Impact of Approaches to Handling Dropout on Value-Added Analysis

A paper presented at the annual meeting of the National Council on Measurement in Education
Philadelphia, PA

Paula Cunningham

Catherine Welch

Stephen Dunbar

April 2014

Abstract

Using a multi-cohort longitudinal data set from a school district in a Midwestern state, several treatments for handling missing data were evaluated for their effects on teacher value-added. Spearman rank correlations between corresponding rankings derived from data sets treated differently for their missing data were calculated, and the average of the correlations between methods for these rankings ranged from .851 to .997. Mean imputation resulted in rankings markedly different from the others, while the pairwise deletion, regression imputation, and multiple imputation rankings correlated very highly. Teacher rankings were further examined to determine what percentage of teachers under three categorization schemes would move from one category of effectiveness to another due to the handling of missing data. The results of this study suggest that pairwise deletion is the preferable method if the use of imputed data is viewed negatively, but regression or multiple imputation are superior if imputed data can be used.

Impact of Approaches to Handling Dropout on Value-Added Analysis

States and school districts are increasingly turning to evaluation systems that incorporate students' standardized test scores to some degree in consequential decisions about teacher salaries, promotions, tenure, and dismissal (Braun, 2005). Value-added models are being applied as a way to quantify deviations from expected student test performance after a year of instruction in a teacher's classroom. These models take into account the different "starting points" of students by incorporating characteristics such as prior scores on achievement tests and perhaps other student and school factors as well. Teachers in elementary grades whose students take standardized tests in reading and mathematics can be held accountable for getting them to achieve their expected scores in those subjects. The movement toward tying student performance on tests to teacher evaluations gained considerable momentum when the Obama administration endorsed the practice, through the awarding of points in its Race to the Top competition to states that linked them (Braun, 2012).

While value-added models are being applied more widely for the purpose of teacher evaluation, the validity of the inferences made from such analyses has not been unequivocally established. A positive view is that value-added models will add objectivity to teacher evaluation systems that have heretofore relied on seniority, attainment of credentials, and principal observations of classroom performance (Braun, 2012); likely the first two of these do not directly measure teacher effectiveness in the classroom, and in too many cases principal observations occur infrequently and result in satisfactory ratings for virtually all teachers (Papay, 2012). Nevertheless, the degree of confidence one can place in value-added estimates of teacher effectiveness depends in part on factors such as how well assumptions are met and the integrity of the data inputs (National Research Council and National Academy of Education, 2010).

Despite the enthusiasm of state legislatures for mandating such teacher evaluation systems, there is not a clear set of "best practices" available for those charged with implementing them. There are numerous decisions facing states and school districts during the process of implementing teacher evaluation systems based on student achievement test scores. A decision must be made about how to weigh the teachers' value-added scores with any other measures of teacher effectiveness being used, and there are also numerous choices to be made concerning the value-added modeling itself. Some factors to consider include how to characterize growth, how many years of data to use in the model, how teachers influence achievement over time, and which student or teacher characteristics to control for in the model. In addition, states and school districts need to be concerned with the integrity of the inputs and must decide about how to deal with test scores missing from the longitudinal data set.

Missing longitudinal data can be considered either missing-at-random (MAR) or missing-not-at-random (MNAR), where the probability of a score being observed is dependent upon the value of the score (Singer & Willett, 2003). The assumption that longitudinal data are missing-at-random—that the unobserved data are similar to the data that are observed—is commonly made in value-added modeling (Wainer, 2011). However, this assumption is not well supported, as it is known that children who miss exams are not just like those who take them, depending on the reason for the dropout. In a recent study investigating this issue, McCaffrey and Lockwood (2011) modeled missing data as either MAR or MNAR and concluded that the MNAR model better fit the longitudinal data set but that varying this condition had little impact on the resulting teacher effects, as they were very highly correlated with one another. The present study aims to further investigate the issue by exploring the effect on value-added estimates of several empirical methods of handling missing data that could be employed by states and school districts.

Objectives

A data set can be treated for missing data in two general ways—deletion or imputation. In pairwise deletion, only students with scores in adjacent years are included in the computation of teacher value-added. This method results in maximizing the amount of real data used in the analysis. In listwise deletion, only complete cases are used in the computation. This approach has the drawback of using fewer observations in the analysis. Missing data can also be imputed by several methods that include single (mean or regression imputation) or multiple imputation procedures (Fitzmaurice, Laird, & Ware, 2011). The principal drawback of imputation methods is that not all the data used in the analysis is real. An important concern is that whenever data is missing, bias can be introduced by whichever method is used to handle it.

