How Many Students Need to be Replaced to Invalidate a Teacher's Evaluation Based on Value-added?: an Approach to Characterize the Uncertainty, Interpret and Make Use of Value-added

Abstract

Value added measures have been used to evaluate teacher effectiveness, informing high-stake decision-making for individual teachers, such as determining hiring, tenure, compensation and directing professional development. For such high stakes decisions it is essential to know the uncertainty of the measure. Firing or promoting a teacher based on only an uncertain point estimate may be unfair or create a loss of investment or resources for a school.

However, the uncertainty of value-added measures is not well represented for administrators and policy-makers to make such high stakes decisions. Current approaches quantify uncertainty in terms of the standard error of the estimated effect, which only focuses on the variation caused by randomness or sampling error. Additionally, this concept of standard error is not well-understood as how it is calculated. This even includes the fact that the sample size can vary depending on how you conceptualize the level of analysis for calculating the standard error. Moreover, the standard error is not easy to understand by policy-makers since it requires thinking about a repeated sampling framework that conjures a scenario which is beyond the observed data. Typically, it gets interpreted with confidence intervals to make an evaluation relative to a threshold but the confidence interval may create more cognitive demand for policy-makers. In this study, we conceptualize uncertainty and bias in value added measures in terms of random or purposeful resampling of the students in a teacher's class. Importantly, this will allow us to represent in a framework that policy-makers can understand more easily and intuitively.

We begin by reviewing potential sources of bias in value-added estimation under conditional random assignment. Almost all the potential inconsistencies of value added, such as those caused by test unreliability, missing data or model specification, can be represented in terms of violations of conditional random assignment assumption. For instance, measurement error could be thought as an unobserved confounding variable that brings inconsistency in estimating the teacher effect by correlating both with teacher assignment and with changes in students' test scores.

Our approach to quantify the part of uncertainty of value-added leverages the potential for non-random student-teacher assignment to create a general framework to quantify sources of bias/inconsistencies. By design, this framework recognizes the agency of the administrator, teacher and parents in assigning students to teachers rather than relinquishing the agency to a random mechanism that rarely applies in everyday contexts. Specifically, we will characterize the uncertainty of value-added measures in terms of the number of students that would have to be replaced with other students to change the evaluation of a teacher relative to a threshold for defining competency.

We will first present how this student replacement approach works in an ideal situation where students are assumed independent of one another. Then we will discuss how this replacement idea can

be generalized for different scenarios. For example, we will present a replacement procedure to quantify the potential bias accounting for spillover effects among students, which is a violation of the independence assumption (and SUTVA) that the repeated sampling framework relies on. We will also propose several purposeful replacement approaches when there is a concern about possible violation of the constant effect assumption.

In general, our framework leads to statements such as "For a teacher evaluated below a threshold for effectiveness, xx of her students would have to be replaced with average students in the grade to move her above the threshold". The number of replacement students here cannot only be used to quantify the value-added measure's robustness to potential sources of bias but also be applied to provide an intuitive description for how far a teacher is from a threshold. In some special scenarios, this number may provide extra information to inform comparisons between teachers who receive equivalent value added scores.

Importantly, our framework seeks to provide an intuitive framework that complements the traditional standard error approach by accounting for the component of uncertainty due to potential bias/ inconsistencies. Meanwhile, once we allow randomness in choosing students, we can accommodate the repeated sampling framework employed in interpreting standard errors.

In the last part of this paper, we will study how our approach can help administrators and teachers make use of value-added measures for better prospective decisions based on anticipated student test scores. For example, the teacher may adjust her teaching strategies for particular students so that she can move her value added scores above a threshold in the future.

1. Introduction

Value added models have gained increasing popularity with the 2002 No Child Left Behind (NCLB) to measure school and teachers' effectiveness on students' progress. By comparing students' expected test scores to their actual ones, the "deflections" are then inferred to be the "added value" from the school and teacher (Raudenbush & Bryk, 2002). The federal government's Race to the Top competition, under President Obama's administration, further promoted the adoption of the value-added measures to inform teacher evaluation.

Proponents of value-added models cite research that show teachers' considerable and long-lasting influences on students' achievements (Chetty, Friedman, & Rockoff, 2011; Hill, Kapitula, & Umland, 2011; Rivkin, Hanushek, & Kain, 2005). They argue that there is important variation in teachers' effectiveness (Aaronson, Barrow, & Sander, 2007) that can be better identified by value added measures (Hanushek & Rivkin, 2010). By selecting or deselecting teachers based on value-added we can improve teacher quality and increase student achievement and long-term outcomes (M. A. Winters & Cowen, 2013; Gordon, Kane, & Staiger, 2006; Marcus A. Winters & Cowen, 2013; Chetty et al., 2011). It is also shown that value-added measures provide a better prediction for future student achievement than observed teacher attributes that are currently applied in the labor market (D. Goldhaber & Hansen, 2010). The Measurement of Effective Teaching Project discussed several approaches to combine the value-added with other measures to generate a composite measure (Kane & Staiger, 2012).

However, various concerns have been raised about the validity and reliability of value added measures (VAM) as a basis to inform high stake decisions (e.g., hiring, retention, and professional development) for a specific teacher (Guarino, Reckase, & Wooldridge, 2014; Harris, 2009; Raudenbush, 2015). We will begin our review of these concerns with the conditional random assignment assumption. Almost all the potential inconsistencies of value added, such as those caused by test unreliability, missing data or model specification, can be represented in terms of violations of conditional random assignment assumption. Our approach to quantify the uncertainty of value-added also draws on this framework of the student-teacher assignment mechanism.

The conditional random assignment assumption illustrates that students are randomly assigned to every teacher conditional on the other variables (Rothstein, 2009, 2010). However, research have shown that there is a nontrivial amount of sorting based on students' prior test scores as well as a nontrivial amount of non-random assignment of teachers to classrooms. For example, recent research has shown that teachers who are nominated as help-providers to other teachers and with leadership positions are assigned better students (Kim, Frank, & Spillane, 2018). These nonrandom assignments may cause substantial bias in value added estimates if not captured by the controls in the model specification (Paufler & Amrein-Beardsley, 2014; Rothstein, 2010). In some cases, the estimates based on value added may even have the opposite sign of the true teacher effect (Dieterle, Guarino, Reckase, & Wooldridge, 2015). Hiring or dismissing a teacher based on this flipped ranking can be unfair for teachers and cause unwanted competition that can lead to test-driven teaching.

