Abstract Title Page
Not included in page count.

Title: Using Test Scores from Students with Disabilities in Teacher Effectiveness Indicators

Authors and Affiliations:

Heather M. Buzick
Educational Testing Service
Email: hbuzick@ets.org

Nathan D. Jones
Boston University
Email: ndjones@bu.edu

Background / Context:

The increased emphasis on using student growth measures in teacher evaluation has raised questions about how to treat test scores from students with disabilities. One important question is how to ensure that, when these students' scores are incorporated into estimates of teacher effects, they lead to valid interpretations about teacher effectiveness (Jones, Buzick, & Turkan, 2013; Warren, Thurlow, Lazarus, Christensen, Chartrand, & Rieke, 2012). As outlined by Buzick and Laitusis (2010), several challenges emerge when measuring growth for students with disabilities. Students with disabilities frequently perform at the low end of the scoring distribution (Center on Education Policy, 2009; Wu, Liu, Thurlow, Lazarus, Altman, & Christian, 2012). In some cases, their disabilities create barriers to accessing item content, raising validity concerns about interpretations based on their scores. Lastly, testing accommodations, which are intended to improve the accessibility of standardized tests for students with disabilities, can impact test performance (e.g., Sireci, Scarpeti, & Li, 2005); consequently, inconsistent accommodation use across years may increase or decrease measured growth regardless of the teacher's inputs. Research has not examined the policy and practical implications of these challenges, despite the fact that these threats to validity are relevant for special educators and general educators alike.

In the context of teacher evaluation, threats to validity associated with test scores from students with disabilities may undermine the credibility of teacher effectiveness indicators based on student growth. On the other hand, excluding students with disabilities in indicators of teacher effectiveness can have unintended consequences on teaching (e.g., lack of differentiated instruction or lack of incentives to improve instructional quality for students with disabilities). A number of studies have examined the sensitivity of estimated teachers effects to various model specifications (Ballou, Sanders, & Wright, 2004; Ehlert, Koedel, Parsons, & Podgursky, 2012; Goldhaber, Walch, & Gabele, 2012; Lockwood et al., 2007; McCaffrey, Lockwood, Koretz, Louis, & Hamilton, 2004; Wright, 2010). Despite the contribution of these studies, none have looked specifically at how model results vary depending on the way that scores from students with disabilities are treated. In fact, in research studies, a common convention is to exclude test scores students with disabilities from the models entirely, given potential threats to validity.

Therefore, in the current study, we explore practical questions related to including scores from students with disabilities in statistical approaches to estimating teacher effectiveness, namely

Purpose / Objective / Research Question / Focus of Study:

In this study, we explore the consequences of three common approaches for treating scores from students with disabilities: a) including scores from students with disabilities with no additional disability-related covariates, b) including scores from students with disability as well as disability-related covariates, and c) removing these students entirely when estimating teachers' value-added scores. We estimated a series of value-added models (using teacher fixed effects), with each model adding additional covariates. We estimated all models with and without scores from students with disabilities; and, we examined whether model choice functioned differently depending on the proportion of students with disabilities in the classroom. We chose several different metrics to understand how and for whom model choice might matter, including Pearson correlation coefficients, percentile ranks, and percent movement across quintiles.

Setting:

For this study, we drew on a single state's administrative data over the period of 2007-2009, which provided us with information about all students in grades 3-5 who took the state general summative assessment; vertically-scaled student test scores were available in reading and mathematics. In our sample we included data from classes of between 10 and 30 students. 30 is the maximum class size for grade 5 in the state; we considered teachers linked with more than 30 students to have served as test proctors and excluded them from the analysis. Like many states, we put a lower bound on the number of students required for calculating teacher scores to improve the accuracy of the measures. We used 10 students as our cut off to include in the sample more teachers of students with disabilities.

