Measuring Student Achievement in the Washington DC IMPACT Evaluation System: A Review

Paper commissioned by the National Academy of Sciences

Cory Koedel

May 2014

1. Introduction

Teachers in Washington, DC Public Schools (DCPS) are evaluated using the IMPACT evaluation system, which was first implemented in 2009. Teachers' IMPACT scores depend in part on direct measures of student achievement. Two types of teacher-level measures are used in IMPACT: (1) value-added and (2) teacher-assessed student achievement (TAS). Teacher value-added is estimated using a statistical model that aims to isolate teacher contributions to student test-score growth on standardized assessments. The standardized assessments given in DCPS are the DC Comprehensive Assessment System (DC-CAS) tests. TAS metrics rely on non-standardized assessments and measure student success as defined by learning goals that are chosen by teachers and approved by administrators.

This paper reviews the estimation of teacher value-added in IMPACT. Value-added is used as part of the IMPACT evaluation for all teachers for whom it can be estimated. Currently teachers who teach math in grades 4-8 and/or English language arts (ELA) in grades 4-10 have value-added scores. For these "group 1" teachers, 35 percent of the total IMPACT score depends on value-added to math and/or ELA standardized tests and 15 percent depends on TAS. Value-added is unavailable for teachers in other grades and subjects. For example, for group-2 teachers, who make up the majority of teachers in DCPS, 15 percent of their total IMPACT scores are based on TAS. The remaining 85 percent of their IMPACT scores are based on evaluation metrics that are not directly tied to student achievement.

2. Teacher Value-Added in IMPACT

2.1 Model Overview

Value-added models (VAMs) are estimated for DCPS teachers based on student performance in math and English language arts (ELA). The most recently estimated model takes the following form (Isenberg and Walsh, 2014):

$$Y_{ticjg} = \lambda_{jg} S_{i(j-1)} + \omega_{jg} O_{i(j-1)} + \boldsymbol{\beta}^{'} \mathbf{X}_{ij} + \boldsymbol{\pi}^{'} \mathbf{C}_{ij} + \boldsymbol{\delta}^{'} \mathbf{T}_{tijg} + \boldsymbol{\theta}^{'} \mathbf{T}_{2tijg} + \varepsilon_{ticjg} \qquad (1)$$

In (1), Y_{ticjg} is a test-score outcome for student *i* taught by teacher *t* in classroom *c* in school-year *j* and grade *g*, $S_{i(j-1)}$ is the prior-year test score for student *i* in the same subject, and $O_{i(j-1)}$ is the prior-year test score in the other subject. So, for example, in the math model $S_{i(j-1)}$ is the lagged math score and $O_{i(j-1)}$ is the lagged ELA score; and vice versa for the ELA model. $\mathbf{X}_{ij}$ and $\mathbf{C}_{ij}$ are vectors of student characteristics and classroom-aggregated student characteristics, respectively. The vector $\mathbf{T}_{tijg}$ includes binary teacher-grade-subject-year indicator variables that track student-teacher assignments – correspondingly, $\boldsymbol{\delta}$ is a vector of parameter estimates that is used to construct the estimated teacher effects and is of primary interest. The vector $\mathbf{T}_{2tijg}$ contains "shadow" teacher-grade-subject-year indicators. These shadow indicators do not have any substantive value. Instead, they are used as a statistical tool to obtain more accurate parameter estimates for the other, non-teacher control variables (see discussion below). Finally, ε_{ticjg} is the error term and represents the residual value of the test score outcome (dependent variable) after accounting for the contributions of the control variables in the model.

2.2 Control Variables

The model in equation (1) controls for a number of student ($\mathbf{X}_{ij}$) and classroom ($\mathbf{C}_{ij}$) characteristics. At the student level controls are included for free-lunch eligibility, reduced-price lunch eligibility, limited-English-proficiency status, learning disabilities, prior-year attendance and student mobility (Isenberg and Walsh, 2014). These controls are similar to the controls that are used in VAMs throughout the research literature (e.g., see Aaronson, Barrow and Sander, 2007; Chetty, Friedman and Rockoff, forthcoming; Goldhaber and Hansen, 2013; Koedel and Betts, 2011; Rivkin, Hanushek and Kain, 2005; Sass et al., 2012), although prior-year attendance is rarely included and its inclusion in the DC IMPACT VAM seems valuable. Two noteworthy omitted variables are (1) race and (2) gender. It is likely that information on race and gender was omitted for political reasons as there is no sound statistical basis for its exclusion. Without access to the data or direct diagnostic information it is difficult to ascertain the importance of the exclusion of these variables from the model. It may be the case that the other control variables largely capture variability in test scores attributable to these omitted variables. It is also notable that Chetty, Friedman and Rockoff (forthcoming), who have access to exceptionally rich data on students from tax records, find little scope for bias in value-added models that rely on sets of control variables similar to those used in the DC IMPACT VAM (also see Kane and Staiger, 2008).[1] Based on the available research evidence,

[1] The Chetty, Friedman and Rockoff (2011) specification differs slightly from the DC IMPACT VAM specification in that the DC IMPACT VAM is estimated in one step while Chetty, Friedman and Rockoff use a two-step procedure. This difference is unlikely to influence the general applicability of the Chetty, Friedman and Rockoff findings.

the likelihood that teachers' value-added estimates are significantly biased due to insufficient student-level control variables in the DC IMPACT VAM is small.