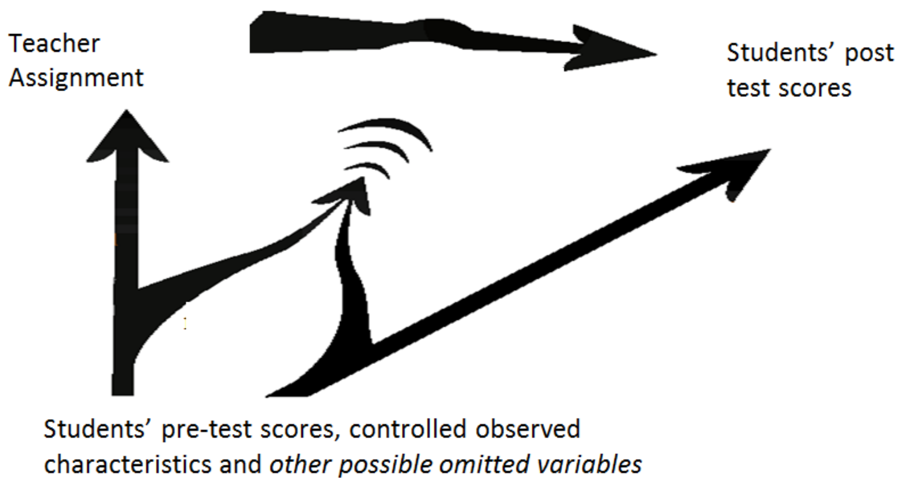
Other commonly-discussed concerns about value-added measures include missing student testing data and unreliability in testing scores. In order to get a more reliable value added measure, researchers recommend model specification with two years of prior tests (D. Goldhaber & Hansen, 2010; Kane & Staiger, 2012; Rothstein, 2009). But it is not uncommon for teachers to change grades or students to change schools within three-year duration. For example, 6.7% K-12 students in Michigan changed schools during the school year 2015-2016[1] and this is just for a single year. Unreliability in test scores can appear as measurement errors or differences among different achievement measures. Previous research have shown bias in value-added caused by measurement errors alone (Lockwood, Louis, & McCaffrey, 2002) as well as large variation in the estimated effects of applying different achievement measures (Lockwood et al., 2007). All this uncertainty could lead to lack of legitimacy in those high stakes decisions.

From a regression framework, we can regard these concerns of missing student data and unreliability in test scores as learning the effects of possible omitted variables that can violate the conditional random assignment assumption. In value-added models, this assumption is realized by controlling as many factors as we can that may affect both students' post-test scores and teacher assignment to students. But in empirical research we can never know whether there are other crucial

[1] This percentage is calculated based on the report from https://goo.gl/3UCD53. There are 99,750 mobile students and 1,383,815 stable students during the school year 2015-2016.

confounding factors that are not included in our model. Though those concerns of missing data and unreliability are rarely understood as a confounding variable, the way they bring bias is through affecting both the outcome variable (post-test scores) and the key predictor of interest (teacher assignment), just as the possible omitted variables in Figure 1(as in Frank, 2000). Or we can think all these concerns as worrying about teacher assignment being an endogenous variable. For example, we can think about measurement error as an unobserved confounding variable (as in Frank, 2000). Then the bias caused by measurement error is only possible if this variable correlates both with teacher assignment and with changes in students' test scores. The previous correlation can be possible due to a tracking system and the fact that some groups of students may be more likely to show measurement error in their test-scores (Koretz et al., 2016).

Figure 1. Effects of Confounding Variables in Value Added Models



Therefore, we summarized these sources of bias as they can lead to violation of the conditional random assignment assumption and thus bring uncertainty to the estimated effect.

In spite of the uncertainty of measurement, , value added measures have been used to evaluate teacher effectiveness in many school districts, such as the Chicago Public Schools, New York City Department of Education, District of Columbia Public Schools and some districts in North Carolina, Tennessee and Ohio.[2] Inspired by the Race to the Top grant competition, some states are using value-added measures, together with other measures, to inform high-stake decision-making for individual teachers, such as determining hiring, tenure, compensation and directing professional developments.[3] For such high stakes decisions it is essential to know the uncertainty of the measure. Firing or promoting a teacher based on only an uncertain point estimate may be unfair or create a loss of investment or resources for a school.

[2] The information is from https://en.wikipedia.org/wiki/Value-added_modeling.
[3] See footnote 2.

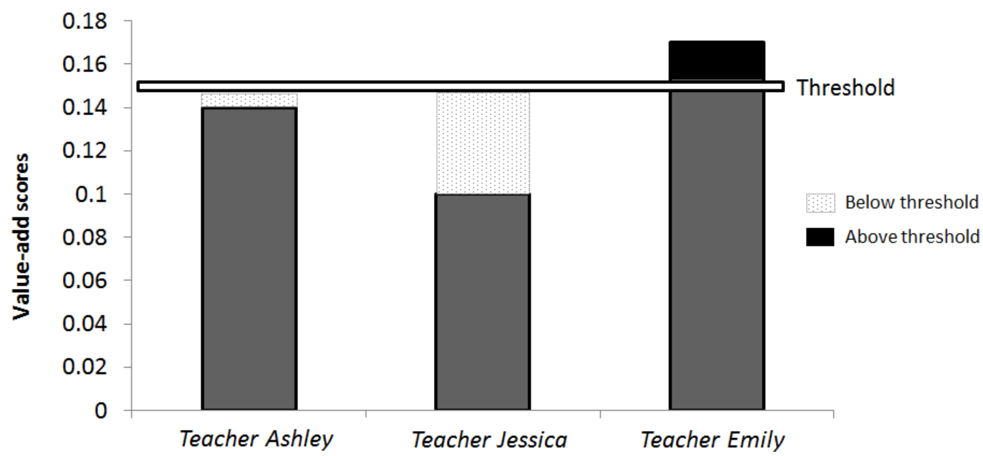
However, the uncertainty of value-added measures is not well represented for administrators and policy-makers to make such high stakes decisions. Current approaches quantify uncertainty in terms of the standard error of the estimated effect, which only focuses on the variation caused by randomness or sampling error. Additionally, this concept of standard error is not well-understood as how it is calculated. This even includes the fact that the sample size can vary depending on how you conceptualize the level of analysis for calculating the standard error. Moreover, the standard error is not easy to understand by policy-makers since it requires thinking about a repeated sampling framework that conjures a scenario which is beyond the observed data. Typically, it gets interpreted with confidence intervals to make an evaluation relative to a threshold but the confidence interval may create more cognitive demand for policy-makers.

Our approach to quantify the part of uncertainty of value-added leverages the potential for non-random student-teacher assignment to create a general framework to quantify sources of bias/inconsistencies. By design, this framework recognizes the agency of the administrator, teacher and parents in assigning students to teachers rather than relinquishing the agency to a random mechanism that rarely applies in everyday contexts. Specifically, we will characterize the uncertainty of value-added measures in terms of the number of students that would have to be replaced with other students to change the evaluation of a teacher relative to a threshold for defining competency. This generates statements such as "For a teacher evaluated as below a threshold for effectiveness, xx of her students would have to be replaced with average students in the grade to move her above the threshold". Thus, the sensitivity analysis replaces the expression of uncertainty based on repeated random sampling with a measure of uncertainty based on the single deliberate resample necessary to change a decision. In this sense, the sensitivity framework takes care of the component of uncertainty due to potential bias. Meanwhile, once we allow randomness in choosing students, we can accommodate the repeated sampling framework employed in interpreting standard errors. We apply our framework to retrospective decisions of the relevant actors based on known student test scores or prospective decisions based on anticipated student test scores.