Population / Participants / Subjects:

To prepare the sample for analysis, we created both a single-cohort sample and a two-cohort sample. Descriptive information related to both grades appears in the Results section. The original dataset had 73,276 student records for grade 5 students in 2009. Twelve percent of the students had a disability and 3% were English learners (ELs; 0.28% were ELs with a disability). Among the 3,894 uniquely identified grade 5 teachers of record in 2009, we considered 83% of them as valid teachers. We defined valid teachers as those linked to between 10 and 30 students. The final analysis sample for grade 5 students in 2009 included 3,189 teachers, 61,091 students with current and prior reading scores, and 61,139 students with current and prior math scores. Ten percent of students in the final grade 5 sample were students with disabilities. Two-thirds of the teachers had at least one student with a disability in their classroom. Among those, general education teachers had an average of 2 to 3 students with disabilities (13% of the class).

Intervention / Program / Practice:

Existing research has also provided little information on the consequences (intended or unintended) of including growth scores from students with disabilities in estimates of teachers' effectiveness, and there is currently no common standard for treating test scores from students with disabilities. There are three different approaches that can be used: a) including scores from students with disabilities with no additional disability-related covariates, b) including scores from students with disability as well as disability-related covariates, and c) removing these students entirely when estimating teachers' value-added scores. While controlling for student disability is an increasingly common approach used in practice, there is no existing research examining the consequences of this decision for teacher effectiveness scores.

Research Design:

Our goal was to describe if and how teachers' scores would change depending on model choice and whether or not scores from students with disabilities were included in the estimation. To accomplish this, we estimated teacher scores with several different VAMs and also median SGPs, using math and reading as separate outcomes. We chose a set of models that are relatively simple in terms of estimation, given that this choice is often made in practice so that methods are cost-effective, transparent and straightforward to explain to stakeholders. We estimated all models both with and without test scores from students with disabilities. To understand how model choice is related to the number of students with disabilities in the classroom, in presenting our results we divided teachers into three categories: teachers linked to no students with disabilities ($n \frown 900$), teachers linked to some students with disabilities ($n \frown 2{,}000$), and

teachers linked to all students with disabilities (n ⁀ 65)[i]. We chose several different metrics to understand how and for whom model choice might matter; these included Pearson correlation coefficients, average percentile ranks, and percent movement across quintiles. We also report the correlations between teachers' scores and aggregate covariates—school poverty, classroom percent of students with disabilities, and average classroom prior same-subject achievement.

Data Collection and Analysis:

We estimated a series of VAMs (using teacher fixed effects), with each successive model adding an additional set of conditioning variables. The general form for the models for student *i*, in grade *g*, taught by teacher *j,* in year *t* is

$$y_{ijt,g} = \beta_0 + y'_{i(t-1),g-1}\beta_1 + u'_{jt,g}\beta_2 + x'_{ijt,g}\beta_3 + z'_{ijt,g}\beta_4 + \psi_{j,g} + \varepsilon_{ijt,g},$$

where $y_{ijt,5}$ denotes a vector of grade *g* standardized achievement scores, β_0 is the grand mean across all teachers, $y'_{i(t-1),g-1}$ denotes a vector of prior standardized achievement scores, $u'_{jt,g}$ is a vector of classroom characteristics associated with teacher *j* at time *t*, x'_{ijtg} is a vector of general student background characteristics, $z'_{ijt,g}$ is a vector of disability-specific student variables, $\psi_{j,g}$ is the teacher effect for $g = (5^{th}, 4^{th})$ grade teachers, and $\varepsilon_{ijt,g}$ is the *iid* error term with mean equal to 0. We standardized current and prior scores within grade level across all students in each analysis sample. We used the *fese* command in Stata (Nichols, 2008) to estimate the models. The specific variables included in each model are described in Table 1. Note that without scores from students with disabilities, VAM2B is equivalent to VAM2A and the disability-related covariates in VAM3 are omitted.

