

	Sample Size	Estimation Method			
		PC	TI	TT	FR
DGP=Partial Credit Method					
Average Rank Correlation	6,000	0.827	0.826	0.827	0.827
Median Rank Change (Abs.)	300,000	5	5	5	5
Median Rank Change (Net)	300,000	0	0	0	0
P(Est. Contrib. Top 50%|True Contrib. Bottom 50%)	150,000	17.9%	17.9%	17.9%	17.9%
P(Est. Contrib. Top 80%|True Contrib. Bottom 20%)	60,000	31.7%	31.7%	31.6%	31.6%
DGP=Teacher Interaction Method					
Average Rank Correlation	6,000	0.827	0.826	0.828	0.828
Median Rank Change (Abs.)	300,000	5	5	5	5
Median Rank Change (Net)	300,000	0	0	0	0
P(Est. Contrib. Top 50%|True Contrib. Bottom 50%)	150,000	17.9%	17.9%	17.9%	17.9%
P(Est. Contrib. Top 80%|True Contrib. Bottom 20%)	60,000	31.4%	31.4%	31.4%	31.4%
DGP=Teacher Team Method					
Average Rank Correlation	6,000	0.824	0.824	0.831	0.831
Median Rank Change (Abs.)	300,000	5	5	5	5
Median Rank Change (Net)	300,000	0	0	0	0
P(Est. Contrib. Top 50%|True Contrib. Bottom 50%)	150,000	18.1%	18.1%	17.9%	17.9%
P(Est. Contrib. Top 80%|True Contrib. Bottom 20%)	60,000	32.0%	32.0%	31.3%	31.3%

Table 2. Average Differences between Estimated and True *Contributions* under the Partial Credit DGP

Quintile	Unconditional	Quintile Partner *Individual* Effect					Shared Students			Percent Responsible		
Individual												
Effect		1	2	3	4	5	7	12	17	10-30%	40-60%	70-90%
Individual Performance Metric = Partial Credit (Psi-hat)												
1	+	+	-	-	-	-	+	-	-	-	-	+
2	-	-	-	+	-	-	-	-	-	-	-*	+
3	+	+	+	-	+	-	+	+	-	-	-	+
4	-	-	+	-	+	-	-	-	+	-	-	+
5	+	+*	-	+	+	-	-	-	+*	+	-	+
Individual Performance Metric = Teacher Interaction (Theta-hat)												
1	-	+	-	-	-	-	+	-	-	-	-	+
2	-	-	-	+	-	-	-	-	-	-	-	+
3	+	+	+	-	+	-	+	+	-	-	-	+
4	-	-	+	-	+	-	-	-	+	-	-	+
5	-	+*	-	+	+	-	-	-	+*	+	-	+
Individual Performance Metric = Teacher Team (Omega-hat)												
1	+**	+	+**	+**	+**	+**	+**	+**	+**	+**	+**	+**
2	+**	-**	-	+**	+**	+**	+	+**	+**	+**	+**	+**
3	+	-**	-**	-	+**	+**	+	+	-	-	-	+
4	-**	-**	-**	-**	+	+**	-**	-**	-**	-**	-**	-**
5	-**	-**	-**	-**	-**	-	-**	+**	-**	-**	-**	-**
Individual Performance Metric = Full Roster (Beta-hat)												
1	+**	+	+**	+**	+**	+**	+**	+**	+**	+**	+**	+**
2	+**	-**	-	+**	+**	+**	+	+**	+**	+**	+**	+**
3	+	-**	-**	-	+**	+**	+	+	-	-	-	+
4	-**	-**	-**	-**	+	+**	-**	-**	-**	-**	-**	-**
5	-**	-**	-**	-**	-**	-	-**	-**	-**	-**	-**	-**

* indicates significance at the 0.05 level and ** at the 0.01 level. Significance levels are inferred from t-statistics using bootstrapped standard errors based on 5,000 resamples.