2. Aims of this study

The ultimate goal of any educational research is to inform decision-makings in practical education terms. So is the value added estimate for teacher effectiveness. By comparing a teacher's value added score to a certain proficiency threshold, personnel decisions will be informed and resources will be reallocated. For example, school administrators may decide to dismiss a teacher with a below-threshold value added score or send some of these teachers to professional developments. In other words, the purpose of value-added is to inform all these decisions in practice.

Figure 2. Teacher Effects Estimated by Value-added (Hypothetical)



However, there are many concerns about the validity and reliability of value added and we need an intuition to inform the debate on value-added as basis for high stakes decision-making. To illustrate, three teachers' value added scores are presented in Figure 2. Both Ashley and Jessica are below the threshold of 0.15. If the threshold represents a serious lack of proficiency, the administrator may decide to dismiss both of them. However, we can see that Ashley is much closer to the threshold than Jessica. This indicates that an evaluation of Ashley as ineffective is much less robust than that for Jessica. It might be some bias in value added score estimation causes teacher Ashley to be below the threshold. As a result, the personnel decision should be considered more seriously or other measures should be referred to. Or the administrator may direct Ashley to professional development if the school resources can only support one teacher for this opportunity. Similarly, an administrator may want to provide professional development to teacher Emily who is above the threshold, but just barely so.

This study puts forth a non-parametric approach not only to characterize the uncertainty of the value added measures but also to formalize the interpretation of value-added. To do this, we ask an intuitive question: how many students need to be changed to alter or invalidate the teacher evaluation based on value added? We then use the answer to quantify the robustness of evaluations based on value added measures.

This question "how many students need to be changed to alter or invalidate the teacher evaluation based on value added?" is derived from a framework for quantifying sources of bias for both internal and external validity (Frank, Maroulis, Duong, & Kelcey, 2013). Importantly, this framework allows us to identify a "switch point" where the bias we are concerned about is large enough to invalidate the teacher evaluation. It is of great significance to quantify the switch point in an intuitive way for certain sources of bias that we are concerned about. Because this enables policy-makers to better evaluate whether there is potentially large enough bias to invalidate our teacher evaluation result. This evaluation process can add legitimacy to those high-stake decision-makings based on the value added results.

Equally important, Frank et al's (2013) approach can provide an intuitive way to formalize the discourse about how far a teacher is from a threshold. This is a critical step if we expect to improve teacher quality through promoting the use of value-added measures. The way we interpret the measure directly affects teachers' understandings and perceptions of the measure's accuracy and fairness. As proponents for value-added argue, in addition to the measure accuracy, we are supposed to care about the causal impacts on the teacher workforce quality: how will teachers (and potential teachers who may enter the labor market) react to this? What are their behavior responses (D. Goldhaber, 2015; D. D. Goldhaber, Goldschmidt, & Tseng, 2013)? These responses are highly dependent on their understandings of the measures.

This study also extends Frank et al's (2013) work by discussing how we can quantify the uncertainty to the Stable Unit Treatment Value Assumption (SUTVA). In the value-added context, this assumption could be violated if there are spillover effects in one classroom or teacher's having varying effects on different students. We provide ways to quantify the uncertainty to these concerns by generalizing the replacement framework. These discussions can also be applied beyond the value-added context as long as we have theoretical reasons to concern about the SUTVA assumption.

The first part of this study (Section 4) adopts a retrospective point of view. We will answer the question to quantify the uncertainty of an existing teacher evaluation result and to provide an intuitive interpretation of the "switch point". The second part of this research (Section 5) is from a forecasting/planning view by studying how value-added could be applied as tools for school administrators and teachers. The approach is similar to the first part, which is to ask the same question about how many students need to be changed. But the goal is to realize better teacher assignment (for school administrators) or to achieve higher value added (for teachers) in the future rather than characterize the uncertainty of value added in an evaluation framework. Section 6 will discuss the implications of this study from different perspectives and section.

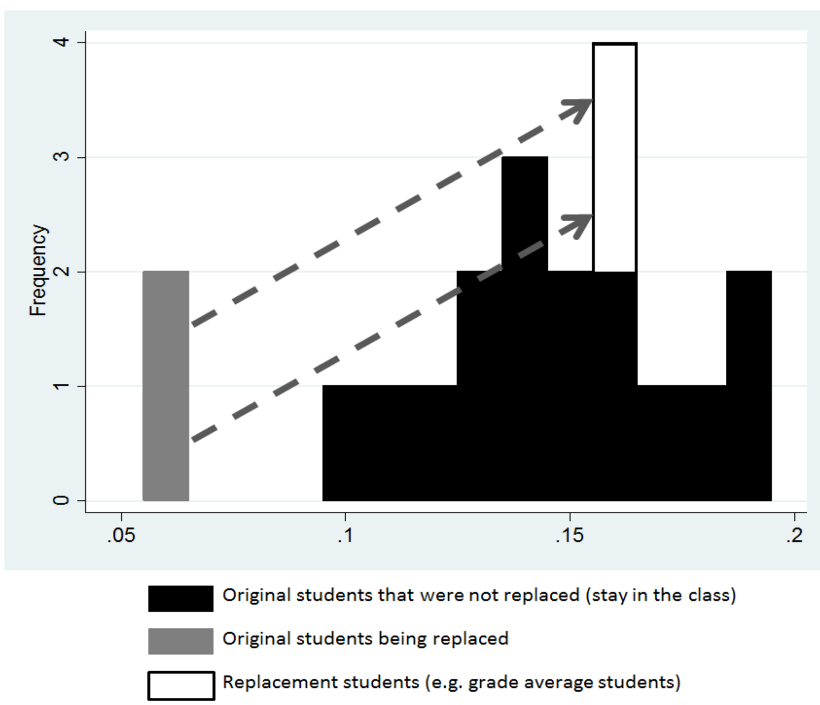
3. Theoretical framework of this study

This research draws the idea from a study that provides a method for quantifying the robustness of an inference(Frank et al., 2013). The authors show an approach to quantify how much bias there must be in an estimate to invalidate an inference. Then this bias is interpreted in terms of sample replacement to be more intuitive for interpretation. In other words, to show how robust an inference is, we ask a question based on a thought experiment which is counterfactual: what percentage of the samples should be replaced with counterfactual (unobserved) no-effect cases to invalidate an inference made from the data. Or if the concern is about external validity, we think about what percentage of the samples should be replaced with unsampled no-effect cases. The larger the percentage is, the more robust the conclusion/inference is, the less likely that the finding is only due to chance.

This case replacement idea can be applied in various ways to characterize the uncertainty of value added and to realize better application of value added. The general idea, however, is always about replacing some of a teacher's students with other students so that the teacher's VAM is moved above the threshold. We build our framework with this teacher-student assignment because it is directly related to the fundamental assumption of conditional random assignment. For the purpose of this study, we focus on the mechanism of assigning students to teachers and how the assignment can affect the estimated effect. This is general since we can understand those specific sources of bias as a potential confounding variable that can violate the conditional random assignment assumption. Additionally, applying this framework for prospective decisions based on anticipated student test scores help us recognize the agency of teachers, administrators and parents in authentic settings.