At the classroom level the model includes three control variables: average prior-year test scores in the same subject, the standard deviation of prior-year test scores in the same subject, and the fraction of students eligible for free or reduced-price lunch.[2] The conceptual benefit of the classroom-level controls is that they allow for a more accurate accounting for contextual factors that are outside of the control of teachers. The lagged test-score measures (average and standard deviation) are included to account for peer achievement and the dispersion of achievement within the classroom, and the free/reduced-price lunch control offers additional contextual information. It is not clear how the classroom level controls were selected for inclusion, or why the list of classroom-level controls is shorter than the list of student-level controls. For example, an alternative strategy would have been to include classroom-aggregated elements in $\mathbf{C}_{ij}$ that directly correspond to the elements of $\mathbf{X}_{ij}$. A reason for the discrepancy may be that estimation issues were encountered when trying to specify a fuller $\mathbf{C}_{ij}$ vector, although this is purely speculative as it is not directly addressed in available documentation. As discussed in Sections 4.4 and 5.3 below, the variation that is used to identify the parameter vector $\boldsymbol{\pi}$ is limited in the model, which presents a challenge when attempting to account for classroom context.

2.3 Operational Details

This section discusses a number of operational decisions that were made in estimating the DC IMPACT VAM.

2.3.1 Roster Confirmation

The DC IMPACT VAM is based on linked student-teacher records in relevant subjects. These links are based on student rosters that are created by administrators for teachers in eligible subjects (math grades 4-8; ELA grades 4-10), and confirmed by teachers themselves. In some cases teachers are responsible for creating their own rosters. Teachers report the proportion of time that they teach each student in each subject. Teachers who claim a student for less than 100 percent time are asked to indicate the reason for the reduction. Administrators verify teacher-confirmed rosters at each school and central office staff at DCPS follow up with teachers as necessary.

[2] Due to data limitations, the 2012-2013 model was only able to include a classroom-level measure of free/reduced-price lunch eligibility for students in grade-6 or later (Isenberg and Walsh, 2014).

The roster verification process is thorough and a positive aspect of the DC IMPACT VAM.

2.3.2 Teacher Teams and Dosage

Conceptually, it is useful to think about developing VAMs in a context where each student receives instruction in each subject from a single teacher. However, in reality teaching assignments in K-12 schools can be complex. For example, during the 2012-2013 school year in Washington DC, approximately 26 percent of math teachers and 40 percent of ELA teachers were involved in a student sharing arrangement with at least one other teacher for at least some of their students (Isenberg and Walsh, 2014).

Team teaching, and teacher dosage more generally, is carefully considered in the DC IMPACT VAM. By design, co-taught students contribute to teachers' value-added estimates in the same way as solo-taught students. Put differently, teachers are held fully responsible for both solo- and co-taught students. Operationally, the model equates the weighting by using a unique record for each student-teacher assignment with a single binary variable to indicate the match. Students who are taught by more than one teacher in a given subject have multiple records in the dataset. At the point of estimation, the record corresponding to each student-teacher assignment is weighted to reflect the share of instructional time for that student credited to that teacher. Equal-share team teaching is accommodated by allowing the weights to sum to more than one for individual students. For example, consider a pure team-teaching situation where two teachers jointly co-teach math to a classroom of grade-5 students. Every student in the class would have two records in the math model (one for each teacher assignment), and each record would receive "100 percent" weight (this reflects the purposeful equating of the value of team- and solo-taught students for teachers). If the two teachers did not teach any other students, then they would both receive the same value-added estimate, which would reflect value-added for the two-person team.

The weighting approach to dealing with team teaching in the DC IMPACT VAM is clever and useful. There is a separate technical add-on to the model (the vector of "shadow" teacher assignments, $\mathbf{T}_{2tijg}$) that accommodates the weighted-observations approach without influencing the identification of the other parameters in the model. The adaptation of the model to handle complex student-teacher assignments is a clear advantage of the DC IMPACT VAM.

2.3.3 Test Score Reliability and Standardization of Test Score Outcomes

The DC Comprehensive Assessment System (DC-CAS) tests serve as the achievement assessments in The District. The reliabilities of the tests are high. CTB/McGraw-Hill (2013) reports Cronbach's Alpha coefficients, which are commonly-reported measures of test reliability, for DC-

CAS tests in relevant grades and subjects between 0.90 and 0.93. These coefficients are similar to, if not slightly higher than, reliability coefficients for standardized tests used in other locales.[3] In fact, if the DC-CAS reliability coefficients were any higher it might be viewed as undesirable, as it would raise concerns about a lack of breadth in the test items (Tolmie, Muijs and McAteer, 2011).

