Statistics, Politics, and Policy

Volume 3, Issue 2 2012 Article 4

Teacher Effect Estimates and Decision Rules for Establishing Student-Teacher Linkages: What are the Implications for High-Stakes Personnel Policies in an Urban School District?

Bing Han, RAND Corporation
Daniel McCaffrey, RAND Corporation
Matthew G. Springer, Vanderbilt University
Michael Gottfried, RAND Corporation and Loyola Marymount University

Recommended Citation:
DOI: 10.1515/2151-7509.1035

©2012 De Gruyter. All rights reserved.
Teacher Effect Estimates and Decision Rules for Establishing Student-Teacher Linkages: What are the Implications for High-Stakes Personnel Policies in an Urban School District?

Bing Han, Daniel McCaffrey, Matthew G. Springer, and Michael Gottfried

Abstract

Both research and practice of value-added models (VAM) have been growing in recent years due to the widespread effort to quantify teacher effectiveness. The existing VAM literature has not yet tested the sensitivity of value-added estimates to the rules that define which students contribute to each teacher’s value-added estimate. Student-teacher linkages are often a complex network due to various transfers, students taking multiple courses in the same subject, and students receiving special education or other “pull-out” services. Complex linkages are often considered as among the main threats to the validity of VAM. In this paper we conducted a case study to examine the sensitivity of VAM to the alternative link definitions. We examined three popular VAM approaches and applied alternative rules for linking students to teachers with each method. We found no overall sensitivity of estimated teacher effects to the linking rules. Even though more teachers had value-added estimates under more inclusive rules, for a teacher with estimates under all three rules, the correlation among pairs of estimates created using different linking rules was always above 0.95 and generally above 0.98 for each VAM approach. The value-added estimates of a small number of teachers were affected by highly different link definitions and these tended to be teachers with small numbers of students. Restricting the minimal sample size for calculating individual teachers’ value-added largely reduced the sensitivity in link definition.

KEYWORDS: value-added model, student teacher linkage, sensitivity
Introduction

In recent years, the widespread effort to increase student testing in conjunction with the accountability rhetoric of No Child Left Behind has brought into the policy-making spotlight the notion of quantifying teacher effectiveness. As such, both research and practice regarding teacher retention and compensation have relied more heavily on utilizing improvement in student achievement on standardized exams as the weathervane by which teacher performance can be measured. Hence, over the last decade, researchers have begun using and stressing the importance of longitudinal datasets to link student achievement with teacher and classroom information for conducting what is commonly known as “value-added” modeling to estimate teacher effectiveness. Essentially, because test scores are typically available for each student in every academic year, it is possible to implement a value-added approach in which the effectiveness of teachers can be measured by growth in standardized student test scores.

Beginning with William Sanders’ seminal work (Sanders, Saxton, and Horn, 1997) in developing the Tennessee Value-Added Assessment System (TVAAS), teacher value-added systems have since expanded to several states and to an even larger number of school districts or consortia of districts. An increasingly larger amount of weight (and press) is being placed on research findings from value-added studies throughout the country and value-added estimates of teacher effectiveness increasingly are being considered as the yardstick for high-stakes decision-making in a number of policies. However, as research and practice proceed with the use of value-added modeling, it is important to understand if and how changes in the assumptions and data underlying these models can impair their capability to accurately measure teacher performance.

Researchers have made efforts to evaluate the robustness of value-added estimates of teacher effectiveness to many factors that could cause errors in the estimates. Some studies have compared the stability of teacher effects across time (e.g., McCaffrey, Sass, Lockwood, & Mihaly, 2010; Koedel & Betts, 2005; Ballou, 2005; Bock et al., 1996) or the volatility of estimates to different statistical models and methods (e.g., McCaffrey et al., 2004; Harris & Sass, 2006; Thomas et al., 2007; Buddin et al., 2007; McCaffrey, Han, & Lockwood, 2008). Other studies have focused on the sensitivity of estimates to teacher mobility (Harris & Sass, 2006; Thomas et al., 2007; Buddin et al., 2007), to different mathematics achievement measures (Lockwood et al., 2007; Corcoran, Jennings, Beveridge, 2010), to test-score ceiling effects (Koedel & Betts, 2008), and to heterogeneity in student growth rates (Rothstein, 2010). Researchers are also increasingly interested in understanding the relative strengths and weaknesses of
these estimates within the context of policies such as teacher pay for performance programs (McCaffrey, Han, and Lockwood, 2008, Staiger and Rockoff, 2010).