Figure 3 illustrates the student replacement idea with a toy data. Teacher Ashley in Figure 2 has a value added score of 0.14, which could be the average of her students' gain scores in a very simple value added setting. The proficiency of threshold is 0.15. Assume Ashley has 20 students, whose gain scores have a distribution represented as black and grey parts shown in Figure 3. Hypothetically, to improve Ashley's value added score from 0.14 to 0.15, we can replace two students (the grey parts) with two grade average students whose gain scores are 0.16 (the white parts with black outline). This counterfactual thought experiment tells us that via replacing only two students with grade-average students, Ashely could achieve the threshold of proficiency. We can also say that 2 out of 20, that is about 10% of Ashley's students need to be replaced with grade average students to alter the evaluation.

Figure 3. Example replacement of students to invalidate teacher's evaluation based on value added



The hypothetical example above only gives one possible replacement to move the teacher above the threshold. This can be generalized depending on: (1) how we select students from the teacher's class to conduct the replacement (also the distribution of students' gain scores) (2) what students are regarded as replacement cases. In later analysis, we will discuss different ways to think about the hypothetical replacement and how each way helps inform the debate on applying value added for high stakes decision-making in different contexts.

4. Retrospective: characterize the uncertainty of value-added with an intuitive interpretation

The goal of value-added is to have a measure for a teacher's effect on students' achievement. Ideally, we hope that we can learn about each individual teacher's effect on all students. This is analogous to the classical model for causal inference. The teacher is playing the role of "treatment" in this context. The validity of value-added then relies on to what extent we can satisfy the classical assumptions in causal inference, such as independence (or conditional independence in nonrandom observational studies) and the SUTVA assumption (Holland, 1986). In following discussions, we will look at these assumptions in the context of value-added and discuss how we can quantify the uncertainty of value-added by accommodating these assumptions in the replacement framework.

The conditional independence assumption in observational studies is equivalent to conditional random assignment assumption in value-added models. Many specific sources of uncertainty may relate to student-teacher assignment to contribute to final bias in the estimated teacher effects. By discussing the uncertainty through the mechanism of teacher-student assignment, we are not only studying uncertainty caused by non-random assignment but also other concerns that cause bias through correlating with the assignment. To do this, we quantify the uncertainty in terms of the number of students that would have to be replaced with other students to change the evaluation made about a particular teacher.

It is crucial to point out that all of the discussions in this part are counterfactual thought experiments. We ask a straightforward question in terms of student replacement to quantify the uncertainty of an existing value added estimate. The intuition of the idea is illustrated in the toy example in Figure 3, in which the teacher Ashley needs to replace two students with grade average students to achieve the threshold. The number two here is just a result in our thought experiment to quantify how uncertain the value added is or how far the teacher is from the threshold.

We now formalize the intuition in Figure 3. For the following discussion, assume that all grade nine math teachers in one middle school were evaluated based on their students' achievement scores. Suppose we have a general value-added model as follows:

$$A_{it} = \tau_t + \lambda A_{i,t-1} + T_{it}\gamma + X_{it}\beta + \mu_{it}$$ [4]

where

[4] There are various value added models. This particular (simplified) model is only used as an example to illustrate that: all the following discussions in this study are based on the "purified" or adjusted "gain scores". In other words, the gain score here is after adjusting for student characteristics that are included in the value-added model. Theoretically, the change in students' test scores can be decomposed to teacher effectiveness and student heterogeneity. By removing the latter part, we can estimate the teacher effectiveness.

A_{it} is student i's test score at time t (post-test score);
τ_t is the intercept;
λ is the coefficient (scaler) for the pre-test score $A_{i,t-1}$;
$A_{i,t-1}$ is student i's test score at time $t-1$ (pre-test score);
T_{it} is a row vector of teacher indicators;[5]
γ is a column vector of teacher fixed-effects[6];
X_{it} is a row vector that include covariates to control student heterogeneity such as student family backgrounds;
β is a column vector that include the coefficients for the covariates X_{it};
μ_{it} is an unobserved error term.

After estimating those parameters, we obtain a "purified" gain score s_{il} for each student i in teacher l's class at time t after removing the effects from those observed characteristics ($A_{i,t-1}$ and X_{it}) included in the value-added model. This is shown in the following equation:

$$s_{il} = A_{it} - (\hat{\tau_t} + \hat{\lambda} A_{i,t-1} + X_{it}\hat{\beta})$$

This gain score s_{il} can also be understood as a "deflection" score which is the difference between a student's expected score (based on those covariates $A_{i,t-1}$ and X_{it} that are outside teacher l's control) and actual score. We assume that this deflection is caused by teacher l [7] who teaches student i. To clarify, all the gain scores in following discussions refer to this s_{il}. We can decompose s_{il} by using ANOVA parameterization: $s_{il} = \mu + \alpha_l + e_{il} = VAM_l + e_{il}$

where

μ is the grand mean gain score of all students in this grade. For simplicity, we assume that each teacher teaches one class where each class has the same number of students (n). Then $\mu = \overline{VAM}$, which is the average value added score of all teachers in this grade;

α_l is how far teacher l's value added score (VAM_l) departures from the $\overline{VAM}$;

VAM_l[8] is the value added score for teacher l.

One implication of the conditional random assumption is that teacher has a constant effect on all the students who could be in her classroom. That is the teacher's effect is the same regardless of whether or not the student is assigned to her. This is related to the constant effect assumption in SUTVA.

---

[5] If there are 10 teachers in this grade, then T_{it} is a 1*10 row vector for each student (observation). Correspondently, γ is a 10*1 column vector, with each element representing one teacher's fixed effect.
[6] Here "fixed-effect" means that we are NOT viewing all teachers as a population and then getting an estimate based on a random sample drawn from this population of teachers. Instead, we are interested in learning individual teacher's effects on students' achievements. Therefore, we just use dummy variables to indicate each teacher (see footnote 5). There are also other estimation methods in value added literature. This study will focus on the teacher-fixed effect as an example for now. (This is different from the "fixed effect" in panel data context.)
[7] This study mainly discusses using student-level data to evaluate teachers within a school. Therefore, the school-level effect is not included.
[8] If the pre-test score (test score for time 1) is set before the teacher encounters the student, then we can think about this VAM as a function of the post-test score (test score for time 2).

Another crucial part of SUTVA is to assume that the teacher's effect on one student will not affect the other students' performance. In this section 4, we will first discuss scenarios where we assume there is no spillover effect. Then we will take a look at how we can quantify the violation of the no spillover effect assumption.

By discussing how to quantify the uncertainty to these assumptions, we aim to achieve three goals with this student replacement thought experiment in this evaluation framework. They are listed as follows and different replacement schemes in later analysis may serve for parts or all of these goals.

Goal (a) is to obtain an intuitive interpretation of how uncertain the value-added is: how robust is the inference that a teacher's effectiveness is below a threshold? This uncertainty may come from various omitted variables in a regression framework as shown in Figure 1 or other violations of assumptions underlying the value-added models.

Goal (b) is to formalize the discourse about how far a teacher is from the threshold.

Goal (c) is to develop some other measures that may help rank teachers, that is, to have an idea of how uncertain/sensitive the teacher evaluation is to students' heterogeneity (as a violation of the constant effect assumption).

