

Measurement Issues Associated with Value-added Methods¹

Mark D. Reckase
Michigan State University

Value-added models are extremely complex mathematical/statistical expressions. They include numerous student, school, and environmental variables. Rather than begin the discussion of measurement issues related to the use of these models with that extreme level of complexity, a much simpler analogous example to growth in educational achievement is used to lay out the measurement issues.

A Simple Example

Suppose that instead of considering the variables that explain the increase in educational achievement, there is interest in the influences on a more easily measured characteristic of students – their heights. Height is selected here because it can be measured very accurately so an initial part of the example can avoid the issue of measurement error. However, measurement error will not be totally ignored with the example. It will be considered later.

Suppose that a random sample of 1000 nine-year-old girls is obtained from the school attending population of girls in the United States and that the height of these girls is carefully measured in inches. Then, one year later these same girls have their height measured a second time. After the second measurement, it would be an easy task to determine how much each girl grew in one year – the first measurement is simply subtracted from the second one. Summary statistics could also be determined from these measurements. Table 1 shows the mean and standard deviations for the heights for the girls along with the mean and standard deviation of the increase in height. The correlation between the heights over the two years is .75. All girls are not gaining in height by the same amount even though there is a mean gain of three inches.

Table 1
Mean and Standard Deviation
of Heights in Inches of 9- and 10-year Old Girls

	Age in Years		Gain in Inches
	9	10	
Mean	52.07	55.10	3.03
Standard Deviation	1.78	2.33	1.56

There is one small problem that becomes evident from looking at the summary statistics in the table. The mean growth is 3.03 and the standard deviation of the gain is 1.56. If the gain were normally distributed, this would imply that about 2.5% of the girls lost height over the period of the year. The fact that this is very unlikely implies that the

¹ Paper written at the request of the National Research Council Committee on Value-added Methodology for Instructional Improvement, Program Evaluations, and Educational Accountability, October, 2008.

growth is not normally distributed. For this example, the growth is slightly positively skewed and a little flatter than a normal distribution. Even though the height at age nine was modeled here to have a normal distribution, the fact that growth is not normal means that the distribution of age ten height is not normal. When considering the generalization of this basic model to value-added in education, the assumption of distributional forms needs to be considered.

The transfer of this example to the case of educational achievement would be very straightforward except that for the educational case, the measurement tools at the two ages do not have the same units. The measurements usually come from different tests with different score scales. The equivalent in the example for height would be that the first year heights are measured in inches and the second year they are measured in centimeters. However, in the case for educational achievement, the conversion from the one set of units to the other is not known.

To model this situation, suppose that all of the heights for the ten-year-olds are measured in centimeters. This means that the gain in height can not be determined directly through subtraction of the first height from the second. One way to approach this problem is to fit a linear regression model to the data. Then the regression function can be used to predict the results on the new scale from the heights collected from the nine-year-olds. The gain in height over the two years is a linear function of the residual in prediction. That is, the students with the largest positive residual had the largest gain in height and the largest negative residual had the least gain in height.

It may also be possible to interpret the slope and the intercept of the regression equation for the data. The intercept term is the predicted value for when the height for the nine-year-olds is zero. This could be interpreted as the gain in centimeters from age nine to age ten. Because the true mean gain was 3.03 inches, the gain in centimeters should be 7.7. The slope could be interpreted as the constant for change in measurement units from age nine to age ten. That constant should be 2.54. The actual results for the estimated slope and intercept from the data are 2.48 and 10.69 respectively. The slope was fairly well estimated, but the intercept term is quite different than the true value, although the 95% confidence interval for the intercept contains the true value. Overall, the interval estimates are accurate, but the point estimates might not be.

The analysis of the height measurements is a useful analog for the analysis of gains in achievements except for one aspect of the data. These data do not have any measurement error, but actual test data has less than perfect reliability. Suppose that instead of measuring the height with accurate instruments, the height data is collected by teacher judgment with a standard error of .8. This is equivalent to a reliability of .8 for the measures of height for age nine.