The value-added literature has not yet tested the sensitivity of value-added estimates to the rules that define which students contribute to each teacher’s value-added estimate. The students who might contribute to a teacher’s value-added are determined by an interconnected network of student-teacher linkages defined by all the teachers who instruct each student in a given subject area such as mathematics or English language arts and reading. The rhetoric around value-added modeling seems predicated on student-teacher linkages in which each student is taught by one teacher per year and for the entire year, leading to unambiguous rules for determining which students should be used for calculating a teacher’s value-added. However, in reality, students’ instructional inputs are far more complicated than people imagine. Student-teacher linkages are often complex, with individual students receiving instruction from multiple teachers, due to an array of factors including: students transferring among schools mid-year, students transferring between classrooms mid-year, students transferring between classes mid-year, students taking multiple courses in the same subject, and students receiving special education or other “pull-out” services. For example, Battelle for Kids (2010) conducted a study of 125,000 class rosters in South Carolina, Texas, Ohio, and Oklahoma. They found 33 and 11 percent of students were linked to more than one teacher in the same subject in urban schools and rural schools, respectively. In an earlier report, Battelle for Kids (2009) suggested that student assignments to teachers change constantly due to factors such as student mobility, teacher reassignment and pull-outs, which result in error-prone administrative data systems and complicated modeling.

Given the complications in forming student-teacher links, well-defined rules are necessary for determining when students with complex linkages will contribute to a teacher’s value-added effect estimate for a given year. For example, the law that governs the TVAAS in Tennessee dictates that a student must be present for 150 days of classroom instruction per year or 75 days of classroom instruction per semester before that student’s record is attributable to a specific teacher (see, Tennessee Code 49-1-606). Additionally, records from any student who is eligible for special education services can not to be used as part of a teacher’s effectiveness estimate under TVAAS. However, a broad set of alternative rules could just as easily be applied and justified. Each alternative rule could in turn lead to differing conclusions about teacher effectiveness.

In a recent Washington Post article (5/9/2011), Valerie Strauss references a paper by John Ewing (2011) that suggests that complex student-teacher linkages, resulting from transferring and team-teaching, could undermine the validity of value-added analysis. In particular, she cited the rules used to link
students to teachers for modeling (e.g., 150 days or more in a teacher’s class for the current year) as arbitrarily determined, and without any solid rationale.

We too found that there are not general rules and guidelines for linking students and teachers in the presence of complex linkages and no investigations of the sensitivity to results to the linking rules that were used in a review of more than 90 journal articles, book chapters, and working papers that used value-added models of teacher effectiveness to varying degrees. A list of these references is available upon request. Studies ranged from those focused on the methodological aspects of value-added modeling to studies exploring the relationships between teacher certification and teacher effectiveness. In all of these articles, however, there was one consistent feature: a lack of detailed information on the sensitivity of specific rules used for selecting students to contribute to teacher value-added. This is not to say researchers did not take care in establishing linkages. Rather, it underscores the dearth of attention paid to this potentially important issue of defining and testing student-teacher linkage rules.

Because student-teacher linkage rules provide the basis for determining a teacher’s value-added score, it is important to understand the degree of sensitivity of a teacher’s value-added effectiveness to changes to those rules. Hence, this study will be among the first to conduct sensitivity analyses to determine if teacher value-added estimates are robust against changes in the student-teacher linkage rules. If it is the case that teacher value-added estimates change substantially in response to changes in the student-teacher linkage rules, then determination of the rules will be a key component of developing value-added evaluation programs. On the other hand, robustness in the estimates would help build confidence in the identification strategy of teacher value-added models.

**Student-Teacher Linkage Rules**

Given that complex student-teacher linkages are widespread, there are many potential student-teacher linkage rules. This section offers three plausible definitions, which represent strategies likely used in practice by states or school districts when making teacher-student linkages in complex longitudinal data systems. These definitions can be readily extended to a wider range of plausible definitions. We also describe several value-added models for examining the existence of the sensitivity of linkage definitions.

**Definitions of Student-Teacher Linkages**

In all subsequent discussions we let the subject of the test be fixed and the links and value-added are all created with respect to the selected subject. We use the term *link* to denote a pair, which includes one teacher and a student who received
instruction from this teacher in the school year the value-added estimates are calculated.

We define two basic types of student-teacher links: simple and complex. A simple link is the case where a student was taught exclusively by a single teacher during the whole school year. A student can have one and only one simple link. On the contrary, a transfer student does not have a simple link since his or her teacher changed at a point during the school year.

We define complex links as those cases where a student was taught by multiple teachers during the school year. A student can receive instruction from multiple teachers either sequentially or simultaneously or both. A complex link arises typically when a student has transferred between schools or classes, or a student took multiple courses in the same subject area, or both. They also arise when teachers change midyear as a result of reassignments or leaves of absence. By definition, a student with complex links has at least two of these. However, because the data for value-added analysis often have geographical boundaries, such as a school district or a state, some students may have more links than are recorded, so that, for students who transfer midyear across the geographic boundary, only a subset the complex links are collected in the data used for value-added modeling. Consequently a student might have only one observed complex link in the analysis data set.