4.1 Three replacement approaches when there is no spillover effects

First we assume that students are not affected by others in the same class. When we replace students, the change of teacher l's value-added is only from the difference between the new students' and original students' gain scores after the replacement. For example, in the hypothetical example in Figure 3, teacher Ashley needs $(0.15 - 0.14) \times 20\ (the\ teacher\ has\ nine\ students) = 0.2$ total increase to achieve the threshold. This 0.2 comes from $(0.16 - 0.06)\ +\ (0.16 - 0.06)\ =\ 0.2$, which is the difference between the two original students in the class (denoted in grey parts) and two replacement students (represented as the white parts). Those original students who are not replaced do not change their scores in the replacement process.

To simplify the discussion for now, assume that we are evaluating teachers for one grade within one school. One way to set the threshold is to use a certain percentile such as the 5th percentile in all teachers' value added distribution.

For teacher l, we can only observe her effect on the students in her class. For the other students taught by other teachers, we cannot know their scores if they were taught by teacher l because this is counterfactual. Because of this, teacher l may argue that her value-added VAM_l is below the threshold (Thr) because of the students she is assigned. She may argue she actually has the average teacher effect in this grade and she will be able to achieve the threshold if she is assigned with more grade average students (this could well be the argument of a beginning teacher – see Kim, Frank and Spillane, forthcoming). However, the evaluator, such as the principal, may argue that this low value-added VAM_l reflects that this teacher l has a low effect. While the dispute is about the point estimate of the VAM, the debate about the teacher's evaluation is informed by understanding the uncertainty of the VAM.

To formalize the discourse above for the uncertainty of value-added, we can think about how many grade average students need to be replaced to alter teacher l's evaluation. Another argument for replacing with grade-average students is that if we randomly choose one student from the grade then a grade-average student will be the expectation for a student being selected.

In order to do the replacement analysis, we need to know the grade average student's gain score. As before, this gain score is achieved after adjusting for all those covariates included in the value-added model. Two possible ways are presented as follows to get an estimate for this grade-average student's gain score g_t.

In the first approach, we can just use μ as an estimate for g_t. This approach is convenient and the resulting g_t will be the same for all teachers in the replacement thought experiment. From the teacher's argument illustrated before, she has the average teacher effect in this grade and this g_t might be a good estimate for an average teacher's effect on a grade average student. The disadvantage is that this average gain score is under the observed teacher-student assignment condition and we are assuming that this grade average student will keep the grade average score if taught by this teacher.

Another possible way to estimate this grade average student's gain score g_t is still conditioning on covariates in the current model and the observed teacher-student assignment but is more conservative. Specifically, rather than look for an estimate for an average teacher's effect on a grade average student, we try to estimate this particular teacher l's fixed effect on a grade average student. The "average" here refers to having the grade average pre-test scores and other controlled characteristics. This means looking for a student j in teacher l's class so that the value of $\left|\left(A_{j,t-1} + X_{jt}\right) - \left(\overline{A_{t-1}} + \overline{X_t}\right)\right|$ is minimized (where $\overline{A_{t-1}}$ and $\overline{X_t}$ are the grade average covariates). Then we use this student j's gain score s_{jl} as g_t. The second approach here seems to provide a "closer guess" for a grade average student's gain score if taught by teacher l. However, since we only use one particular student's observed gain score for the replacement, the reliability will be a more serious issue than the first approach.

Once we get g_t, we can conduct following three replacement thought experiments to quantify the uncertainty of value-added.

4.1.1 Random replacement – Goals (a) and (b)

If we consider randomly selecting students from teacher l's class to be replaced with grade average students, then the formula is shown as follows.

$$Thr = (1 - \pi) \cdot VAM_l + \pi g_t = (g_t - VAM_l)\pi + VAM_l$$

From this we can get:

$$\pi = \frac{Thr - VAM_l}{g_t - VAM_l} = 1 - \frac{g_t - Thr}{g_t - VAM_l}$$

where

π is the percentage of students need to be changed/replaced,
Thr is the threshold of value-added above which the teacher will be evaluated as eligible. In this paper, we assume that threshold(Thr) is below the average value added score ($\overline{VAM}$) and all the VAM_l we are interested in is below the threshold(Thr). Therefore, we have $VAM_l < Thr < \overline{VAM}$.

Suppose g_t is bigger than Thr and VAM_l. Also the g_t is the same for all teachers (the first approach discussed previously). Then with higher VAM_l, the π gets smaller. This makes sense intuitively because teachers who are closer to the threshold only need to change fewer students. However, we can see that the relationship between VAM_l and π is not linear.

If we treat π as a function of VAM_l (and assume g_t as a known constant for now), we can apply delta method to get a standard error for π as follows.

$$\frac{dVAM_l}{d\pi} = (Thr - g_t) \cdot (g_t - VAM_l)^{-2}$$

$$(\frac{dVAM_l}{d\pi})^2 = (Thr - g_t)^2 \cdot (g_t - VAM_l)^{-4}$$

Then we can get:

$$AVar[\sqrt{n}(\hat{\pi} - \pi)] = (Thr - g_t)^2 \cdot (g_t - VAM_l)^{-4} \cdot AVar[\sqrt{n}(\widehat{VAM}_l - VAM_l)]$$

From this formula, we note that with the value-added measure getting further away from the grade average gain score (that is, the $(g_t - VAM_l)$ gets larger), the asymptotic variance of $\hat{\pi}$ approaches 0 because of the term $(g_t - VAM_l)^{-4}$. This indicates that for teacher with a relatively low value-added measure, we can get a quite precise estimate for π. In that case, a relative large $\hat{\pi}$ can represent a quite robust evaluation for a teacher's incompetence relative to a threshold.

In order to account for randomness in estimating the $\widehat{g_t}$, we may consider a bootstrap approach. In this way, we can get a confidence interval for $\hat{\pi}$ that accommodates both randomness and potential bias. If this confidence interval covers 0, then this may indicate that the evaluation for this teacher is not robust.

4.1.2 Constant effect assumption and selective replacement – Goals (a), (b) and (c)

Constant effect is another crucial assumption in causal inference. As illustrated by Holland (1986), the average causal effect is an average and thus it "enjoys all of the advantages and disadvantages of averages". The constant effect says that all the units in the population of interest experience the same effect caused by the treatment. This assumption will then allow the average treatment effect to be used to draw causal inference at the unit level.

One thing to note here is that this constant effect assumption does not necessarily relate to student grouping based on prior test results. Prior test scores may reflect students' ability but the constant effect assumption is about teachers' effects on students. Students' ability may or may not relate to their improvement affected by the teachers. The treatment effect in this context is more a problem of whether this teacher's teaching works for one student (matching problem). Therefore, even in the most homogeneous case where students are grouped based on their pre-test scores we still need to think about violation of the constant effect assumption.