If the same regression analysis procedure is run on the height data with error of measurement, the resulting slope and intercept are 2.09 and 31.05 respectively and the correlation between the measures is .64. The confidence intervals around these values do not include the true values used to generate the data. The intercept is much larger than

the true value of gain of 7.7 and the slope term is much smaller than the change in units of 2.54. The correlation is much lower than the value when there was no error of measurement of .75. The correlation between the true gains in height and the residuals from the regression equations is .81 instead of 1.0 when there was no error in measurement, but there was a change in units of measurement.

To build an analogy that totally matches the case for achievement gains, the measurement at age ten would have to measure something that is highly related to height, but that is not quite the same thing as height. For example, weight could be measured because it is highly related to height for this age of student. This would be equivalent to having the test at one grade level measure something slightly different than the test at the previous level. This would add more uncertainty and error to the relationship between the measures at the two age levels.

Connection to Gains in Achievement

The example of the analysis of gains in height was carefully constructed to mirror the assessment of gains in achievement for students. Instead of considering the height of nine-year-old girls, consider the mathematics achievement of 4th grade students. The level of achievement could be assessed using a carefully aligned state assessment program. This would not be the highly accurate measurement of achievement like the accurate measurement of height. Instead it would be similar to the judgmental estimate of height. The assessment of mathematics achievement would have reliability less than 1.0 so there would be some error in measurement.

The assessment of mathematics achievement at grade 5 would typically be done with a different form of the test than was used for grade 4. That means that the two tests would have different scales of measurement so the gains in achievement could not be determine by simply subtracting the scores from grade 4 from those at grade 5. This is similar to the case of measuring height using inches and centimeters for different age groups. Of course, the grade 5 test would also have reliability less then 1.0.

Given this situation, gains in achievement can be estimated in two different ways. One is to use the regression equation relating grade 5 to grade 4 to estimate the conversion constant for the different units of measurement and to estimate the average change in achievement. This approach would have the same difficulties that were evident from the height example presented earlier.

The second approach is to vertically scale the two tests so that the reported scores are on the same scale. Then the gain in achievement can be determined by subtracting the grade 4 score from the grade 5 score for each person. The estimate of gain would still be affected by the error of measurement on the two tests and any error from the scaling process.

Measurement Requirements for Value-Added Assessment Models

Value-added assessment models are basically elaborate regression models. The left hand side of the equation is the level of achievement at a particular point in time. The right hand side of the equation contains the predictor variables including the achievement measure at the earlier grade in a general linear model. Articles such as McCafrey, Lockwood, Koretz and Hamilton (2003) describe the general forms of these models.

The use of this form of general linear model requires that the data being modeled meet certain assumptions. The first is that the residual distribution for the achievement variable given the levels of the predictor variables is normal with mean zero. The fact that a specific form of distribution is assumed means that the model also requires that the scale for the achievement measure also support the interpretation of equal intervals on the scale. Without equal intervals, it is not possible to specify the shape of a distribution.

Other assumptions are that the variances of the residual distributions are equal across levels of the predictors, that there is a linear relationship between predictors and the achievement measure, and that the residuals are uncorrelated with the predictor variables. All of these assumptions require that the achievement measure have interval scale properties. That is, that equal numerical differences on the scale of the achievement measure mean equal differences on the underlying construct.

There are two ways that interval scale properties are argued for achievement test data. The first is by reporting results on a score scale that is a linear transformation of item response theory θ -estimates. Item response theory models also require that the θ -scale have interval scale properties because the form of the item response theory function is not defined unless the θ -scale has interval properties. If the item response theory model fits the item response data, then it can be argued that the θ -scale has interval properties.