This study assesses the sensitivity of value-added estimates to the inclusions of links for modeling. If a link is included in value-added model (usually used to define the design matrix of regression equations), such a link is referred to as an accepted link. Simple links are certainly accepted, but some complex links may not be, as is the case with TVAAS. In this study we use three rules to define accepted links. We then test the robustness of the value-added estimates in response to considerable changes in the rules defining accepted links.

The major concern for accepting a student-teacher link in the value-added model is whether the student received a sufficient proportion of his or her instruction from a teacher for his or her achievement to reflect that teacher’s inputs. Thus, accepted links will depend on the amount of instruction a teacher provided to the student. There are two components of instructional time that we consider to be important in the calculation. First is the amount of a teacher’s instruction a student received. If a student receives little of the instruction from a teacher then the student’s achievement might not result from that instruction. Second is the total proportion of a student’s instruction that a teacher provided. If a teacher provides little of the student’s total instruction then the effect of that instruction may be very hard to sort out from all the other inputs the student received.

We define a weight between 0 and 1 for each link that reflects the instructional time a student received from a teacher in a given school year. 0 is no
instruction time, and 1 is 100 percent instruction time. We consider two alternative weights: The first weight equals the number of class days that a student was enrolled with the teacher of the link divided by the total number of class days of all links belonging to that same student. Under this definition, weights of all links of a student sum to one. This weight reflects a teacher’s share of a student’s total instruction. The second weight equals the number of class days that a student was enrolled with the teacher of the link divided by the total number of days in the school year. The sum of these weights across links for a student can be greater than one for students with simultaneous instruction by multiple teachers. By either weight definition, simple links always have weight equal to one.

In our data students only enroll in one course in each subject at a time. Consequently, no students receive instruction from two teachers in the same subject at the same time and the two alternative weights are equal in our study sample. We define three alternative rules for accepted links using these weights:

1. **Strong definition**: all simple links and complex links with very high weight (>0.9) are accepted.
2. **Moderate definition**: in addition to the strict definition, a predominant complex link is also accepted, i.e., an accepted link has weight greater than a cutoff of 0.5.
3. **Weak definition**: all existing links are accepted, i.e., any accepted link has weight greater than zero.

Some students had class changes in the first few days of a semester and stayed in a single classroom for the rest of the semester. It seems reasonable to hold the last teacher accountable in these cases and we expect that in most applications of value-added modeling this would be the case. The student’s link to this last teacher will have a weight very close to one. Hence, all definitions above do not differentiate simple links and complex links with very high weight. We could have chosen an alternative cutoff for the moderate definition. However, a cutoff of 0.5 requires the teacher provides the majority of the instruction to a student for a link to be accepted.

Our strong and moderate definitions result in students having only one accepted link and partition the student sample in non-overlapping, mutually exclusive groups defined by the teacher links. The weak definition results in overlapping structure, or networks, where we cannot partition students into disconnected groups. To apply the link definitions (except for the specific weak definition above) detailed enrollment information on a daily basis for all classrooms is required.
Value-added estimates

Let $y_i$ denote the current year test score of student $i$ in a subject, and $y_{i,k}$, $k=1,2,$ and 3 denote the lag 1, 2, and 3 prior year scores in the same subject. Let $X$ denote a vector of student characteristics including race, gender, lunch subsidy status, and special education status. Let $Z_{j,i}$ denote the link between student $i$ and teacher $j$ in the current year. $Z_{j,i}$ can be a dichotomous indicator for an accepted link (depending on the definition in use). Some approaches also allow $Z_{j,i}$ to be continuous, in which case the weight of contribution can be used as $Z_{j,i}$. We refer the reader to McCaffrey et al. (2008) for more technical details of the fixed-effect ANCOVA and residual methods below.

1. Fixed-effect ANCOVA

This is a popular approach to value-added estimates, in which the teacher valued-added are modeled as fixed effects in an ANCOVA model and the covariates include prior test scores and student characteristics. Our model is defined as follows:

$$y_i = \mu_g + \beta_1 y_{i,-1} + \beta_2 y_{i,-2} + \beta_3 y_{i,-3} + \gamma X_i + \sum_j Z_{j,i} \theta_j + \epsilon_i.$$  

(1)

In (1), $\mu_g$ accounts for the mean of grade $g$ and $\epsilon_i$ is the random error. The fixed-effect $\theta_j$ are teacher value-added estimates with the restriction $\sum_j \theta_j = 0$ within a grade. The observed score history differs among cohorts, e.g., the cohort of grade 6 in the current year does not have $y_{i,-3}$. The fixed-effect ANCOVA further implements a pattern mixture approach to missing data (Springer et al., 2010). This allows a joint estimate of teacher value-added from all grades. In this paper, we set $Z_{j,i}$ equal to the share of instruction weight for complex linkage with the weak linking rule, and use dichotomous $Z_{j,i}$ for the other two link definitions.