The DC-CAS tests are not designed to be equated across grades and/or subjects (CTB/McGraw-Hill, 2013). To facilitate "common metric" comparisons, student test-score outcomes are standardized to have a mean of zero and a standard deviation of one within grades and subjects. This standardization equates the value of equal-distanced scores within each distribution of student scores; e.g., a student who scores one standard deviation above the average in the grade-4 mathematics distribution is treated the same as a student who scores one standard deviation above the average in the grade-6 ELA distribution. This type of standardization is the established norm in the research literature and is properly applied in the DC IMPACT VAM (e.g., see Aaronson, Barrow and Sander, 2007; Chetty, Friedman and Rockoff, forthcoming; Goldhaber and Hansen, 2013; Koedel and Betts, 2011; Rivkin, Hanushek and Kain, 2005; Sass et al., 2012).[4]

2.3.4 Model Flexibility

The DC IMPACT VAM is flexible in allowing for various prior test scores and control variables to differentially predict the current year test-score outcome. Unique coefficients on the lagged test-score variables are estimated for each subject-grade test. This is a useful feature of the DC IMPACT VAM in that it helps to guard against test design or alignment differences across grades and subjects throughout the DC-CAS tests.

The coefficients corresponding to the student- and classroom-level characteristics are allowed to vary by subject and grade span – i.e., elementary school (grades 4, 5), middle school (grades 6, 7, 8) and high school (grades 9, 10). This approach represents a compromise between competing objectives. On the one hand, pooling the data across all grades and subjects allows for the most precision in identifying these parameters. On the other hand, estimating the coefficients separately for each subject and grade would maximize the flexibility of the model in terms of

[3] For example, see Arizona Department of Education (2011), CTB/McGraw-Hill (2012), Massachusetts Department of Elementary and Secondary Education (2012). It should be noted that there is some disagreement in the research community regarding the value of the Cronbach's Alpha coefficient as a measure of reliability; for a critical view, see Sijtsma (2009). I do not examine the criticisms of this or other common-practice reliability measures in this paper, but simply note that based on commonly reported measures of reliability available in technical reports, the DC-CAS tests compare well with other standardized tests.

[4] Even if DCPS used a test for which a vertical scale was attempted, it is unclear whether this would be valuable (Ballou, 2009). Furthermore, if a test were effectively vertically scaled within a subject across grades, comparing teachers across subjects would remain problematic and likely require a norming procedure similar to what is implemented in the DC IMPACT VAM.

allowing the control variables to have differential effects on different tests. The DC IMPACT VAM strikes a fair balance in weighing these competing objectives.

2.3.5 Measurement Error Correction

The DC IMPACT VAM uses a classical errors-in-variables (CEV) correction to account for measurement error in pretest scores. Although this is a common procedure, it should be noted that the assumptions that underlie this correction are not consistent with the test measurement error properties of most standardized tests (Boyd et al., 2013; Koedel, Leatherman and Parsons, 2012; Lockwood and McCaffrey, 2014), including the DC-CAS tests (CTB/McGraw-Hill, 2013). More specifically, the CEV correction assumes that the measurement error is homoscedastic, but in reality the measurement error on DC-CAS tests is heteroskedastic with the error variance being much larger in the tails of the distribution. Boyd et al. (2013), drawing on a longstanding literature (e.g., see Thorndike, 1951), further note that measurement error attributable to the testing instrument itself represents just one aspect of total measurement error. There is no "industry standard" approach to dealing with test measurement error in value-added models at present. As noted by Lockwood and McCaffrey (2014), this issue is "regularly overlooked" (p. 22) in the context of VAM estimation.

Although the treatment of test measurement error in the DC IMPACT VAM is conceptually unappealing, it should be noted that the consequences may be small. For example, despite the fact that the conditions for the CEV correction are not met in standard VAMs, Koedel, Leatherman and Parsons (2012) show that teacher value-added rankings are only moderately affected by applying the inappropriate correction. That said, it is fair to question the value of the approach to dealing with test measurement error in the DC IMPACT VAM, and to be concerned that as presently implemented it is doing more harm than good.[5]

One simple solution that may improve model performance would be to avoid making a measurement error correction at all, and instead rely on the inclusion of more control variables in the model to reduce the impact of measurement error. Lockwood and McCaffrey (2014) note that student covariates aggregated to the teacher level may be valuable in this regard.[6] The DC IMPACT VAM already includes classroom-level controls, which are likely to sufficiently approximate teacher-

[5] The errors in the value-added scores for some DC IMPACT teachers, which were popularized in the media (e.g., see Anderson, 2013; Strauss, 2013), were caused because the initial run of the model did not initialize the code that performed the measurement-error correction (this information was obtained in correspondence with researchers at Mathematica Policy Research).