The primary goals of this study are to treat a school district's longitudinal achievement test data by a variety of methods to handle missing information and to use the new data sets created either by deletion of cases or by imputation of values to generate value-added estimates of teacher effectiveness. In this way the impact of the various approaches to handling dropout can be compared in the context of teacher value-added. The teacher rankings that result from the different methods can be compared using Spearman rank correlation coefficients, and the impact on teacher evaluation ratings can be evaluated by considering the percentage of teachers who change ratings due to the manner in which missing data is handled.

Data

The analysis utilized scores on the reading and mathematics subtests of an achievement test battery, used by a Midwestern state for accountability and aligned with its standards in those subjects. The subtest reliabilities over the levels used in this study ranged from .89 to .92 (average .90; Hoover et al., 2003).

Three longitudinal cohorts from a single school district in the state were used in the study, with each student having up to four years of matched achievement test scores that spanned the years 2004-2008, 2005-2009, or 2006-2010. Each longitudinal data set was comprised of a cohort of students who were in the third grade during the first school year (2004, 2005, or 2006); their scores for that school year and the three subsequent years were included. As a consequence, each data set could potentially provide information about the students' growth in grades three, four and five. The total number of students whose data were included in this study was 3,771.

Methods

Using the multi-cohort longitudinal data set described above, five procedures for handling missing data were evaluated for their effects on the value-added estimates derived from them using a fixed-effects covariate adjustment model (McCaffrey, Lockwood, Koretz, & Hamilton, 2003). Value-added estimates were calculated for the 3rd, 4th, and 5th grade teachers in the district in both subjects and for all cohorts, after the data had been treated in turn by each of the methods described.

Two deletion approaches for dealing with the dropout were considered. For the pairwise deletion condition, a reduced data set containing only students with test scores in adjacent years was made. For the listwise deletion condition, only students with scores in all years were included in the data set; hence, this set had the smallest number of observations.

Two single imputation methods were evaluated in the study. For the mean imputation data set, sample mean values of the test scores for all available cases replaced the missing values. As a result of this procedure, the data set was complete, and the means of the test scores did not change; however, mean values were unlikely to approximate well what the missing values would have been. The regression imputation procedure substituted fitted values from the regression

model for the missing data and resulted in a complete dataset; while this method gives more likely values for the missing data than mean imputation does, it does not account for their uncertainty.

The final method used for handling dropout in this study was multiple imputation, which like single imputation uses observed data to impute missing values and gives a complete data set; however, multiple imputation reflects the variance in imputed values by generating plausible values for missing data. Complete case analysis is performed on each data set generated, and the results are combined. Multiple imputation for this study was carried out in SAS/STAT® 9.3 software (SAS Institute Inc., 2011) with PROC MI, using the Markov chain Monte Carlo (MCMC) method to create five imputations for each cohort data set.

Results

The multi-cohort data set was first examined for patterns of missingness: an arbitrary, rather than a monotone, missing pattern was confirmed. In addition, the mean scores of students with complete records were compared to those of students who had one or more missing scores. In Table 1 and Table 2 these mean scores are displayed for the subtests in reading and mathematics, respectively. The means for students with incomplete data records are consistently lower than those of students who have complete data records.

Value-added estimates were calculated for classroom teachers who had at least fifteen students with test scores in reading and mathematics. Spearman rank correlation coefficients between corresponding rankings (same cohort, grade, and subject) derived from data sets treated differently for their missing data were calculated. A set of average correlations over all cohorts and grades in each subject was computed using a Fisher transformation for each method compared with every other. These are presented in Table 3, where the correlations for reading

are presented above the diagonal and those for mathematics are shown below the diagonal. The average correlations between methods ranged from a low of .851 to a high of .997. The mean imputation procedure resulted in teacher value-added rankings that were most different from those derived using other procedures for handling dropout, correlating consistently lower (.851 to .907) than the other methods. On the other hand, the pairwise deletion, regression imputation, and multiple imputation teacher rankings correlated very highly with one another, ranging from .985 to .997. The correlations of these rankings with rankings derived using the listwise deletion procedure ranged from .934 to .944. Thus, the methods tested for handling dropout could be ranked by how consistent they were with one another, with pairwise deletion, regression imputation, and multiple imputation being highly consistent, followed by listwise deletion at moderately consistent, and finally mean imputation as the least consistent.