Findings / Results:

Including vs. excluding scores from students with disabilities

The Pearson correlation coefficients comparing teacher scores when students with disabilities were included and excluded ranged from 0.97 to 0.98 in reading and 0.98 to 0.99 in math across all models estimating one- and two-year effects (the SGP and each of the four VAM models) in both grades. For teachers in classrooms with no students with disabilities, the correlations were all 0.99. For teachers with some students with disabilities, the correlations decreased slightly but were still high (i.e., all above 0.95). The high correlations across the entire sample of teachers indicate that including test scores from students with disabilities does not change the relative ranking of most teachers in the state. Looking at changes across performance quintiles, among teachers with no students with disabilities, almost all remained in the same quintile when scores from students with disabilities were added to the models (see Table 1)

Comparing models with scores from students with disabilities

For all four datasets, correlations among scores from each of the VAM models were all above 0.9 and the correlations between VAM scores and median SGPs ranged from 0.75 to 0.85. The correlations among teacher scores from each of the models were also high when broken out by classroom composition. In particular, the correlations between VAMs were above 0.9 for teachers in classrooms with no students with disabilities, some students with disabilities, and all students with disabilities. These high correlations within the three classroom categories occurred because most teachers maintained their relative ranking within each group of teachers. However,

there is a relationship between model choice and the percent of students with disabilities linked to particular teachers that can be seen looking beyond correlations.

Focusing first on the average percentile ranks by classroom composition (Table 2), there are notable differences across models for teachers linked to all students with disabilities. Specifically, models that do not include variables related to special education rank these teachers below average, and particularly low for math. Models that take into account disability-related covariates (VAM2B and VAM3) produced much higher scores in math for teachers linked to all students with disabilities, moving the average percentile rank close to 50. To explain how the correlations between models are high, even within classroom composition types, while the percentile ranks differ for teachers linked to all students with disabilities, we plotted teachers' scores from VAM2A and VAM2B (single-year estimates) as an example (see Figure 1). The graph shows that VAM2B increased scores for teachers linked to all students with disabilities relative to VAM2A, particularly for math. The increase in scores was generally uniform across all teachers, which caused the high correlations across models, including for teachers linked to all students with disabilities. That is, most teachers maintained their relative ranking, while the scores of the few teachers linked to all students with disabilities improved uniformly when covariates for special education status and accommodation use were added to the VAM model along with other student covariates.

Looking back at Table 2, the average percentile ranks for teachers with some students with disabilities were consistently around average (i.e., 50) across all models. But there is wide variation in the number of students with disabilities linked to teachers in this group. Figure 2 highlights the difference between VAM2A and VAM2B by the percent of students with disabilities linked to teachers. Even though the average percentile ranks were stable across models, the graphs show that as the percent of students with disabilities linked to teachers increases, the difference between models with and without disability-specific covariates increases (except for grade 5 reading, which shows a small change between models for all teachers).

The benefit of including test scores from students with disabilities and accounting for disability-related covariates for teachers linked to some students with disabilities is also apparent when looking at changes across performance quintiles. In other words, including test scores from students with disabilities and accounting for disability status and accommodation use decreases the gap in scores between teachers with few or no students with disabilities and teachers with many students with disabilities.

Conclusions:

Overall, our results provide evidence that the decision to include or exclude test scores from students with disabilities does not appear to affect general education teachers' scores. We see a small amount of movement across performance quintiles, but it is in line with other model comparison studies (e.g., Goldhaber & Theobald, 2013) and can be expected due to sampling variability. With regard to model choice the decision to include disability-related covariates shifted upward special education teachers' scores, appearing to decrease the gap in average rankings relative to general education teachers who teach few or no students with disabilities. Meanwhile the majority of general educators' scores were not affected; but as the number of students with disabilities increased, the greater the likelihood that model specification mattered. Thus, while we suggest that further work be done to replicate our findings, we suggest that concerns of fairness that arise when teachers have larger numbers of students with disabilities can be mitigated through model choice.

Appendices

Appendix A. References

Authors. (2013).

Authors. (2012).