2. The residual method

This is a slight variant of the fixed-effect ANCOVA. This approach does not use the links explicitly as predictors. Rather, teacher value-added are estimated from the regression residuals. This approach has two steps. First, we fit the model:

$$y_i = \mu_g + \beta_1 y_{i,-1} + \beta_2 y_{i,-2} + \beta_3 y_{i,-3} + \gamma X_i + \psi y^*_{j,-1} + \epsilon_i,$$  

(2)

where $y^*_{j,-1}$ is the mean score of students of teacher $j$ in the last year. Between student $i$ and teacher $j$ exists an accepted link in the current year. If student $i$ links to more than one teacher in the current year, there will be multiple records for student $i$ in fitting (2). We fit model (2) separately for different cohorts, since they have different lengths of observed score histories. In the second step, we calculate the raw residuals from model (2) for every student, and estimate value-added as the average of residuals linked to a teacher.
Note that the collection of $Z_{j,i}$ in (1) is a large sparse matrix, regardless of whether the elements are dichotomous or continuous. Without using the explicit link predictors $Z_{j,i}$, this residual approach enjoys better computational efficiency. However, there is no clear method for “splitting” residuals across multiple teachers. Residuals contributed equally to average for every teacher linked to the student even if students linked to multiple teachers.

3. The student growth percentile (SGP) method (Betebenner, 2008).

The SGP models the percentiles of the conditional distribution of student current year scores given their prior achievement in the same subject using nonparametric, quantile regression. Betebenner (2008) suggests summarizing the school performance using the median percentile of students enrolled in the school. We use the median and mean of the percentiles of the students linked to the teacher as the performance measure. We also use the mean and median of the probit transform of the percentiles (i.e., transformed the percentiles by the inverse of the standard normal cumulative distribution function) to estimate teacher effects. We implement SGP using the SGP package in the R environment for statistical computing (R Development Core Team, 2011). This approach is similar to the residual method, but replaces the regression model (2) by SGP. SGP is used in Colorado to evaluate schools and teachers and being is considered or used by additional states for measuring teacher effectiveness.

Sensitivity measure

The first notable difference that arises from different link definitions is the unequal numbers of teachers with value-added estimates. It is possible that a teacher, who has no accepted links under a stringent definition and consequently no value-added estimate, can have accepted links and a value-added estimate under a less stringent definition. For example, in our study there are 45 teachers with value-added estimates by the weak definition and without value-added estimates by the strong definition. Consequently, we focused on the set of “core” teachers who consistently have value-added estimates under all link definitions.

We used both global and individual measures of the sensitivity in value-added estimations. We conducted correlation analysis to examine whether a strong correspondence exists in value-added between any two definitions for acceptable links. We used both Pearson’s and Spearman’s correlation. A good correspondence suggests lack of sensitivity in link definitions because it implies that if the relative size and ordering of value-added estimates is used in teacher evaluations then the evaluations will be similar regardless of the link definition. If the absolute size of value-added estimates is used in teacher evaluations, then the mean differences in value-added among link definitions could result in
changes in teacher evaluations. Hence, we also formally tested whether a shift of mean value-added estimates exists among link definitions. Treating the value-added estimates as the outcome, we fit a two-way ANOVA model, where teachers and link definitions are the two factors. The difference in means is tested by an F-test with two-numerator degrees-of-freedom. A non-zero main effect of link definition indicates in our analysis that the mean value-added or all teachers is shifted between at least one pair of link definitions. This global measure does not apply value-added from the fixed-effect ANCOVA which by construction have mean of zero because of the constraint of \( \sum \theta_j = 0 \) within a grade.

We also used individual-level measures to identify the set of teachers whose value-added estimates have large variability across alternative link definitions. Even if the global measures suggest overall robustness to link definitions, it is possible individual teachers may have large differences in value-added among alternative link definitions. If a linear relationship exists for the value-added estimates between two link definitions, then a simple linear regression is applicable. Namely, we used the value-added under any link definition as response and the valued-added under another link definition as predictor. We used the standardized residuals to determine whether an observation is an outlier, i.e., a teacher has remarkably different value-added estimates between two link definitions. In the regression literature, an absolute standardized residual greater than 3 is often considered as outlier. We used the same empirical rule to determine individual-level sensitivity in value-added estimates.