[6] Lockwood and McCaffrey (2014) also discuss other solutions to the measurement error problem. However, all of the direct methods for accounting for test measurement error that they propose depend on assumptions that "require careful scrutiny" (p. 42).

level controls. However, in pointing to the value of using these aggregated variables to reduce the influence of test measurement error, Lockwood and McCaffrey (2014) cite Chetty, Friedman and Rockoff (forthcoming) and Kane et al. (2013), both of which use VAMs that have a different fundamental structure that is better suited to facilitate the identification of the parameters on the aggregated variables. The extent to which the modeling-structure issue will limit the value of including aggregated variables into the DC IMPACT VAM is not clear. I return to this point in Section 5.3.

2.3.6 Sequential Estimation

Although it is useful conceptually to think of the DC IMPACT VAM as being estimated in a single step, operationally the model is estimated sequentially. There are several reasons for this. A key reason is that the model aims to estimate the separate contributions of classroom characteristics to student achievement, which requires the use of multiple years of data, but the teacher value-added measures of interest are based on performance only in the most recent year (I discuss the use of single-year performance measures in IMPACT below). For the sake of brevity I avoid a lengthy discussion of the sequential estimation approach here and simply note that it is reasonable.

3. Using Model Output to Recover Teacher Value-Added Estimates

3.1 Constructing Singular Teacher Effect Estimates

As noted above, the DC IMPACT VAM produces estimates of teacher-grade-subject effects.[7] To construct a singular VAM measure for each teacher, the grade-subject estimates are averaged for teachers. This is facilitated by first performing an *ex post* normalization of the teacher effects estimated for each grade and subject so that they are comparable in the distributional sense. In particular, after the estimated grade-subject teacher effects are obtained, they are adjusted for each teacher *t* in each grade *g* and subject *k* as follows:

$$\hat{\eta}_{tgk} = \frac{(\hat{\delta}_{tgk} - \overline{\hat{\delta}}_{gk})}{\hat{\sigma}_{gk}} \qquad (2)$$

In equation (2), $\overline{\hat{\delta}}_{gk}$ and $\hat{\sigma}_{gk}$ are the mean and standard deviation of the estimated teacher effects in subject *g* and grade *k*, respectively.

[7] Technically, the estimates are also year specific but there is no variation in the data over years. Put differently, value added estimates for all teachers are based on data from the most recent year only. For ease of presentation I drop the year subscript in the discussion.

The normalized estimates $\hat{\eta}_{tgk}$ are combined across grades, within subjects, using weighted averages.[8] For example, if a math teacher teaches 20 students in grade-7 and 60 students in grade-8, 25 percent of her final value-added score in math will be based on her grade-7 estimate and 75 percent on her grade-8 estimate. For a teacher who teaches a single subject (math or ELA), the weighted average of the grade-level effects is used as the VAM estimate. For teachers who teach both subjects (e.g., self-contained elementary-school teachers), a simple average of the subject-specific estimates is used.

The normalization procedure is effective in facilitating the aggregation of teachers' grade-subject estimates into single effectiveness measures.

3.2 Teacher Inclusion Criteria

A fundamental problem with teacher evaluation is that the structure of schooling means that many teachers are observed with relatively few students, at least in the statistical sense. This limitation is not unique to value-added – it applies for any evaluation metric for teachers. In the DC IMPACT VAM, like value-added models used throughout the research literature and in policy applications elsewhere, "inclusion restrictions" based on minimum required teacher-level sample sizes are used to limit mistakes that may occur when teachers are evaluated using a small number of students.

To receive a value-added estimate in each subject in IMPACT, a teacher must teach at least 15 students in that subject. This is a reasonable threshold and is consistent with prior research (e.g., see Aaronson, Barrow and Sander, 2007). The total student threshold can be met by combining students across grades. However, for a teacher's grade-subject effect to "initialize," a separate threshold of at least seven students in the specific grade-subject combination must be met. This separate "initialization threshold" at the teacher-grade-subject level is imposed to mitigate the influence of wild parameter estimates that may result from attempting to estimate model parameters based on a very small number of students (given that the estimates that come directly out of the model are for teacher-grade-subject effects).

This tiered approach to inclusion is imperfect and can create inclusion inconsistencies. For example, consider a math teacher who teaches eight students in grade-6 and eight students in grade-7. Her effects will "initialize" in both grades and thus her student count will exceed the 15-student threshold. This teacher will receive a value-added score. However, an otherwise similar teacher who

[8] The normalized values in equation (2) are multiplied by a constant to improve interpretability but this has no substantive bearing on the results (see Isenberg and Walsh, 2014).

teaches ten students in grade-6 and six students in grade-7 will not receive a value-added score. This is because the teacher's grade-7 effect will not initialize, which means that her grade-7 students will not be counted toward her total. In fact, these grade-7 students are not incorporated into the value-added model at all.