Authors. (2010).

Ballou, D., Sanders, W., & Wright, P.(2004). Controlling for student background in value-added assessment of teachers. *Journal of Education and Behavioral Statistics, 29*(1), 37–65.

Betebenner, D. W. (2012). An R Package for the Calculation and Visualization of Student Growth Percentiles & Percentile Growth Trajectories. [R package version 0.9.9.0].

Betebenner, D. (2009). Norm- and criterion-referenced student growth. *Educational Measurement: Issues and Practice*, *28,* 42–51.

Betebenner, D. (2008). A primer on student growth percentiles. Dover, NH: National Center for the Improvement of Educational Assessment.

Briggs, D. & Domingue, B. (2011). Due diligence and the evaluation of teachers: A review of the value added analysis underlying the effectiveness rankings of Los Angeles Unified School District teachers by the Los Angeles Times. Boulder, CO: National Education Policy Center. Retrieved from http://nepc.colorado.edu/publication/due-diligence

Castellano, K., E., & Ho, A., D. Simple choices among aggregate-level conditional status metrics: from median student growth percentiles to value-added models. Retrieved from http://scholar.harvard.edu/files/andrewho/files/simple_choices_-_castellano_and_ho.pdf

Center on Education Policy. (2009, November). *Has progress been made in raising achievement for students with disabilities?* Author: Washington, DC.

Ehlert, M., Koedel, C., Parsons, E., & Podgursky, M. (2012). Selecting growth measures for school and teacher evaluations. National Center for Analysis of Longitudinal Data in Education Research (CALDAR). Working Paper #80.

Florida Department of Education. (2011). Recommendations of the Florida Student Growth Implementation Committee: Background and Summary. Retrieved from http://www.fldoe.org/committees/doc/Value-Added-Model-White-Paper.doc

Goldhaber D., & Theobald, R. (2013). *Do different value-added models tell us the same things?* Carnegie Knowledge Network Brief. Retrieved from http://www.carnegieknowledgenetwork.org/wp-content/uploads/2012/10/CKN_2012-10_Goldhaber_Nov2013-Update.pdf

Goldhaber, D., & Walch, J., & Gabele, J. (2012). *Does the model matter? Exploring the relationship between different student achievement-based teacher assessments.* (CEDR Working Paper 2012-6). Seattle, WA: University of Washington.

Harris, D. N. (2011). *Value-added measures in education: What every educator needs to know.* Cambridge, MA: Harvard University Press.

Johnson, M., Lipscomb, S., Gill, B., Booker, K., & Bruch, J. (2012). *Value-added models for the Pittsburgh public schools.* Retrieved from http://www.mathematica-mpr.com/publications/pdfs/education/value-added_pittsburgh.pdf

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

McCaffrey, D. F., Lockwood, J. R., Koretz, D., Louis, T. A. & Hamilton, L. (2004). Models for value-added modeling of teacher effects. *Journal of Educational and Behavioral Statistics*, *29*, 67–101.

McCaffrey, D.F., Lockwood, J.R., Mihaly, K., & Sass, T.R. (2012). A review of stata routines for fixed effects estimation in normal linear models. *The Stata Journal*, *12*(3), 406-432.

Nichols, A. (2008). Fese: Stata module to calculate standard errors for fixed effects. http://ideas.repec.org/c/boc/bocode/s456914.html

Sireci, S. G., Scarpeti, S. E., & Li, S. (2005). Test accommodations for students with disabilities: An analysis of the interaction hypothesis. *Review of Educational Research*, *75*, 457–490.

U.S. Department of Education, National Center for Education Statistics (2012). *Digest of Education Statistics, 2011* (NCES 2012-001), Chapter 2.

Warren, S., Thurlow, M., Lazarus, S., Christensen, L. Chartrand, A., and Rieke, R. (2012). *Forum on evaluating educator effectiveness: Critical considerations for including students with disabilities.* Minneapolis, MN: University of Minnesota, National Center on Educational Outcomes and Washington, DC.: Council of Chief State School Officers.