Teacher rankings were further examined to determine how many teachers would move from one category (e.g., very effective, effective, or ineffective) to another due to the manner in which missing data was dealt with before the analysis was done. This calculation was made under three different but realistic conditions for placing the teachers into rating categories.

Under a rating scheme where the highest 25% of teachers are considered very effective, the middle 50% of teachers are labeled effective, and the lowest 25% of teachers are termed ineffective, comparing the mean imputation rankings with those from all other methods revealed that more than 20% of teachers moved from one category to another (Table 4, in which the percentages for the reading rankings are above the diagonal and those for the mathematics rankings are below the diagonal). Treatment by listwise deletion before analysis resulted in about 15% of teachers changing ratings compared to other ratings. The use of pairwise deletion, regression imputation, or multiple imputation resulted in the least number of occurrences of

teachers changing ratings due to method differences under the 25%-50%-25% scheme, with percentages all at 10% or below.

The same calculations were done for a teacher rating scheme where the highest 15% of teachers are labeled very effective, the middle 70% of teachers are termed effective, and the lowest 15% of teachers are considered ineffective (Table 5). The percentages of teachers changing ratings were lower overall because the upper and lower cutoffs yield smaller numbers of teachers at the extremes of the rankings. Here the use of mean imputation still gave relatively higher percentages of teachers who changed rating categories, but the listwise deletion procedure had similar, although slightly smaller, percentages of teachers changing ratings. Once again, the pairwise deletion, regression imputation, and multiple imputation procedures resulted in the smallest percentages of changing ratings due to method differences, all less than 10%.

A final teacher rating scheme was posited, where the highest 5% of teachers are termed very effective, the middle 90% of teachers are considered effective, and the lowest 5% of teachers are labeled ineffective (Table 6). Here the relative differences in the percentages of teachers changing categories due to methods of handling missing data remained the same as in the other two schemes, but all of the percentages were smaller than 10% because there were so few teachers at the extremes under this condition.

Discussion

Choices must be made by administrators about how to deal with missing data in order to calculate value-added estimates for teacher evaluation systems. This study suggests that pairwise deletion is the best way to proceed if the use of imputed data is viewed negatively, given that its use results in rankings that are highly correlated with those derived from datasets using regression or multiple imputation, which are preferable if imputed data can be used.

Decisions that are made by officials in states and school districts concerning the implementation of teacher evaluation systems impact numerous stakeholders, particularly educators themselves. It is important to note that how missing data is handled can affect the relative position of an educator in his or her ranking and that this fact may result in a teacher being categorized as effective using one method but ineffective using another.

The degree of changes in the ratings of teacher effectiveness due to the method used for handling dropout can vary with the scheme being used for categorizing teachers. This study has shown that the percentages of teachers being mislabeled can be reduced by making the categories at the extremes contain fewer teachers, but that would surely defeat the purpose of meaningful teacher evaluation.

The 15%-70%-15% scheme for teacher ratings may be the most reasonable one of those proposed, but even so the differences due to method for handling dropout would have an impact on teacher ratings. Based on the percentages in Table 5, for a school district with forty fourth grade teachers, one should expect from two to eight fourth grade teachers to be rated differently depending on which method was used to handle missing data. Considering how many teachers at all grade levels would be evaluated on the basis of their students' test scores, a great number of people could be impacted, and the integrity of the evaluation system could be questioned.

References

Braun, H. (2005). *Using student progress to evaluate teachers: A primer to value-added models.* Princeton, NJ: Educational Testing Service.

Braun, H. (2012). *Value-added modeling and the power of magical thinking.* Paper presented at the 2012 Annual Meeting of the National Council on Measurement in Education, Vancouver, BC.

Fitzmaurice, G.M., Laird, N. M, & Ware, J. H. (2011). *Applied Longitudinal Analysis, 2nd Edition.* New York, NY: John Wiley and Sons.

Hoover, H.D., Dunbar, S.B., Frisbie, D.A., Oberley, K.R., Ordman, V.L., Naylor, R.J., Bray, G.B., Lewis, J.C., Qualls, A.L., Mengeling, M.A., & Shannon, G.P. (2003). *The Iowa tests: Guide to research and development.* Itasca, IL: Riverside Publishing.

McCaffrey, D.F. & Lockwood, J.R. (2011). Missing data in value-added modeling of teacher effects. *Annals of Applied Statistics*, *5*(2A), 773-797.

McCaffrey, D.F., Lockwood, J.R., Koretz, D.M., & Hamilton, L.S. (2003). *Evaluating value-added models for teacher accountability*. Santa Monica, CA: RAND Corporation.