Legend	≤ -0.03	$(-0.03, -0.02]$	$(-0.02, -0.01]$	$(-0.01, 0.01)$	[0.01, 0.02)	[0.02, 0.03)	$\geq$ 0.03

Table 3. Effects of Bias under the Partial Credit DGP

	Own Individual Effect	Co-Teacher Individual Effect	Sample Size	Estimation Method			
				PC	TI	TT	FR
Type=Partner							
Med. Rank Change (Net)	Bottom 20%	Top 20%	1516	1	1	4	4
				[1, 2]	[1, 2]	[4, 5]*	[4, 5]*
P(Est. Contrib. Top 80%)	Bottom 20%	Top 20%	1516	32.7%	32.7%	47.4%	47.7%
				(30.3, 35.0)	(30.3, 35.0)	(44.7, 49.7)*	(44.7, 49.7)*
Type=Mainstream							
Med. Rank Change (Net)	Bottom 20%	Top 20%	702	1	1	2	2
				[1, 2]	[1, 2]	[2, 3]	[2, 3]
P(Est. Contrib. Top 80%)	Bottom 20%	Top 20%	702	31.2%	31.3%	37.9%	37.9%
				(27.8, 34.6)	(27.9, 34.8)	(34.2, 41.5)	(34.2, 41.5)
Type=Intervention							
Med. Rank Change (Net)	Bottom 20%	Med. Eff. Top **50%**	637	1	1	5	5
				[1, 2]	[1, 2]	[4, 5]*	[4, 5]*
P(Est. Contrib. Top 80%)	Bottom 20%	Med. Eff. Top **50%**	637	32.3%	33.1%	50.5%	50.5%
				(28.6, 35.6)	(29.4, 36.8)	(46.5, 54.2)*	(46.5, 54.2)*

95% bootstrapped percentile confidence intervals reported in brackets. 95% BCa confidence intervals reported in parentheses. Confidence intervals calculated from 5,000 resamples. * indicates a confidence interval that does not overlap with the confidence interval for the corresponding statistic associated with the Partial Credit metric.

Table 4. Average Differences between Estimated and True *Contributions* under the Teacher Interaction DGP

Quintile *Individual* Effect	Unconditional	Quintile Partner *Individual* Effect					Shared Students			Percent Responsible		
		1	2	3	4	5	7	12	17	10-30%	40-60%	70-90%
Individual Performance Metric = Partial Credit (Psi-hat)												
1	-	+	-	+	-	-	-	-	-	-	-	+
2	+	-	-	+	+	-	-	-	+	+	+	-
3	+*	+	+	+	+	+	+	+*	+	+*	+	+
4	-	+	-	-	-	+	-	+	-	+	-	+
5	-	-	-	-	-	+	-	-	-	-	-	-
Individual Performance Metric = Teacher Interaction (Theta-hat)												
1	-	+	-	+	-	-	-	-	-	-	-	+
2	+	-	-	+	+	+	-	-	+	+	+	-
3	+*	+	+	+	+	+	+	+*	+	+**	+	+
4	-	+	-	-	-	+	-	+	-	-	-	+
5	-	-	-	-	-	+	-	-	-	-	-	-
Individual Performance Metric = Teacher Team (Omega-hat)												
1	+**	+	+**	+**	+**	+**	+**	+**	+**	+**	+**	+**
2	+**	-**	-	+**	+**	+**	+*	+**	+**	+**	+**	+**
3	+*	-**	-**	+	+**	+**	+	+*	+	+**	+	+
4	-**	-**	-**	-**	-	+**	-**	-**	-**	-**	-**	-**
5	-**	-**	-**	-**	-**	+	-**	-**	-**	-**	-**	-**
Individual Performance Metric = Full Roster (Beta-hat)												
1	+**	+	+**	+**	+**	+**	+**	+**	+**	+**	+**	+**
2	+**	-**	-	+**	+**	+**	+*	+**	+**	+**	+**	+**
3	+*	-**	-**	+	+**	+**	+	+*	+	+**	+	+
4	-**	-**	-**	-**	-	+**	-**	-**	-**	-**	-**	-**
5	-**	-**	-**	-**	-**	+	-**	-**	-**	-**	-**	-**

* indicates significance at the 0.05 level and ** at the 0.01 level. Significance levels are inferred from t-statistics using bootstrapped standard errors based on 5,000 resamples.

Legend	≤ −0.03	(−0.03, −0.02]	(−0.02, −0.01]	(−0.01, 0.01)	[0.01, 0.02)	[0.02, 0.03)	≥ 0.03

Table 5. Effects of Bias under the Teacher Interaction DGP

	Own	Co-Teacher	Sample	Estimation Method			
	Individual Effect	Individual Effect	Size	PC	TI	TT	FR
Type=Partner							
Med. Rank Change (Net)	Bottom 20%	Top 20%	1,515	1	1	4	4
				[1, 2]	[1, 2]	[4, 5]*	[4, 5]*
Type=Mainstream							
Med. Rank Change (Net)	Bottom 20%	Top 20%	729	2	2	3	3
				[1, 2]	[1, 2]	[2, 3]	[2, 3]
Type=Intervention							
Med. Rank Change (Net)	Bottom 20%	Med. Eff. Top **50%**	613	1	2	5	5
				[1, 2]	[1, 2]	[5, 6]*	[5, 6]*

95% bootstrapped percentile confidence intervals reported in brackets. Confidence intervals calculated from 5,000 resamples; * indicates a confidence interval that does not overlap with the confidence interval for the corresponding statistic associated with the Teacher Interaction metric.