The problem described in the previous paragraph is not simply limited to teacher inclusion. For example, suppose the latter teacher had eighteen students in grade-6 and six students in grade-7. It is still the case that her grade-7 effect will not initialize, but now she will receive a value-added score because she has enough grade-6 students to qualify. However, the performance of her grade-7 students, who make up 25 percent of her workload, is not incorporated into her value-added score.

To be clear, the inclusion issue discussed in this section is unlikely to be substantively important for most teachers (although this cannot be directly inferred from available documentation). Put differently, few teachers are likely to be in a situation where they teach a very small number of students in a particular subject-grade. However, options for adjusting the inclusion-restriction criteria in a way that creates more consistent inclusion rules should be considered. One possibility is to build a sufficiently rich interacted VAM that facilitates the direct estimation of total teacher value-added, bypassing the teacher-subject-grade specific estimates. A limitation of doing this is that it would complicate the grade-subject alignment procedure discussed in Section 3.1 (because some teachers' total value-added estimates will be the product of performance in different grades and/or subjects), but it may be possible to develop a modified alignment procedure that will function with output from such a model.

4. Key Changes to the DC IMPACT VAM Over Time and Their Implications

Thus far this paper has reviewed the most recent iteration of the DC IMPACT VAM. Like with any new technology, the current version of the model reflects a cumulative development process. Since DC IMPACT was first implemented the model has undergone a number of changes. On the whole, these changes have improved the quality of the model. In this section I briefly discuss some of the more interesting and important changes. The changes that I discuss, along with some other changes that I omit for brevity, are also covered in Isenberg and Walsh (2014).[9]

4.1 The Use of School-Level Value-Added

The most substantial change to how value-added is used in DC IMPACT has nothing to do with specification details related to the model. Absent from the current evaluation structure is the application of any school-level measure of value-added. In previous iterations of IMPACT, five

[9] With the exception of the school-level value-added issue discussed in Section 4.1.

percent of each teacher's IMPACT score, regardless of which subject and/or grade was taught, depended on a school-level value-added measure (Dee and Wyckoff, 2013).[10]

It is not clear from available documentation why school-level value-added was removed as a component of teachers' IMPACT evaluations. There are a number of candidate explanations. One is that many teachers were assigned a school-level value-added score that was very far removed from their own work (e.g., a grade-11 history teacher, grade-1 teacher, etc.).

While school-level value-added was never a major component of DC IMPACT, its absence from the current formula does merit consideration. The use of school-level metrics in teacher evaluation systems has been advocated in previous research (Ahn and Vigdor, 2011), and there are well-documented productivity benefits associated with team incentives (although the best evidence on the value of team incentives comes from outside of the education context – e.g., see Hamilton, Nickerson and Owan, 2003). It may be that school-level value-added is not the best metric to use to incorporate team incentives. However, in future iterations of IMPACT, and in other systems nationwide, the question of whether a school-level or alternative team-level component to the incentive structure is desirable should be considered.

4.2 *Model Inclusiveness*

The current DC IMPACT VAM covers more teachers than earlier iterations – specifically, the current model produces value-added scores for ELA teachers in grades 9 and 10 (in addition to math and ELA teachers in grades 4-8), whereas previous versions did not. This is a useful step and future efforts should continue to work to incorporate value-added into the IMPACT scores for more teachers. One reason that this is important is that for non-value-added teachers, just 15 percent of the total IMPACT scores are currently based on a direct measure of student achievement (TAS).

4.3 *Accounting for Student Mobility and Improved Accounting for Student Poverty Status*

The current DC IMPACT VAM is an improvement over earlier versions in terms of controlling for student characteristics that are likely to influence outcomes. In particular, the 2012-2013 version of the model includes a direct control to allow for differential test-score growth for students who transfer schools mid-year. Mid-year school transfers are a disruptive event, and accounting for transfers in the model is useful to avoid the misattribution of any adverse consequences to teachers.

[10] Isenberg and Hock (2012) provide a technical discussion of the school-level model used in IMPACT in 2011-2012.

The current VAM also uses a more thorough procedure for accounting for student poverty status than in previous versions. The new version of the DC IMPACT VAM uses data on poverty status for up to four school years for each student, whereas in previous versions the model used poverty-status data only from the current year (this is how VAMs estimated in research typically operate). This change appears to have been prompted in part by the inclusion of an increasing number of students and teachers into the model from schools that do not collect poverty information for students. The poverty information from previous years is particularly valuable for students at these schools, who may have poverty information from prior years when they attended a different school. However, in addition to being valuable for these students, the improved poverty measures are likely of value for predicting test-score performance for all students because the data from additional years improves the accuracy of the poverty information.