We also considered a policy-oriented measure of sensitivity. Based on the estimated value-added measure, we constructed a decision rule to identify outstanding teachers. This rule mimics a merit-pay system in practice. Specifically, a teacher is identified as exceptional if the performance measure exceeds the threshold of the 80th percentile of an assumed reference distribution for teacher performance—a normal distribution with mean zero and variance equal to the estimated variability in teacher value-added. The reference distribution is determined jointly by all teachers, because the variability is estimated based on all teachers’ estimated value-added measures. Therefore, the decision for a teacher is slightly and indirectly influenced by others.

**Data and Sample**

The data are from the management information systems, including test scores, school enrollment histories, and snapshots of daily course enrollments, from a large urban school district in the Southern U.S. In this study we focused on teachers in the 2008-2009 school year who taught students in only one of grades 5-8. The course snapshots and school enrollment files provided data on student-
teacher links and the number of days each student was enrolled with each teacher. Before the statistical analysis, we conducted independent audits of class rosters for a sample of teachers. We focused on mathematics assessment results because typically there is greater variability in teacher effects for mathematics than other subjects (Schochet & Chiang, 2010).

In this school district, some teachers taught in more than one grade. To reduce the complexity in value-added analyses, we restricted our analyses to teachers who taught only a single grade. At the time our study, no students in grade 5 to 8 in the district were simultaneously enrolled in multiple courses of the same subject at the same time. Consequently, a student has only one weight equal to the proportion of the school year that student was enrolled in a teacher’s class.

Table 1 presents the specific link definitions with thresholds we used and the summaries of links in our data. It can be seen that our weak definition accepted roughly 38% more links than our strong definition. Some students with accepted links cannot be used in calculating value-added because they are missing test scores in either the current or prior years. Students may have missed the exams due to transfer out of the district prior to testing or been excluded because they do not speak English or completed alternative tests because of disabilities. Nearly all missing current year scores result from transfer of out the district because federal testing requirements severely limit students who can be excluded from testing. Students may be missing prior achievement scores if they transferred into the school district because testing histories were not transferred to the district, even from other districts in the state. About 8% of students with simple links, i.e., accepted links under the strong definition, are missing either the prior or current year scores required for calculating value-added. The percentage of missing test scores for students with complex links is 29%. It is so high because in this school district, almost all complex links are due to school transfer which is frequently from outside the district.

As noted above, 45 teachers with no value-added under the strong definition, have value-added under the moderate or weak definition. We restricted to the 326 “core” teachers who consistently have accepted links and value-added under all three definitions. On average, each of the “core” teachers has 37 links under the strong definition, 41 links under the moderate definition, and 44 links under the weak definition. These links correspond to students with enough test scores to fit the value-added models. The mean percentage of additional accepted links is 13% between the strong and moderate definition, and 26% between the strong and weak definition, respectively. Figure 1 plots the percentage and number of additional accepted links versus the number of simple links (with test scores) for the 326 teachers. It can be seen that the percentage change is large for a few teachers who have only a small number of students under the strong definition.
Additional descriptive statistics of student characteristics are presented in Table 2. There are a few notable differences among subgroups of students defined by our three link definitions. Students with complex links have lower test scores in both the current and the past year compared to the students under the strong link definition. In the students with complex links, there are more students with lunch subsidy, more students with special education status, and more black students. This is consistent with the literature on student transfer which finds that low income and minority students are more likely to transfer and that transfer students tend to be lower-achieving (Engec, 2006; Kerbow, 1996; Mao et al., 1998; Mehana & Reynolds, 1995; Rumberger, 2003; U.S. General Accounting Office, 1994). The pattern is the same when restricting to the 326 core teachers.

Table 1: Link Definitions and their Distribution. Some students with accepted links do not have test scores in the administrate records and are excluded from value-added analysis. The upper table lists the counts for all teachers with value-added under each definition. The lower table lists the counts for only the 326 “core” teachers with value-added across all link definitions with test scores on file.

<table>
<thead>
<tr>
<th>Link definition (weight)</th>
<th>#accepted links</th>
<th># students with accepted links</th>
<th># teachers with accepted links</th>
<th>#accepted links with test scores</th>
<th># students with test scores</th>
<th># teachers with value-added</th>
</tr>
</thead>
<tbody>
<tr>
<td>Strong (&gt;0.9)</td>
<td>13,234</td>
<td>13,234</td>
<td>328</td>
<td>12,177</td>
<td>12,177</td>
<td>326</td>
</tr>
<tr>
<td>Moderate (&gt;0.5)</td>
<td>15,581</td>
<td>15,581</td>
<td>351</td>
<td>13,864</td>
<td>13,864</td>
<td>349</td>
</tr>
<tr>
<td>Weak (&gt;0)</td>
<td>18,309</td>
<td>16,882</td>
<td>375</td>
<td>15,650</td>
<td>14,395</td>
<td>373</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>326 “Core” teachers with value-added and under all definitions</th>
</tr>
</thead>
<tbody>
<tr>
<td>Link definition (weight)</td>
</tr>
<tr>
<td>------------------------</td>
</tr>
<tr>
<td>Strong (&gt;0.9)</td>
</tr>
<tr>
<td>Moderate (&gt;0.5)</td>
</tr>
<tr>
<td>Weak (&gt;0)</td>
</tr>
</tbody>
</table>
Han et al.: Student-Teacher Linkages

Table 2: Descriptive statistics for students with test scores on file.