Wright, P. S. (2010). An investigation of two nonparametric regression models for value added assessment in education. Retrieved from http://www.sas.com/resources/whitepaper/wp_16975.pdf

Wu, Y. C., Liu, K. K., Thurlow, M. L., Lazarus, S. S., Altman, J., & Christian, E. (2012). *Characteristics of low performing special education and nonspecial education students on large scale assessments* (Technical Report 60). Minneapolis, MN: University of Minnesota, National Center on Educational Outcomes.

Appendix B. Tables and Figures
Not included in page count.

Table 1
Model covariates

Model	Covariates
SGP	Prior year same subject standardized score
VAM1	Prior year same subject standardized score
VAM2A	Prior year math and reading standardized scores plus general student covariates: gender and ethnicity
VAM2B	VAM 2A plus special education specific-student covariates: special education status and consistent or inconsistent accommodation use
VAM3	VAM 2B plus classroom covariates: average prior reading standardized scores, average prior math standardized scores, percent nonwhite, school poverty level, class size, percent of students receiving special education services, percent of students with consistent or inconsistent accommodation use

Table 2
Average percentile rankings of teachers, by classroom composition

	Median SGP	VAM1	VAM2A	VAM2B	VAM3
Grade 4 Math:		*One-year estimates*			
No students w/ disabilities	52.7	53.7	52.6	50.3	
Some students w/ disabilities	49.8	49.1	49.2	49.7	
All students w/ disabilities	28.8	21.8	26.9	51.7	
		Two-year estimates			
No students w/ disabilities	52.7	53.9	52.8	50.8	51.4
Some students w/ disabilities	49.8	49.1	49.3	49.6	49.9
All students w/ disabilities	28.8	20.3	25.1	49.8	37.4
Grade 5 Math:		*One-year estimates*			
No students w/ disabilities	51.2	50.8	50.1	48.0	
Some students w/ disabilities	51.3	50.8	50.8	51.3	
All students w/ disabilities	28.8	26.2	28.8	49.4	
		Two-year estimates			
No students w/ disabilities	50.6	50.5	49.7	47.8	48.1
Some students w/ disabilities	51.7	51.1	51.1	51.5	51.8
All students w/ disabilities	27.7	24.3	28.0	50.1	37.7
Grade 4 Reading:		*One-year estimates*			
No students w/ disabilities	53.4	54.3	52.4	50.9	
Some students w/ disabilities	50.0	48.9	49.3	49.6	
All students w/ disabilities	41.5	34.4	46.3	60.3	
		Two-year estimates			
No students w/ disabilities	53.4	54.3	52.8	51.4	51.6
Some students w/ disabilities	50.0	49.1	49.4	49.6	49.7
All students w/ disabilities	41.5	31.5	42.6	60.1	55.0
Grade 5 Reading:		*One-year estimates*			
No students w/ disabilities	49.1	50.8	48.6	47.9	
Some students w/ disabilities	51.9	50.7	51.2	51.4	
All students w/ disabilities	44.1	33.0	46.7	52.7	
		Two-year estimates			
No students w/ disabilities	49.5	50.8	48.4	47.2	45.8
Some students w/ disabilities	52.0	50.8	51.3	51.5	52.0
All students w/ disabilities	43.7	32.6	46.7	61.0	61.6

Table 3
Correlation between covariates and teacher scores from different models (two-year estimates)