4.4 Accounting for Classroom Characteristics

Unlike its predecessors, the current DC IMPACT VAM accounts for classroom characteristics in addition to student-level characteristics. The rationale of including direct classroom-level controls is to better distinguish teachers' contributions to student test scores from contextual classroom information. As noted above, the DC IMPACT VAM controls for classroom-averaged prior test scores in the same subject, the standard deviation of prior year scores in the same subject, and classroom level poverty status (the latter measure is used only in grades 6 and above due to data limitations).

The spirit behind the inclusion of these controls is in the right place, and their inclusion likely improves the performance of the model relative to previous versions. However, it is important to recognize that these controls may not be achieving their full intended purpose, in part because of the modeling structure. A concern is whether the available within-teacher variation in classroom-level characteristics is sufficient to properly identify the parameters in the vector $\boldsymbol{\pi}$ from equation (1). It is beyond the scope of this review to delve deeply into technical detail on this issue, but in short, the model leverages variation across years and/or classrooms within years for individual teachers to identify these parameters. However, this raises concerns about extrapolating from small differences across classrooms within teachers to potentially much larger differences across teachers (including teachers who teach in different schools). The within-teacher identification strategy also exacerbates concerns about the attenuating effect of measurement error in the classroom-level variables. Ehlert et al. (forthcoming) provide a detailed discussion of these issues in the VAM context.

Notable recent studies by Chetty, Friedman and Rockoff (forthcoming) and Kane et al. (2013) do not rely solely on within-teacher variation to identify the other parameters in the model. This choice comes with a tradeoff: these studies may conflate teacher quality with student and classroom characteristics if teacher quality is systematically related to these characteristics. Unfortunately no clearly preferred method for accounting for contextual, group-level factors in VAMs has been established. The approach taken by the DC IMPACT VAM is reasonable and has some precedent in the research literature, but it has limitations.

5. Topics for Consideration in the Future Development of VAMs to be Used to Inform Teacher Evaluation Systems

This section discusses conceptual issues that merit attention for the future development of IMPACT and similar systems.

5.1 Should VAM estimates be based on multiple years of data?

Value-added for teachers in Washington DC is estimated based entirely on performance in the most-recent year. This approach has strong conceptual appeal – it ensures that teacher evaluations in IMPACT are based on current performance measures. However, given that value-added is a statistical construct and thus subject to estimation error, there are benefits to including multiple years of data.[11] For example, research shows that using multiple years of value-added data reduces estimation error and the influence of transitory student sorting bias, thereby improving inference (Koedel and Betts, 2011; McCaffrey et al., 2009; Schochet and Chiang, 2013).

The tradeoff is as follows: VAMs can be used to produce either (1) a less-reliable measure of value-added that depends entirely on performance in the most recent year, or (2) a more-reliable measure that depends on performance during the most recent several years. DCPS has elected for the former, and although neither approach is clearly preferred, this is an important conceptual aspect of system design that merits careful consideration. Alternatives to the DC IMPACT approach include using two or three years of value-added data to produce a "moving average" performance metric for teachers, or a similar approach that down-weights older value-added data but uses weights larger than zero.

Again, there is no clearly preferred approach to this problem, but the decision of how best to use historical value-added data represents an important aspect of system design.

[11] This is also true for non-value-added measures. Although the research literature on the statistical properties of non-value-added performance measures for teachers is considerably less developed than the value-added literature, *all of the DC IMPACT performance measures are estimated with error.*

5.2 *Should Teacher Evaluation Systems Incorporate School-Level Value-Added or Alternative Team-Level Performance Measures?*

See Section 4.1.

5.3 *What is the Best Way to Structure Value-Added Models for Use in Teacher Evaluation Systems?*

The modeling structure used to estimate teacher value-added in recent studies by Chetty, Friedman and Rockoff (forthcoming) and Kane et al. (2013) differs from the approach taken in the DC IMPACT VAM. Without delving too deeply into the statistical details, the approach in the former studies uses variation within and across teachers to identify the coefficients on the control variables in the model, whereas the DC IMPACT VAM uses only within-teacher variation to identify the coefficients on these parameters. Ehlert et al. (2014, forthcoming) describe the class of models to which the DC IMPACT VAM belongs as "one-step" VAMs and the class of models to which the Chetty, Friedman and Rockoff and Kane et al. models belong as "two-step" VAMs. Ehlert et al. (forthcoming) raise concerns about relying entirely on within-teacher variation to identify the other parameters in value-added models, as is done in one-step VAMs. As mentioned in Section 4.4, there are reasons to expect that the control variable coefficients in the DC IMPACT VAM, and in particular those that correspond to the classroom-level control variables, are attenuated.

In addition, Ehlert et al. look beyond purely statistical considerations in their examination of modeling structure, and discuss how well different value-added approaches satisfy the likely policy objectives of teacher evaluation systems. They argue that a model along the lines of that used by Chetty, Friedman and Rockoff and Kane et al. is better-suited to achieve policy objectives than available alternatives, including the one-step VAM used in IMPACT.