The second justification for interval scale properties is through the assumption of a specified form for the true score distribution on the test. If the observed score distribution matches the assumed distribution, then it can be argued that the resulting score scale has interval properties. Typically, normal distributions are assumed for the achievement of unselected populations of school children. If the number correct score has a normal distribution, this is considered support for interval properties for the score scale.

For some state assessments, neither of these arguments for interval scale properties can be supported. The reporting score scale may have non-normal distributions and the scaling model may be based on classical test theory. In other cases, there may be misfit to the item response theory model. Little research has been done to determine the quality of fit needed to support interval interpretations of the θ -scales. In most cases, the issue of the nature of the score scale is not considered when doing value-added modeling. The interval scale nature of the data is assumed, but not checked.

When the left side of the side of the value-added model has a gain score rather than the achievement measure, the measurement requirements are more stringent than a model with an achievement measure on the left side. For the gain score case, the two measures that are used to compute the difference need to measure the same construct and use the same units of measurement. The example at the beginning of this paper met these requirements because height was measured at both times using inches. For achievement measures to meet these requirements, the tests at both grade levels need to measure the same construct (the content of the tests can not shift from one grade to the next), and the test results need to be put on the same scale. This means that the tests have to be carefully constructed according to test specifications designed to keep the same construct, and the tests need to be put on the same scale using methodology called vertical scaling.

Consideration of Measurement Issues in the Value-added Literature

The literature on value-added modeling makes little mention of the measurement requirements for using the models. For example, a summary of value-added research published by the American Educational Research Association (Zurawsky, 2004) only indicates that the tests need to be aligned to the state curriculum for them to be used for value-added modeling. This article does not mention any of the measurement requirements for value-added modeling summarized above. Other articles do not mention any measurement requirements for the achievement measures at all. For example, Dee (2004) uses test data from the Stanford Achievement Test in a study of teacher and race effects, but he does not considered any of the issues of test alignment, test construct equivalence over grades, or vertical scaling. In fact, the study uses percentile ranks as the test scores being analyzed. This type of test score does not have a normal distribution and they do not have equivalent units across grade levels.

Articles that provide more guidance about measurement issues than most others were written by Doran and Fleischman (2005) and McCaffrey, Lockwood, Koretz and Hamilton (2003). These articles indicates that the tests used in value-added models need to be measuring the same thing at different grade levels and that the scores need to be put on the same scale using vertical scaling methods. The article also states that test specifications change over grades so the common construct requirement is not usually met, and that vertical scaling is a difficult technical process that may not always work properly. The point out that there are multiple methods for test design and development, and for vertical scaling and the different methods yield different results.

Although there is little consensus in the measurement community on the desired features of an assessment program for use with value-added models, there is clearly a trend within state assessment programs toward developing testing programs that use vertical scaling to connect the score scales from different grades. The clear intent of developing these vertically scaled tests is to allow the direct computation of gains in achievement from one year to another.

An Ideal Testing Program for Value-added Models

The example provided at the beginning of this paper and the advice provided by various the authors referenced earlier suggest requirements for a grade level testing program that will be analyzed using value added models. First, the test scores within a grade need to at least approximate an interval scale. This can be done by either using an IRT model that fits the data to determine the score scale, or design the test to yield an approximately normal distribution of scores. Another alternative is to make an argument that the score distribution should be normal and then perform a normalizing transformation on the test scores. These requirements would need to be met for each grade level test in the program.

The second requirement is that the tests need to be highly reliable. As shown in the height example at the beginning of this paper, the amount of error in the estimates of the scores affects the estimates of the value-added model parameters. The third requirement is that the tests at the different grade levels need to be measuring the same construct. This is a challenging requirement because there is usually a desire to shift the test content at each grade level to match the shift in curriculum. These shifts in content should be minimized to support the use of the models.

The height example also showed that changes in the units for the measurements at different grade levels can affect the estimates of model parameters. This result argues for vertically scaling the test scores at the different grade levels to put the results on scales with the same units of measurement.