	SGP	VAM1	VAM2A	VAM2B	VAM3
In a high poverty school?					
Grade 4 math	-0.03	-0.17	-0.11	-0.13	-0.27
Grade 4 reading	-0.23	-0.36	-0.27	-0.28	-0.33
Grade 5 math	-0.13	-0.19	-0.09	-0.10	-0.08
Grade 5 reading	-0.2	-0.33	-0.22	-0.23	-0.14
% of students with disabilities					
Grade 4 math	-0.12	-0.19	-0.14	0.01	-0.06
Grade 4 reading	-0.05	-0.12	-0.05	0.04	0.02
Grade 5 math	-0.12	-0.16	-0.13	0.00	-0.06
Grade 5 reading	-0.04	-0.11	-0.02	0.07	0.08
Average prior same-subject achievement					
Grade 4 math	0.10	0.31	0.22	0.20	0.34
Grade 4 reading	0.18	0.41	0.29	0.27	0.31
Grade 5 math	0.20	0.32	0.22	0.20	0.26
Grade 5 reading	0.29	0.46	0.31	0.29	0.21

Note. Correlations for models using one-year estimates were similar.

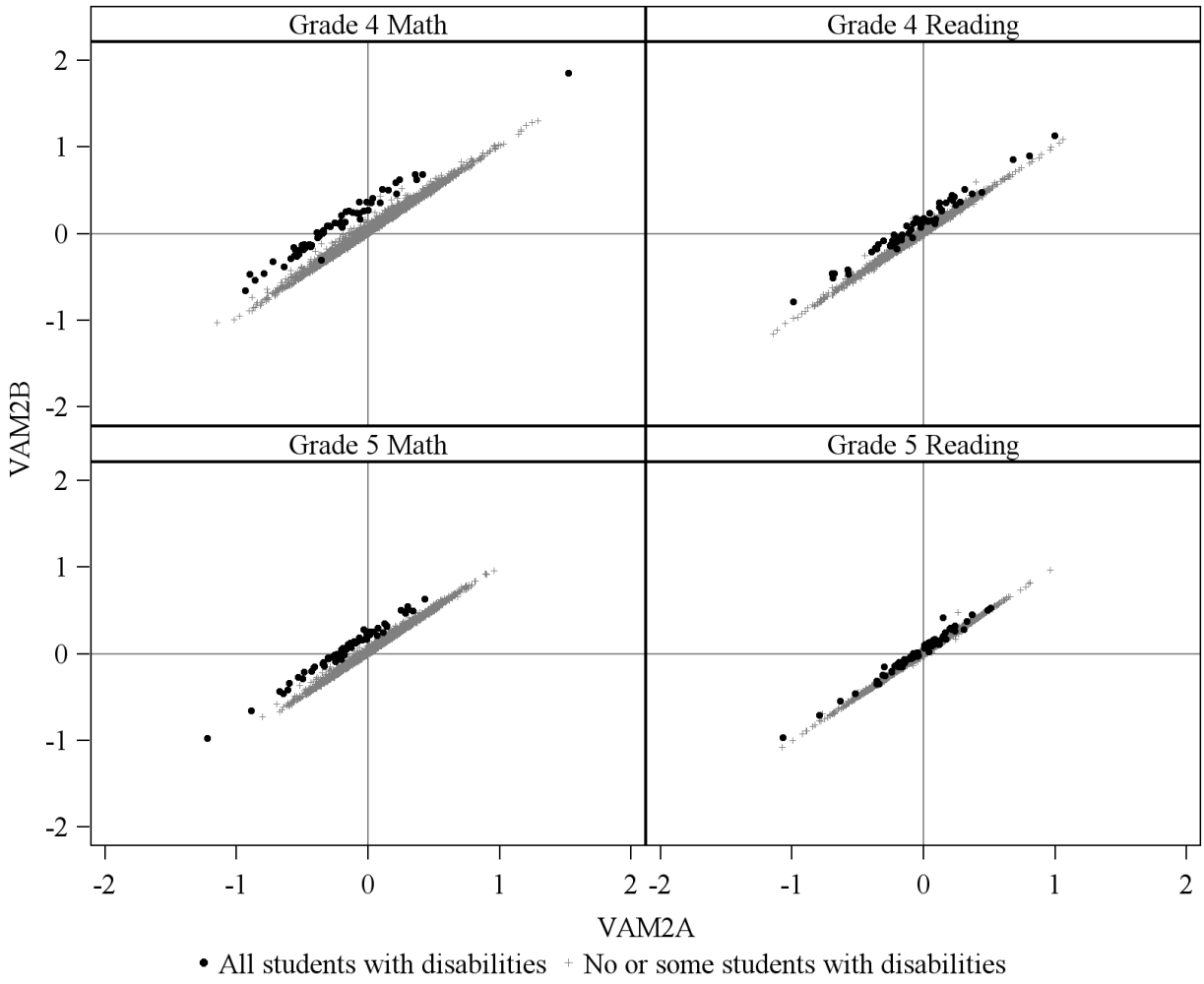