No consensus has emerged in the research literature regarding the choice of a one-step or two-step VAM for use in teacher evaluation systems. Although the recent above-referenced studies use a two-step model, the one-step model remains the predominant model in the research literature to date. Although a consensus has not emerged, the tradeoffs between modeling structures have been well-documented and this is another important aspect of system design that merits careful attention.

6. Summary Assessment of the DC IMPACT VAM

The technical aspects of the DC IMPACT VAM are generally well-supported by available research evidence. My summary assessment of the model is quite positive. Here are some aspects of the model that I reviewed favorably:

1. The roster confirmation process for teachers seems thorough and likely reduces errors in value-added estimation.
2. Available research suggests that the control variables in the model are adequate for reducing bias in estimated teacher value added to a likely negligible level (Chetty, Friedman and Rockoff, forthcoming; Kane and Staiger, 2008). The additional flexibility in the model above and beyond standard models, which allows for key predictor variables to have differential effects on tests in different grades and subjects, is a strength of the DC IMPACT VAM.
3. The standardization and normalization procedures used to facilitate teacher comparisons across grades and subjects are effective.
4. The treatment of teacher teams, and teacher dosage more generally, is clever and a valuable innovation for the practical application of VAMs for teacher evaluation.
5. With the possible exception of the complete removal of school-level value added measures from the DC IMPACT evaluation system, which was likely an administrative decision unrelated to modeling issues, it is my assessment that the DC IMPACT VAM has improved over time, which is an encouraging albeit expected outcome (like with any new technology).

During my review I noted two technical dimensions along which the DC IMPACT VAM can be improved, although the substantive gains from improvement in both cases will likely be modest:

1. Empirical evidence does not support the notion that the measurement error correction used in the DC IMPACT VAM leads to improved inference. The properties of test measurement error on standardized tests, and in particular its heteroskedasticity, are not consistent with the correction that is applied in the model (which assumes homoscedastic measurement error). While it is not clear that the correction is doing significant harm, it is also not clear that it is helpful.
2. The question of how to handle teachers who teach relatively few students is one that must be addressed in any VAM that is used for high-stakes teacher evaluation. The number of students required for teachers to receive a value-added score in each subject in IMPACT – 15 – is reasonable. However, it may be possible to increase the number of teachers that meet this total student threshold by adjusting how the model is estimated. In particular, if teacher-grade-subject effects could be consolidated during the estimation procedure, some teachers

who teach small numbers of students across multiple grades who do not currently receive a value-added score could receive one.

Finally, I reiterate three conceptual issues that merit future consideration for IMPACT and other developing evaluation systems:

1. There is a tradeoff between the recency of information contained in teachers' value-added estimates and their accuracy. Teacher value-added in IMPACT is estimated using data from a single year. Using a single year of data is appealing in that performance in previous years does not influence a teacher's current evaluation score. However, bringing in data from multiple years improves the precision of value-added estimates. It may be optimal to use multiple years of value-added data for teachers when possible, perhaps unequally weighted so that recent performance is emphasized, to improve the accuracy of teachers' value-added estimates.
2. The absence of a school-level value added measure, or alternative "team" performance measure, is a weakness of IMPACT. Available research points to team incentives as being generally effective in encouraging worker effort. Note that in earlier iterations of IMPACT, five percent of the IMPACT score for all teachers was determined based on school-level value added.
3. The "one-step" value-added model used in IMPACT is based on the predominant modeling structure in the research literature. However, several recent studies have adopted an alternative "two-step" or "averaged residuals" approach. The two-step approach offers some benefits in terms of estimation and targeting key policy objectives, but there are tradeoffs. Neither the one-step or two-step approach is clearly preferred, but as these systems continue to evolve the question of what modeling structure is most useful for achieving stated system objectives merits careful consideration.

References

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

Ahn, Thomas and Jacob L. Vigdor. 2011. Making Teacher Incentives Work: Lessons from North Carolina's Teacher Bonus Program. *Education Outlook* No. 5 (June). American Enterprise Institute for Public Policy Research.

Anderson, Nick. 2013. D.C. Schools Gave 44 Teachers Mistaken Job Evaluations. *The Washington Post* (12.23.2013).

Arizona Department of Education. 2011. Arizona's Instrument to Measure Standards: 2011 Technical Report.

Ballou, Dale. 2009. Test Scaling and Value-Added Measurement. *Education Finance and Policy* 4(4), 351-383.

Boyd, Donald, Hamilton Lankford, Susanna Loeb and James Wyckoff. 2013. Measuring Test Measurement Error: A General Approach. *Journal of Educational and Behavioral Statistics* 38(6), 629-663.

Chetty, Raj, John N. Friedman and Jonah E. Rockoff (forthcoming). Measuring the Impacts of Teachers I: Evaluating Bias in Teacher Value-Added Estimates. *American Economic Review*.