For all teachers

<table>
<thead>
<tr>
<th>Links</th>
<th>( Y )</th>
<th>( y_{-1} )</th>
<th>% male</th>
<th>% lunch</th>
<th>% special edu</th>
<th>% black</th>
<th>% Hispanic</th>
<th>Weight</th>
</tr>
</thead>
<tbody>
<tr>
<td>Strong definition only</td>
<td>526.0 (43.3)</td>
<td>509.8 (40.4)</td>
<td>49.8</td>
<td>67.3</td>
<td>6.2</td>
<td>45.9</td>
<td>16.7</td>
<td>1.00</td>
</tr>
<tr>
<td>Moderate definition only</td>
<td>506.3 (39.6)</td>
<td>494.2 (36.6)</td>
<td>50.6</td>
<td>84.7</td>
<td>10.0</td>
<td>57.3</td>
<td>15.9</td>
<td>.68</td>
</tr>
<tr>
<td>Weak definition only</td>
<td>504.3 (38.9)</td>
<td>492.5 (41.6)</td>
<td>50.0</td>
<td>86.6</td>
<td>28.1</td>
<td>49.5</td>
<td>26.4</td>
<td>.25</td>
</tr>
</tbody>
</table>

For the 326 “core” teachers with value-added under all definitions

<table>
<thead>
<tr>
<th>Links</th>
<th>( Y )</th>
<th>( y_{-1} )</th>
<th>% male</th>
<th>% lunch</th>
<th>% special edu</th>
<th>% black</th>
<th>% Hispanic</th>
<th>Weight</th>
</tr>
</thead>
<tbody>
<tr>
<td>Strong</td>
<td>526.0 (43.3)</td>
<td>509.8 (40.4)</td>
<td>49.8</td>
<td>67.3</td>
<td>6.2</td>
<td>45.9</td>
<td>16.7</td>
<td>1.00</td>
</tr>
<tr>
<td>Moderate</td>
<td>504.3 (41.7)</td>
<td>491.6 (40.0)</td>
<td>53.1</td>
<td>88.2</td>
<td>11.5</td>
<td>57.1</td>
<td>17.3</td>
<td>.68</td>
</tr>
<tr>
<td>Weak</td>
<td>504.5 (40.0)</td>
<td>493.3 (37.3)</td>
<td>51.9</td>
<td>89.3</td>
<td>14.0</td>
<td>64.4</td>
<td>15.4</td>
<td>.34</td>
</tr>
</tbody>
</table>
Results

Global measure

Table 3 presents the results of the correlation analysis. For all presented value-added models and linking rules, the correlations are well above .95 except for the Pearson correlation between the strong and weak definition estimates for the residual method. For each value-added method, the smallest correlation occurs between the strong and weak definitions. The correlations are slightly larger for the SGP method than the other two methods. The correlation analysis suggests both the values and the ranks of value-added estimates are insensitive to the choice of link definitions.

The test of mean shifts found no significant shift for the residual method. However, the SGP method has significant shift when the teacher effect is estimated by the mean percentile (p<.01) and mean probit (p<.01), but not when it is estimated by either of the two median-based methods. A closer examination
reveals that in both cases where the overall significant shift occurs, the significant pairwise difference is between the weak and the strong definition, where the weak definition corresponds to a significantly lower mean value-added estimate than the strong definition. However, the effect size is very small with little practical significance (less than .5 on the percentile scale).