One may argue that we only need to draw causal inference at an average level rather than at a unit level. But if our goal is to rank all the math teachers, we should consider the heterogeneity of the students in their classes and ideally the teacher whose teaching works out for more students should be more favored. Think about two teachers who have the equivalent value-added. In teacher l's class, there is only one student who gets extremely low gain score and it is this score that makes the teacher's value-added below the threshold. However, teacher m has several students who get quite low gain scores. Apart from the value-added, we can also use this information of mismatching as another measure for teacher's effectiveness. Even we are only interested in the average level, we may still concern about the effects of outliers.

To quantify how sensitive the value-added measure is to this heterogeneous effect or the outlier effect, we discuss another three selective replacement approaches in the following part. These methods can be applied in any other scenario where we are concerned about potential outliers' effects on our causal inference. The value-added model is just one example.

(1) Successive extreme replacement: this process is data-dependent and there is no closed formula. We start our replacement from the student who has the lowest gain score in teacher l's class. If the teacher's value-added is still lower than the threshold, then we replace the student with the second lowest gain score. We continue this process until teacher l's value-added achieves the threshold and we record how many students need to be replaced with g_t.

(2) Purposeful sampling process: the lower the student i's gain score is, the probability that this student is selected to be replaced gets higher. This is shown in the following formula:

$$For\ all\ s_{il} < VAM_l, \Pr(s_{il}\ is\ selected\ to\ be\ replaced) = \frac{VAM_l - s_{il}}{\sum(VAM_l - s_{il})}$$

Then the formula for replacement is shown as follows:

$$Thr = \sum_{s_{il}<VAM_l} \left[(g_t - s_{il}) \cdot \frac{VAM_l - s_{il}}{\sum(VAM_l - s_{il})}\right] \cdot \pi + VAM_l$$

From this we can get:

$$\pi = \frac{Thr - VAM_l}{\sum_{s_{il}<VAM_l}\left[(g_t - s_{il}) \cdot \frac{VAM_l - s_{il}}{\sum(VAM_l - s_{il})}\right]}$$

For now we are only replacing students who are below the class average ($for\ all\ s_{il} < VAM_l$). But we can also consider including those students who are above the class average but below the threshold ($Thr > s_{il} > VAM_l$). In this case, the formula will be the following one.

$$\pi = \frac{Thr - VAM_l}{\sum_{s_{il} < Thr} \left[(g_t - s_{il}) \cdot \frac{Thr - s_{il}}{\sum (Thr - s_{il})} \right]}$$

(3) Replace the teacher's median student(s) with grade average students. This approach is proposed due to the fact that median is less sensitive to extreme values than mean. Instead of replacing the mean student gain score in the teacher's class, we select the median student to think about the replacement (Med_l). The formula is represented as follows.

$$\pi = \frac{Thr - VAM_l}{g_t - Med_l}$$

All these four replacement schemes in 4.1 can help with goals (a) and (b). The magnitude of π gives us an intuitive understanding about how far the teacher l is from the threshold if we assume the value-added is a valid and reliable evaluation. It also quantifies how much bias there needs to be to invalidate this evaluation. A small π indicates a lack of robustness or a small departure from the threshold. Similarly, we may get a confidence interval for π by applying the delta method or bootstrap.

Additionally, the three selective replacement schemes can help with the goal (c) by providing a supplemental measure for teacher evaluation. For instance, when two teachers have the same value-added, we can use π from 4.1.2 as another measure for evaluation purpose. The teacher with a smaller π may be more favored because her value added score is more likely to be negatively affected by just a few outlier students.

4.2 Spillover effects – Goal (a)

In the previous section 4.1, we assume that there is no spillover effect among students. This means that when we replace students, those students who remain in the teacher's class will not be affected by the new students coming in. But what if this assumption is violated? This means that in the example in Figure 3, we need to think about whether those original students who are not replaced (black parts) keep the same gain scores after the replacement.

This is actually an essential part in the Stable Unit Treatment Value Assumption (SUTVA) (Rubin, 1986, 1990). If there is spillover effect, then the students' test scores are not only determined by teachers and themselves, but also by the classmates in the same class. Examples of spillover effects include peer effects and some value of mixed class.

To study how sensitive one teacher's value-added is to this spillover effect, we ask the question: how many students should be replaced with other students who have the same gain scores but will bring spillover effects to other students in the class so that the teacher's value-added can achieve the threshold after the replacement?

Specifically, we assume that once we replace students, the change in the teacher's value-added is from the students who stay in the classroom, rather than the students who are brought to this classroom in exchange. Following is the formula for this thought experiment.

$$Thr = (1 - \pi) \cdot (VAM_l + s_{se}) + \pi VAM_l$$

From this we get:

$$\pi = 1 - \frac{Thr - VAM_l}{s_{se}}$$

where

s_{se} is the spillover effect for each student who stays in the class during the replacement. In real application, we can get this information from previous research. Therefore, it is regarded as known here.

From this expression for π, we can tell that when the spillover effect s_{se} is large, then we can replace a large percentage of students with comparable value added score and get the teacher above the threshold because the replacement students trigger large changes in the few students remaining in the class.

One note here is that this formula may seem to tell that the bigger $(Thr - VAM_l)$ is the smaller π is. But this is not necessarily the case. For instance, s_{se} might be a function of π such as $s_{se} = R * n\pi$ $(R > 0)$, where R is the response of a remaining student to a typical new student. With more students having the same background in class, the spillover effect s_{se} for each remaining student gets greater. In this case, we can get: $nR(\pi - \pi^2) = Thr - VAM_l$. This equation can be rewritten as follows, which could be seen as a function of π.

$$Thr - VAM_l = -nR(\pi - 0.5)^2 + 0.25nR$$

Figure 4

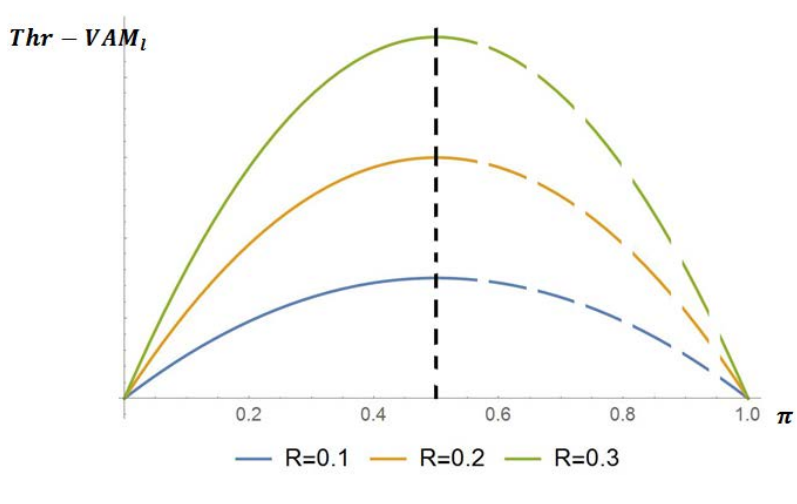


Figure 4 shows this quadratic function of π at different levels of R. The axis of symmetry is $\pi = 0.5$. We just look at situations when $0 < \pi < 0.5$. If $\pi > 0.5$, we can always find another π from (0,0.5) that gets the same $(Thr - VAM_l)$ by symmetry. Additionally, we might say that the teacher needs to replace too many students to invalidate the evaluation if $\pi > 0.5$. Similar conclusions hold for teachers whose VAM score is so low that the distance to the threshold ($Thr - VAM_l$) is larger than the maximum value of this function $0.25nR$. For scenarios where $0 < \pi < 0.5$, the value of π can tell us how sensitive the value-added is to the spillover effect based on s_{se}. The larger the value is, the less sensitive/the more robust this value-added measure is.