National Research Council and National Academy of Education. (2010). *Getting value out of value-added: Report of a workshop.* Committee on Value-Added Methodology for Instructional Improvement, Program Evaluation, and Educational Accountability, Henry Braun, Naomi Chudowsky, and Judith Koenig, Editors. Center for Education, Division of Behavioral and Social Sciences and Education. Washington, DC: The National Academies Press.

Papay, J.P. (2012). Refocusing the debate: Assessing the purposes and tools of teacher evaluation. *Harvard Educational Review*, *82*(1), 123-141.

SAS Institute Inc. (2011). *SAS/STAT® 9.3 User's Guide.* Cary, NC: SAS Institute Inc.

Singer, J.D. & Willett, J.B. (2003). *Applied Longitudinal Data Analysis.* New York, NY: Oxford University Press.

Wainer, H. (2011). *Uneducated guesses: Using evidence to uncover misguided education policies*. Princeton, NJ: Princeton University Press.

Table 1

Mean Scores on Reading Subtests by Students with Complete or Incomplete Records

			Grade			
		N	3rd	4th	5th	6th
Cohort 1	Complete	761	187.6	207.8	224.7	230.8
	Incomplete	425	177.9	199.9	215.5	219.8
Cohort 2	Complete	830	186.1	210.2	221.7	227.7
	Incomplete	455	181.1	202.0	210.4	219.9
Cohort 3	Complete	841	187.1	205.4	222.3	229.4
	Incomplete	459	179.6	192.9	214.1	218.4

Table 2

Mean Scores on Mathematics Subtests by Students with Complete or Incomplete Records

			Grade			
		N	3rd	4th	5th	6th
Cohort 1	Complete	759	185.0	206.2	226.2	238.6
	Incomplete	424	177.5	200.3	216.5	225.3
Cohort 2	Complete	827	187.3	208.9	225.5	234.5
	Incomplete	458	180.2	201.5	215.6	224.0
Cohort 3	Complete	836	186.2	205.8	224.0	237.5
	Incomplete	461	179.3	195.7	216.6	226.2

Table 3

Average Spearman Correlations of Rankings Between Methods for Handling Dropout

	Pairwise Deletion	Listwise Deletion	Mean Imputation	Regression Imputation	Multiple Imputation
Pairwise Deletion		0.944	0.905	0.997	0.987
Listwise Deletion	0.942		0.855	0.942	0.934
Mean Imputation	0.895	0.851		0.907	0.900
Regression Imputation	0.997	0.942	0.895		0.988
Multiple Imputation	0.985	0.934	0.891	0.988	

Note. The correlations for the reading rankings are above the diagonal and those for the mathematics rankings are below the diagonal.

Table 4

Average Percentage of Teachers who Switch Ratings under Different Conditions of Handling Dropout in a 25%-50%-25% Scheme

	Pairwise Deletion	Listwise Deletion	Mean Imputation	Regression Imputation	Multiple Imputation
Pairwise Deletion		15%	23%	4%	6%
Listwise Deletion	14%		28%	15%	17%
Mean Imputation	21%	25%		25%	23%
Regression Imputation	2%	16%	22%		6%
Multiple Imputation	10%	16%	22%	10%	

Note. The percentages for the reading rankings are above the diagonal and those for the mathematics rankings are below the diagonal.

Table 5

Average Percentage of Teachers who Switch Ratings under Different Conditions of Handling Dropout in a 15%-70%-15% Scheme

	Pairwise Deletion	Listwise Deletion	Mean Imputation	Regression Imputation	Multiple Imputation
Pairwise Deletion		11%	16%	5%	8%
Listwise Deletion	13%		20%	13%	13%
Mean Imputation	15%	21%		14%	14%
Regression Imputation	2%	13%	15%		7%
Multiple Imputation	6%	12%	15%	5%	

Note. The percentages for the reading rankings are above the diagonal and those for the mathematics rankings are below the diagonal.

Table 6

Average Percentage of Teachers who Switch Ratings under Different Conditions of Handling Dropout in a 5%-90%-5% Scheme

	Pairwise Deletion	Listwise Deletion	Mean Imputation	Regression Imputation	Multiple Imputation
Pairwise Deletion		7%	7%	4%	4%
Listwise Deletion	5%		8%	7%	8%
Mean Imputation	7%	8%		8%	8%
Regression Imputation	5%	6%	9%		3%
Multiple Imputation	4%	5%	7%	3%	

Note. The percentages for the reading rankings are above the diagonal and those for the mathematics rankings are below the diagonal.