Figure 1. Difference between teacher scores from the model with general student covariates (VAM2A) and the model with additional disability-related covariates (VAM2B). Both models are for single-year estimates. Markers denote classroom composition.

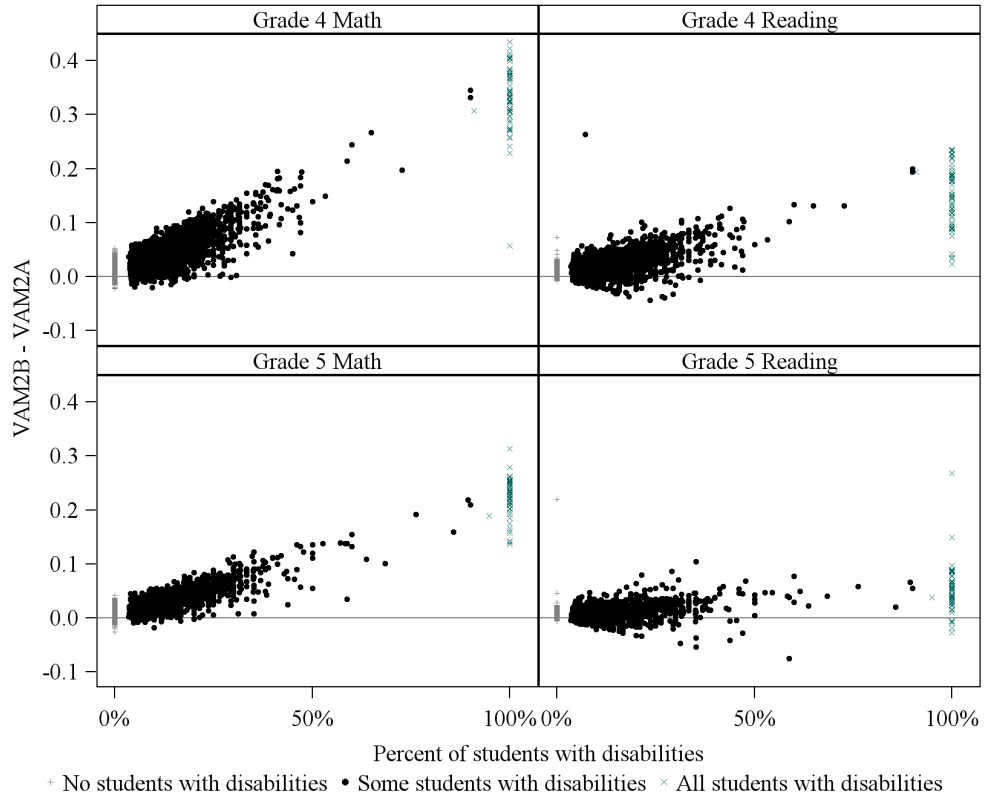

Figure 2. Difference between teacher scores from VAM2A and VAM2B. Markers denote classroom composition.

[i] There were a small number of teachers linked to all EL students with no disabilities (11 grade 5 teachers and 24 grade 4 teachers). There were approximately 200 teachers linked to some ELs (and no students with disabilities) in each grade. Our results for the latter group of teachers were similar to the group of teachers linked to no students with disabilities, but we do not report results for them for brevity and clarity.