4.3 Pairwise exchange – Goals (a) and (b)

Value added measures may also be applied to rank or compare teachers. Suppose there are two teachers l and m. l has a higher value added score than m. Then how robust is the inference that l is a more effective teacher than m? How can we quantify the difference in terms of effectiveness between two teachers?

We can still apply the similar idea. Instead of replacing with grade average students, we can think about how many students need to be exchanged from l's class to m's class to make these two teachers' value added scores equivalent.

From different perspectives, there are several different ways to conduct this thought-experiment specifically. For example, teacher m will argue for randomly reassigning students from these two classes which refers to exchanging l's class average students with m's class average students. Another way that teacher m will favor even more is to exchange l's class best students with m's class worst students. The argument for these two methods is from teacher m's perspective: the reason teacher m has a lower value added score is just because the students she is assigned are not as good as those in teacher l's class. However, teacher l may favor exchanging her mean students with overall average students in these two classes and the argument for this could be randomly reassigning students in these two classes to l's class.

Another issue is about how we should calculate the gain score when we say exchanging "mean" student and this is similar to the discussion before about how to calculate g_t. For example, we can have two different ways to think about exchanging l' mean students with overall average students in these two classes. One simple way is just to exchange VAM_l with $\frac{1}{2}(VAM_l + VAM_m)$. Alternately, we can first find a student in teacher l's class who has the average covariate in the two classes and replace VAM_l with this student's gain score.

5. Prospective: application of value-added

In this part, value-added is applied as a tool for future planning. Because this is not a counterfactual thought experiment anymore, we are replacing teacher l's student with a student from other teachers' classes. We denote this student's gain score as g_c rather than g_t, which represents the average gain score of all the other students in the grade (c is for "complementary"). The way to calculate g_c is the same as that for g_t, except that now we are only averaging across the other students outside teacher l's class.

The main purpose of this section is to provide some strategic suggestions for teachers and administrators. With limited time and resources, how can they achieve better evaluation results through adjusting teacher-student assignment and teaching strategy?

5.1 School administrators: to achieve better school evaluation

(1) Assume that a school administrator plans to adjust teacher assignments to achieve better school evaluation. There are some teachers who are just below the threshold and for some reason (such as classroom observation) the administrator believes the teacher does have a grade average teacher effect. By reassigning students to these teachers, the goal is to have them above the threshold to get a better school evaluation.

In general, we need to replace $n\pi$ students for each teacher (n is the number of students in this teacher's class).

$$\pi = \frac{Thr - VAM_l}{g_c - VAM_l}$$

The aim of this formula is to provide a general idea for how many students each teacher needs to replace to get her above the threshold in the future. In real application, it is possible that very few or even no single student has the gain score of g_c or VAM_l. But as long as the total increase brought by the $n\pi$ students keeps the same, the final result will be the same.

One issue is whether we have enough students to conduct this replacement for all the teachers. If not, then we need to think about from which teacher to start this reassignment. One possibility is to start from the teacher who has the least distance to the threshold (the smallest $(Thr - VAM_l)$). Alternately, we can use our conclusion in the first part under the evaluation framework. We may start from the teacher who needs the smallest number of students to be replaced to achieve the threshold. That is, the teacher who has the smallest π.

(2) Revisit the spillover effect in SUTVA: what is the value of creating a class that can benefit from spillover effects? This is still from the school administrator's perspective. The formula is essentially the same as that in section 4.2. But in this scenario, we focus on the value of creating a new class in which spillover effect will benefit students. The value is the increase in the value-added after the replacement. This increase is from those students who remain in the class, as shown in the following formula.

$$Increase = (1 - \pi) \cdot s_{se}$$

5.2 Teachers: instructional strategy for better future value-added

Assume that teacher l has a below threshold value-added. She hopes to use this value-added to guide her future teaching strategy so that she can improve her value-added in the next round of evaluation. Since this is for future planning, we propose the successive extreme replacement scheme here. Specifically, we start from the lowest gain score student and replace this student with the average gain score of the other students, which is denoted as C_c (C is for class and c is for complementary). We do this for the second, third, fourth lowest gain score student until the teacher gets the threshold. If it turns out that we need to replace the two lowest-score students, then the teacher should spend more time on these two students or she needs to develop a different teaching method for these two students.

Another possible strategy for teacher l is to focus on those students in the bubble: who are just below the proficiency threshold. In fact, there is research showing that in school districts with high-stake accountability system, education resources are reallocated towards these students in the bubble because it is easier for teachers to help these students get an eligible score in tests (Koretz et al., 2016).

6. Discussion: implication of this study from different perspectives

To summarize, there are several goals this paper aims to achieve by applying the idea of replacing students. The first one is to quantify the uncertainty of value-added which is due to potential bias/inconsistencies in an intuitive way. The second is to provide some supplemental measure for teacher evaluation which is also based on the value-added models. If we assume that the value-added is reliable and valid, this paper also shows an intuitive interpretation about how far a teacher is from a certain threshold (or from another teacher) as well as how teachers and school administrators may apply the value-added to plan future teaching more strategically. The approach for all these goals is always to ask the question: how many students need to be replaced to change a value-added score.

These discussions may have different implications from different perspectives. For teachers, we need to recognize that they have the agency to stay or leave the school as well as the agency to change their pedagogies. As Goldhaber illustrated, it is crucial to think about how teachers will react to the accountability system and the incorporation of value added measures. This behavior response is highly dependent on their perceptions on the evaluation systems. The approach discussed in this paper can provide teachers with an intuitive way to interpret, understand as well as applying the value added measures. Recent research has shown that when teachers, especially early career teachers, perceived high evaluation pressure, they tended to move away from enacting ambitious instructions and to only focus on what are valued by the evaluation system (Kim, 2017). The approach provided by this paper can help with this in at least two ways. (1) A better understanding of the value-added measures may help teachers reduce their uncertainty about the evaluation system. (2) This approach provides evaluators with a tool to avoid making high-stake decisions with an obviously invalid value-added score: such as a value-added score which is very close to threshold, indicated by a very small percentage of students needed to be replaced to invalidate the evaluation. This may help reduce teachers' pressure by changing their perceptions in the legitimacy of the evaluation system.

From the school organization perspective, many research have shown bias in value added measures caused by non-random assignment. On one hand, this study provides an approach to quantify this potential bias. On the other hand, we also point out that schools can use the information of value added to achieve better student achievements by reassigning teachers to students strategically. The current study may be too general to lead into real applications directly but this is a direction that deserves more attention. One example like this is the research conducted by Goldhaber, Cowan and

Walch (2013), which studies the correlation in value added scores across subjects and suggests that subject specializing in elementary school may promote student achievement growth.