Table 3: Correlation Coefficients between value-added estimates

<table>
<thead>
<tr>
<th>Pairs of links</th>
<th>Fixed-effect</th>
<th>Residual</th>
<th>SGP: mean percentile</th>
<th>SGP: median percentile</th>
<th>SGP: mean probit</th>
<th>SGP: median probit</th>
</tr>
</thead>
<tbody>
<tr>
<td>Strong-moderate</td>
<td>.9648</td>
<td>.9671</td>
<td>.9812</td>
<td>.9756</td>
<td>.9815</td>
<td>.9732</td>
</tr>
<tr>
<td></td>
<td>.9841</td>
<td>.9859</td>
<td>.9817</td>
<td>.9775</td>
<td>.9832</td>
<td>.9777</td>
</tr>
<tr>
<td>Moderate-weak</td>
<td>.9692</td>
<td>.9517</td>
<td>.9864</td>
<td>.9832</td>
<td>.9840</td>
<td>.9813</td>
</tr>
<tr>
<td></td>
<td>.9909</td>
<td>.9860</td>
<td>.9905</td>
<td>.9865</td>
<td>.9906</td>
<td>.9867</td>
</tr>
<tr>
<td>Strong-weak</td>
<td>.9506</td>
<td>.9313</td>
<td>.9630</td>
<td>.9684</td>
<td>.9621</td>
<td>.9659</td>
</tr>
<tr>
<td></td>
<td>.9799</td>
<td>.9757</td>
<td>.9730</td>
<td>.9685</td>
<td>.9756</td>
<td>.9688</td>
</tr>
</tbody>
</table>

*Note:* The two numbers in each cell are Pearson and Spearman correlations.

**Individual-level measure**

Table 4 presents the count of teachers with high standardized residual in pairwise regression.

Table 4: Number of teachers with high standardized residual in pairwise regression

<table>
<thead>
<tr>
<th>Methods</th>
<th>Fixed-effect</th>
<th>Residual</th>
<th>SGP: mean percentile</th>
<th>SGP: median percentile</th>
<th>SGP: mean probit</th>
<th>SGP: median probit</th>
</tr>
</thead>
<tbody>
<tr>
<td>In all 3 pairs</td>
<td>1</td>
<td>1</td>
<td>3</td>
<td>0</td>
<td>2</td>
<td>0</td>
</tr>
<tr>
<td>In at least 2 pairs</td>
<td>4</td>
<td>7</td>
<td>8</td>
<td>5</td>
<td>9</td>
<td>7</td>
</tr>
<tr>
<td>In at least 1 pair</td>
<td>7</td>
<td>8</td>
<td>14</td>
<td>13</td>
<td>13</td>
<td>16</td>
</tr>
</tbody>
</table>
Residual analyses found a small number of teachers with large standardized residuals (absolute value greater than 3). There are slightly more teachers with high standardized residuals by the SGP methods than by the other two methods. We present the scatter plot matrix of value-added estimates by the fixed-effect ANCOVA in Figure 2. Other methods have very similar patterns.

![Figure 2: Scatter plots of value-added estimates based on the fixed-effect ANCOVA](image)

We also used the policy-oriented decision rule to identify outstanding teachers. Similar to the residual analyses, the decisions differ for only a small number of teachers. For example, by the fixed-effect ANCOVA, 10 of the 326 teachers have different decisions between the strong and moderate link definitions, and 6 teachers have different decisions between the strong and weak definitions.

Furthermore, the Fleiss kappa measures of the agreement among the decisions based on different linking rules are all very high, between .87 and .91, for all the value-added models (see Table 5). A Fleiss kappa between .8 and 1 indicated very good agreement among the decision rules (Sim and Wright, 2005).

Table 5: Results from the bonus decision rule

<table>
<thead>
<tr>
<th>Fixed-effect ANCOVA (Fleiss Kappa = .91)</th>
<th>No bonus by strong def.</th>
<th>Bonus by strong def.</th>
</tr>
</thead>
<tbody>
<tr>
<td>No bonus by moderate def.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>No bonus by weak def.</td>
<td>252</td>
<td>2</td>
</tr>
<tr>
<td>Bonus by weak def.</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>Bonus by moderate def.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>No bonus by weak def.</td>
<td>5</td>
<td>1</td>
</tr>
<tr>
<td>Bonus by weak def.</td>
<td>0</td>
<td>60</td>
</tr>
</tbody>
</table>
### Residual (Fleiss Kappa = .87)

<table>
<thead>
<tr>
<th></th>
<th>No bonus by strong def.</th>
<th>Bonus by strong def.</th>
</tr>
</thead>
<tbody>
<tr>
<td>No bonus by moderate def.</td>
<td>249</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td>5</td>
<td>4</td>
</tr>
<tr>
<td>Bonus by moderate def.</td>
<td>6</td>
<td>3</td>
</tr>
<tr>
<td></td>
<td>0</td>
<td>57</td>
</tr>
</tbody>
</table>

### SGD mean percentile (Fleiss Kappa = .88)

<table>
<thead>
<tr>
<th></th>
<th>No bonus by strong def.</th>
<th>Bonus by strong def.</th>
</tr>
</thead>
<tbody>
<tr>
<td>No bonus by moderate def.</td>
<td>250</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>6</td>
<td>3</td>
</tr>
<tr>
<td>Bonus by moderate def.</td>
<td>4</td>
<td>5</td>
</tr>
<tr>
<td></td>
<td>0</td>
<td>57</td>
</tr>
</tbody>
</table>

### SGD median percentile (Fleiss Kappa = .89)

<table>
<thead>
<tr>
<th></th>
<th>No bonus by strong def.</th>
<th>Bonus by strong def.</th>
</tr>
</thead>
<tbody>
<tr>
<td>No bonus by moderate def.</td>
<td>251</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>6</td>
</tr>
<tr>
<td>Bonus by moderate def.</td>
<td>5</td>
<td>4</td>
</tr>
<tr>
<td></td>
<td>0</td>
<td>57</td>
</tr>
</tbody>
</table>
Why is there no notable sensitivity?