CTB/McGraw-Hill Education. 2012. Missouri Assessment Program Grade-Level Assessments: Technical Report 2012, Final. CTB/McGraw-Hill: Monterey, California.

CTB/McGraw-Hill Education. 2013. Washington, D.C. Comprehensive Assessment System (DC CAS): Technical Report, Spring 2013 Test Administration. CTB/McGraw-Hill: Monterey, California.

Dee, Thomas and James Wyckoff. 2013. Incentives, Selection and Teacher Performance. Evidence from IMPACT. NBER Working Paper No. 19529.

Ehlert, Mark, Cory Koedel, Eric Parsons and Michael Podgursky. 2014. Choosing the Right Growth Measure: Methods Should Compare Similar Schools and Teachers. *Education Next* 14(2): 66-71.

Ehlert, Mark, Cory Koedel, Eric Parsons and Michael Podgursky (forthcoming). Selecting Growth Measures for Use in School Evaluation Systems: Should Proportionality Matter? *Educational Policy*.

Goldhaber, Dan and Michael Hansen. 2013. Is It Just a Bad Class? Assessing the Stability of Measured Teacher Performance. *Economica* 80(319), 589-612.

Hamilton, Barton, Jack A. Nickerson and Hideo Owan. 2003. Team Incentives and Worker Heterogeneity: An Empirical Analysis of the Impact of Teams on Productivity and Participation. *Journal of Political Economy* 111(3), 465-497.

Isenberg, Eric and Heinrich Hock. 2010. Measuring School and Teacher Value Added for IMPACT and TEAM in DC Public Schools: Final Report. Mathematica Policy Research (08.20.2010).

Isenberg, Eric and Heinrich Hock. 2012. Measuring School and Teacher Value Added in DC, 2011-2012 School Year: Final Report. Mathematica Policy Research (08.31.2012).

Isenberg, Eric and Elias Walsh. 2013. Measuring School and Teacher Value Added in DC, 2012-2013 School Year: Final Report. Mathematica Policy Research (09.11.2013).

Isenberg, Eric and Elias Walsh. 2014. Measuring School and Teacher Value Added in DC, 2012-2013 School Year: Final Report. Mathematica Policy Research (01.17.2014).

Kane, Thomas, McCaffrey, Daniel, Trey Miller and Douglas Staiger. 2013. Have We Identified Effective Teachers? Validating Measures of Effective Teaching Using Random Assignment. Bill and Melinda Gates Foundation MET Project Research Paper.

Kane, Tom J. and Douglas O. Staiger. 2008. Estimating Teacher Impacts on Student Achievement: An Experimental Evaluation. NBER Working Paper No. 14607.

Koedel, Cory and Julian R. Betts. 2011. Does Student Sorting Invalidate Value-Added Models of Teacher Effectiveness? An Extended Analysis of the Rothstein Critique. *Education Finance and Policy* 6(1), 18-42.

Koedel, Cory, Rebecca Leatherman and Eric Parsons. 2012. Test Measurement Error and Inference from Value-Added Models. *The B.E. Journal of Economic Analysis & Policy* 12(1).

Lockwood, J.R. and Daniel F. McCaffrey. 2014. Correcting for Test Score Measurement Error in ANCOVA Models for Estimating Treatment Effects. *Journal of Educational and Behavioral Statistics* 39(1), 22-52.

Massachusetts Department of Elementary and Secondary Education. 2012. 2012 MCAS and MCAS-Alt Technical Report.

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

Rivkin, Steven G., Eric A. Hanushek and John F. Kain. 2005. Teachers, Schools and Academic Achievement. *Econometrica* 73(2), 417-58.

Sass, Tim R., Jane Hannaway, Zeyu Xu, David N. Figlio and Li Feng. 2012. Value Added of Teachers in High-Poverty Schools and Lower Poverty Schools. *Journal of Urban Economics* 72, 104-122.

Schochet, Peter Z. and Hanley S. Chiang. 2013. What are Error Rates for Classifying Teacher and School Performance Measures Using Value-Added Models? *Journal of Educational and Behavioral Statistics* 38(2), 142-171.

Sijtsma, Klaas. 2009. On the Use, the Misuse, and the Very Limited Usefulness of Cronbach's Alpha. *Psychometrika* 74(1), 107-120.

Strauss, Valerie. Errors Found in D.C. Teacher Evaluations (2nd Update). 2013. *The Washington Post* (12.23.2013)

Thorndike, Robert L. 1951. Reliability. In E. F. Lindquist (Ed.), *Educational Measurement,* pp.560–620. Washington, DC: American Council on Education.

Tolmie, Andy, Daniel Muijs and Erica McAteer. 2011. Quantitative Methods in Educational and Social Research Using SPSS. McGraw-Hill Education: New York, NY.