Back to the debate on whether value-added scores should be applied to inform high-stake personnel decisions, this study points out a way to quantify the potential bias in value added measures and help decision-makers to know how much confidence they can put in the evaluation result based on value added measures. For teachers who are very close to the threshold, it is unfair to base tenure or key decisions just on value-added measures. This study provides an intuitive way to formalize this discourse.

Last but not least, this study provides a way to think about standard error in a more intuitive framework. We started from a discrete, purposeful replacement and once we allow randomness in choosing both replacement student and student to replace with, we are accommodating the randomness as the standard error argues. Specifically, we may think about standard error in two steps. The first is to randomly select a student from the teacher's class and then replace this student with another student randomly chosen from the grade. But further analysis is needed to study whether this process can generate consistent results as the standard error approach.

Table 1. Formulas for different replacement schemes

	Formula for student replacements
5.1.1 Random replacement	$\pi = \frac{Thr - VAM_l}{g_t - VAM_l} = 1 - \frac{g_t - Thr}{g_t - VAM_l}$
5.1.2 (1) Successive extreme replacement	Start replacement from the student who has the lowest gain score in teacher l's class until the teacher achieves the threshold. No closed formula.
5.1.2 (2) Purposeful sampling process	$\pi = \frac{Thr - VAM_l}{\sum_{s_{il} < VAM_l} \left[(g_t - s_{il}) \cdot \frac{VAM_l - s_{il}}{\sum (VAM_l - s_{il})} \right]}$
	$\pi = \frac{Thr - VAM_l}{\sum_{s_{il} < Thr} \left[(g_t - s_{il}) \cdot \frac{Thr - s_{il}}{\sum (Thr - s_{il})} \right]}$
5.1.2 (3) Replace the median student	$\pi = \frac{Thr - VAM_l}{g_t - Med_l}$
5.2 Spillover effects	$\pi = 1 - \frac{Thr - VAM_l}{s_{se}}$

π is the percentage of students need to be changed/replaced,
Thr is the threshold of value-added above which the teacher will be evaluated as eligible,
VAM_l is the value added score for teacher l,
g_t is a grade average student's gain score,
s_{il} is a "purified" gain score for each student i in teacher l's class at time t after removing the effects from those observed characteristics ($A_{i,t-1}$ and X_{it}) included in the value-added model,
Med_l is the median student gain score for teacher l,
s_{se} is the spillover effect for each student who stays in the class during the replacement.
As a clarification, for all these equations, we have $VAM_l < Thr < \overline{VAM}$.

References

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

Chetty, R., Friedman, J. N., & Rockoff, J. E. (2011). *The Long-Term Impacts of Teachers: Teacher Value-Added and Student Outcomes in Adulthood* (Working Paper No. 17699). National Bureau of Economic Research.

Dieterle, S., Guarino, C. M., Reckase, M. D., & Wooldridge, J. M. (2015). How do Principals Assign Students to Teachers? Finding Evidence in Administrative Data and the Implications for Value Added: How do Principals Assign Students to Teachers? *Journal of Policy Analysis and Management*, *34*(1), 32–58.

Frank, K. A., Maroulis, S. J., Duong, M. Q., & Kelcey, B. M. (2013). What Would It Take to Change an Inference? Using Rubin's Causal Model to Interpret the Robustness of Causal Inferences. *Educational Evaluation and Policy Analysis*, *35*(4), 437–460.

Goldhaber, D. (2015). Exploring the potential of value-added performance measures to affect the quality of the teacher workforce. *Educational Researcher*, *44*(2), 87–95.

Goldhaber, D., Cowan, J., & Walch, J. (2013). Is a good elementary teacher always good? Assessing teacher performance estimates across subjects. *Economics of Education Review*, *36*, 216–228.

Goldhaber, D. D., Goldschmidt, P., & Tseng, F. (2013). Teacher Value-Added at the High-School Level: Different Models, Different Answers? *Educational Evaluation and Policy Analysis*, *35*(2), 220–236.

Goldhaber, D., & Hansen, M. (2010). Assessing the Potential of Using Value-Added Estimates of Teacher Job Performance for Making Tenure Decisions. Working Paper 31. *National Center for Analysis of Longitudinal Data in Education Research*.

Gordon, R. J., Kane, T. J., & Staiger, D. (2006). *Identifying effective teachers using performance on the job*. Brookings Institution Washington, DC.

Guarino, C. M., Reckase, M. D., & Wooldridge, J. M. (2014). Can value-added measures of teacher performance be trusted? *Education Finance and Policy*.

Hanushek, E. A., & Rivkin, S. G. (2010). Generalizations about using value-added measures of teacher quality. *The American Economic Review*, *100*(2), 267–271.

Harris, D. N. (2009). Teacher value-added: Don't end the search before it starts. *Journal of Policy Analysis and Management*, *28*(4), 693–699.

Hill, H. C., Kapitula, L., & Umland, K. (2011). A validity argument approach to evaluating teacher value-added scores. *American Educational Research Journal*, *48*(3), 794–831.

Kane, T. J., & Staiger, D. O. (2012). Gathering Feedback for Teaching: Combining High-Quality Observations with Student Surveys and Achievement Gains. Research Paper. MET Project. *Bill & Melinda Gates Foundation*.

Kim, J. (2017). *Quality Matters: The Influence of Teacher Evaluation Policies and School Context on Teaching Quality* (Ph.D.). Michigan State University, United States -- Michigan.

Koretz, D., Jennings, J. L., Ng, H. L., Yu, C., Braslow, D., & Langi, M. (2016). Auditing for score inflation using self-monitoring assessments: Findings from three pilot studies. *Educational Assessment*, *21*(4), 231–247.

Lockwood, J. R., Louis, T. A., & McCaffrey, D. F. (2002). Uncertainty in rank estimation: Implications for value-added modeling accountability systems. *Journal of Educational and Behavioral Statistics*, *27*(3), 255–270.

Lockwood, J. R., McCaffrey, D. F., Hamilton, L. S., Stecher, B., Le, V.-N., & Martinez, J. F. (2007). The sensitivity of value-added teacher effect estimates to different mathematics achievement measures. *Journal of Educational Measurement*, *44*(1), 47–67.

Paufler, N. A., & Amrein-Beardsley, A. (2014). The random assignment of students into elementary classrooms: Implications for value-added analyses and interpretations. *American Educational Research Journal*, *51*(2), 328–362.

Raudenbush, S. W. (2015). Value added: A case study in the mismatch between education research and policy. *Educational Researcher*, *44*(2), 138–141.

Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis methods* (Vol. 1). Sage. R

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

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

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

Winters, M. A., & Cowen, J. M. (2013). Who Would Stay, Who Would Be Dismissed? An Empirical Consideration of Value-Added Teacher Retention Policies. *Educational Researcher*, *42*(6), 330–337.

Winters, M. A., & Cowen, J. M. (2013). Would a Value-Added System of Retention Improve the Distribution of Teacher Quality? A Simulation of Alternative Policies: Value Added Teacher Retention Policies. *Journal of Policy Analysis and Management*, *32*(3), 634–654.