From all the global and individual measures presented above, we can rather safely conclude that the value-added estimates are generally invariant to linking rules for nearly all of the 326 core teachers. The sensitivity to linking rule is particularly small between the moderate and strong or the moderate and weak rules.

The value-added estimates based on some SGP methods have a mean shift from one definition to another. However, the shift is small and the ranks of the SGP value-added estimate based on different link definitions still have nearly perfect correspondence, judged by Spearman’s correlations. Thus, such a small shift in the mean value-added of teachers is unlikely to have any real consequence. Based on the residual analysis and the classification of exceptional teachers, there are very small proportions of teachers (<5% usually) who would have been affected by changing link definitions for each of the value-added methods. Such a difference is rather common compared to other uncertainties of value-added analysis, e.g., the choice of value-added models.
There are a few potential reasons for the lack of sensitivity due to linking rules in this case study. For most teachers, the links added by the moderate and weak definitions are a relatively small percentage of the total accepted links under the strong definition. More importantly, the differences between the achievement of the added students by moderate and weak definitions and those accepted by the strong rule are generally small even though on average added students have lower scores, and when combined with the small number of added students result in negligible changes to teacher value-added estimates for all but a few teachers. For most teachers the difference in value-added among link definitions is likely due to normal statistical uncertainty of (additional) data.

The lack of sensitivity in this case study may also be due to the lack of simultaneous links. Since no student takes multiple courses at the same time in math, the complex links are only introduced by transfer students, which only account for a relatively small percentage of all students in the analysis data set. If the teaching practice was more flexible, e.g., students take more than one course in math (potentially from two or more teachers), the percentage of complex links could be much larger and possibly result in more sensitivity of estimates for more teachers.

There are a small number of teachers whose individual-level sensitivity measures are large, i.e., large standardized residuals from pairwise regression or different classification as outstanding. A closer examination of the change in value-added suggests that the percentage of added (reduced) links may be a good indication for the sensitivity. For instance, if the weak definition accepts about 50% or more links than the strong link for a teacher, his or her value-added may change to a notable degree according to the individual-level sensitivity measure. Figure 3 illustrates the change in value-added using the fixed-effect ANCOVA and SGP mean percentile methods as an example. Although many teachers with more than 50% links still do not have notable changes in value-added, the large change in value-added often occurs for teachers with 50% more links.

In Figure 1 we see that the percentage of added links is large for those teachers with small numbers of accepted links under the strong definition. For example, a teacher who has only 2 accepted links under the strong definition will have a 100% increase if the weak definition accepts 2 more links. Teachers with 20 or more links under the strong definition never have a large percentage change. This suggests that a lower bound of accepted links may help further eliminate the sensitivity, e.g., a teacher has to have at least 20 accepted links under the strong definition to be evaluated.
Discussion

The alternative link definitions can lead to different analysis data sets for calculating teacher value-added as they did in our case study where about 15 percent of teachers had value-added with the weak linking rule but not with a strong rule. In this paper we present a set of statistical tools to examine the sensitivity in value added estimates. Given that value-added estimates are being tied to high stake decisions in the current education policy, more sensitivity studies including the one on link definitions should be incorporated in the future VAM studies.

In this case study, however, we did not detect notable sensitivity due to alternative link definitions except for a very few teachers. This finding may result from the fact that the school district we studied lacks simultaneous links, or has relatively fewer within-district transfer students, or both. This case study should not be over interpreted to conclude robustness of VAM to alternative link
definitions in general. Instead, we recommend the measures and procedures in this paper for checking sensitivity in the future VAM studies.

This study does suggest that the sensitivity is related to the percentage of added or reduced links by alternative definitions. This can depend on the difference between alternative definitions. To avoid the sensitivity for some teachers, the adopted link definition should not change drastically from one extreme to another (e.g., from strong to weak in our definition). In addition, this study suggests that percentages of added (reduced) links of a teacher is weakly related the change in value-added. Large percentages of added (reduced) links usually occur for teachers with small numbers of links under the most stringent definition. In practice it may be appropriate to exclude teachers with few accepted links from the formal evaluation due to the potential sensitivity of value-added.

References