There are other non-measurement requirements for getting test results that accurately reflect student gains in achievement. The students must be motivated to do their best on the tests and the testing environment must be conducive for them to perform. That is, the test administration setting must be well lit and quiet, and students need to have adequate time to do the test. A well designed testing program will not give accurate results if schools and students do not cooperate in testing process.

Research Issues

Little research is available on the effect of violations of model assumptions on the accuracy of estimates of the parameters of the value-added models. Research on this issue is critical for determining the level of confidence that can be placed in the inferences made from the value-added analyses. Such research might start with evaluations of the fit of the models to the data being analyzed. It is also important to know the proportion of variance in the observed achievement that is accounted for by the models and the actual level of achievement that is predicted by the models. The research now seems to focus on the relationships between explanatory variables, such as type of teacher preparation, on student achievement. But even when a variable is found to be strongly related to student achievement, it is difficult to judge the magnitude of the effect in any practical metric. It would be useful to know the total gain in achievement over a

school year for the target population of students and then determine how much the gain would change as a result of the studied variables.

If it can be shown that the selected value-added model fits the data from an educational system and testing program, then the next important question is to determine the effects of level of test reliability, the type of vertical scaling procedure that is used, and the shape of distributions of gain on the accuracy of estimation of model parameters. My own previous research (Reckase and Li, 2007), suggests that various sources of error in the test scores lead to underestimating the effects of instruction. That work considered the mismatch between the constructs at adjacent grades, but it is likely that unreliability and scaling error would have similar effects. Work is needed to develop corrections for unreliability similar to the corrections for attenuation commonly used in psychological research.

Given the complexity of value-added models and the concern that violations of assumptions and the naturally occurring error levels in test scores and vertical scaling processes have effects on the estimates of model parameters, it is important that the procedures be given a thorough evaluation prior to operational implementation. One way of evaluating a model is to generate simulated data that has the same characteristics as operational data and determine whether the model can accurately capture the relationships that were built into the simulated data. If the model does not estimate parameters with sufficiently accuracy from data that are generated to fit the model and match the characteristics of the test data, there is little likelihood that the model will work well with actual test data.

Discussion

It would not be surprising to discover that error in test scores and in vertical scaling tended to degrade the accuracy of estimates of parameters in value added models. The results of the simulation presented at the beginning of this paper suggest that the practical implication of the reduced accuracy of estimation is that the effects of schools and teachers are probably underestimated. The underestimation may not have serious consequences if the results will only be used to rank order schools or teachers, but there will likely be serious consequences when determining if one teaching treatment is significantly better than another. The various sources of error will reduce the power to detect differences. The effects of error or model misfit may be even greater if the goal is to compare gains to a fixed standard. If the gains are underestimated, schools and teachers will be erroneously categorized as not bringing about the required level of achievement.

Given the direction of influence of error from tests and vertical scaling, observed relationships are probably stronger than indicated by the analyses. We are looking at the educational system through a poor quality lens. The real world is probably more orderly than it appears from the analyses of noisy data.

References

- Dee, T. S. (2004). Teachers, race, and student achievement in a randomized experiment. *The Review of Economics and Statistics*, 86(1), 195-210.
- Doran, H. C. and Fleischman, S. (2005). Challenges of value-added assessment. *Assessment to Promote Learning*, 63(3), 85-87.
- McCaffrey, D. F., Lockwood, J. R., Koretz, D. M., and Hamilton, L. S. (2003). Evaluating value-added models for teacher accountability. Santa Monica, CA: Rand.
- Reckase, M. D. and Li, T. (2007). Estimating gain in achievement when content specifications change: a multidimensional item response theory approach. In R. W. Lissitz (Ed.) *Assessing and Modeling Cognitive Development in School*. Maple Grove, MN: JAM Press.
- Zurawsky, C. (2004). Teachers matter: evidence from value-added assessment. *Research Points: Essential Information for Educational Policy*, 2(2).