Table 6. Average Differences between Estimated and True *Contributions* under Teacher Team DGP

Quintile *Individual* Effect	Unconditional	Quintile Partner *Individual* Effect					Shared Students			Percent Responsible		
		1	2	3	4	5	7	12	17	10-30%	40-60%	70-90%
Individual Performance Metric = Partial Credit (Psi-hat)												
1	-**	-	-**	-**	-**	-**	-**	-**	-**	-**	-**	-**
2	-**	+**	+	-**	-**	-**	-*	-**	-**	-**	-**	-*
3	+*	+**	+**	+	-**	-**	+*	+	+	+	+	+
4	+**	+**	+**	+*	+	-**	+**	+**	+**	+**	+**	+*
5	+**	+**	+**	+**	+**	-*	+**	+**	+**	+**	+**	+**
Individual Performance Metric = Teacher Interaction (Theta-hat)												
1	-**	-	-**	-**	-**	-**	-**	-**	-**	-**	-**	-**
2	-**	+**	+	-**	-**	-**	-*	-**	-**	-**	-**	-**
3	+*	+**	+**	+	-**	-**	+	+	+	+	+	+
4	+**	+**	+**	+*	+	-**	+**	+**	+**	+**	+**	+**
5	+**	+**	+**	+**	+**	-*	+**	+**	+**	+**	+**	+**
Individual Performance Metric = Teacher Team (Omega-hat)												
1	+	-	+	-	+	-	+*	+	-	+	+	-
2	+	-	+	+	+	+	+	+	-	-	+	+
3	+*	+	+	+	-	+	+	+	+	+	+	+
4	-	+	+	-	+	-	+	+	-	+	-	-
5	-*	+	-	-	-	-*	-*	-	-	+	-*	-
Individual Performance Metric = Full Roster (Beta-hat)												
1	+	-	+	-	+	-	+*	+	-	+	+	-
2	+	-	+	+	+	+	+	+	-	-	+	+
3	+*	+	+	+	-	+	+	+	+	+	+	+
4	-	+	+	-	+	-	+	+	-	+	-	-
5	-*	+	-	-	-	-*	-**	-	-	+	-*	-

* indicates significance at the 0.05 level and ** at the 0.01 level. Significance levels are inferred from t-statistics using bootstrapped standard errors based on 5,000 resamples.

Legend	≤ −0.03	(−0.03, −0.02]	(−0.02, −0.01]	(−0.01, 0.01)	[0.01, 0.02)	[0.02, 0.03)	≥ 0.03

Table 7. Effects of Bias under the Teacher Team DGP

	Own	Co-Teacher	Sample	Estimation Method			
	Individual Effect	Individual Effect	Size	PC	TI	TT	FR
Type=Partner							
Med. Rank Change (Net)	Bottom 20%	Top 20%	1,434	-0.5	-1	2	2
				[-1, 0]*	[-1, 0]*	[1, 2]	[1, 2]
	Bottom 20-40%	Top 20%	1,523	-2	-2	1	1
				[-2, -1]*	[-3, -1]*	[0, 2]	[1, 2]
Type=Mainstream							
Med. Rank Change (Net)	Bottom 20%	Top 20%	810	1	0	1	1
				[0, 1]	[0, 1]	[1, 2]	[1, 2]
	Bottom 20-40%	Top 20%	686	0	0	2	2
				[-1, 1]	[-1, 1]	[1, 3]	[1, 3]
Type=Intervention							
Med. Rank Change (Net)	Bottom 20%	Med. Eff. Top **50%**	655	-2	-2	1	1
				[-2, -1]*	[-3, -2]*	[0, 2]	[0, 2]
	Bottom 20-40%	Med. Eff. Top **50%**	606	-2	-2	1	1
				[-3, -1]*	[-3, -1]*	[0, 2]	[0, 2]

95% bootstrapped percentile confidence intervals reported in brackets. Confidence intervals calculated from 5,000 resamples; * indicates a confidence interval that does not overlap with the confidence interval for the corresponding statistic associated with the Teacher Team metric.