Abstract Title Page
Not included in page count.

Title:
Impacts of Multidimensionality and Error: Simulating Explanations for Weak Correlations between Measures of Teacher Quality

Authors and Affiliations:
Mark Chin, Harvard Graduate School of Education
Dan Goldhaber, CALDER, American Institutes of Research

Abstract Body

Background / Context / Purpose:
Over the past decade, states and districts have updated teacher evaluation systems in response to federal incentives, such as waivers to No Child Left Behind and grants from Race to the Top. Many of these newer systems incorporate multiple measures purported to assess the quality of teachers. The use of multiple measures, however, raises an important issue for policymakers and practitioners: to what extent should different measures of a teacher's performance align with one another? We explore this issue here for two measures commonly used to evaluate performance: classroom observations of teacher practices and value-added.

Though school systems have long assessed teachers using classroom observations, many now also use value-added, a relatively new and controversial measure of teachers' contributions to student learning on standardized tests.[‡‡] Both intuition and theory suggests that these measures should align with one another to some degree; teachers who demonstrate stronger instruction in the classroom should also have a greater positive impact on their students' test achievement. In this paradigm, research would tend to show a strong correlation between observation ratings and value-added. Most existing empirical work, however, finds low correlations between these measures (Kane & Staiger, 2012). From a practical perspective, these results can be problematic, as weak correlations suggest to stakeholders that one or both measures may not be valid proxies for teacher quality. This implication can undermine the trust in an evaluation system and make it more politically difficult to use evaluations for personnel decisions, such as informing compensation and tenure.

There are three primary explanations for why correlations between value-added and observations scores may be low. The first is that one or both measures are poor proxies, even when measured without error, for the underlying trait they are intended to measure: teacher quality. A second, related issue, is that value-added and classroom observation scores might be measuring fundamentally different dimensions of teacher quality. Finally, each measure has one or more sources of error. Here the measures themselves may be valid, but their reliability could be poor. Any of these would lead to attenuated correlations between the two measures.

Our paper investigates these explanations using simulated data. Specifically, we explore levels of correlation between simulated value-added and observation scores when varying two factors: the correlation of each measure of performance to one or more unobserved dimensions of "teacher quality", and the amount of error in value-added or observation scores. By examining the multiple causes for low correlations between the two measures, we hope to provide researchers and policymakers a better understanding of the components of new teacher evaluation systems and how they might be expected to relate to one another.

Research Design:

[‡‡] While value-added is new as a measure used in formal evaluations of teachers, it has long been used as a means of assessing both educational productivity and the effects of specific schooling inputs (e.g. Hanushek, 1971; Murnane, 1981), and, in fact, to assess the implications of differences amongst individual teachers and extent to which individual teachers explain the variation in student achievement (e.g. Goldhaber, Brewer, & Anderson., 1999; Hanushek, 1992; Nye, Konstantopoulos, & Hedges, 2004).

To investigate the different reasons that might explain low correlations between value-added (VA) and observation (OBS) scores, we simulate 4500 sets of data for 200 teachers under different parameters, each designed to capture possible conditions where VA and OBS can arise.

To explore the possibility that low correlations exist because VA and OBS are poor proxies for teacher quality (TQ), we vary the correlation between simulated scores for each measure with a single simulated dimension of TQ[§§]. We then explore the extent to which the VA and OBS measures are correlated with one another when we vary the correlations between each measure to TQ, allowing for high, middling, and low correlations between each, resulting in 9 conditions.

To examine the effect of error in VA and OBS on correlations between the two metrics, we add error to the teacher scores created under the above conditions. Score variants include: VA generated from (1) 20 students or (2) 40 students, and; OBS generated from (1) three lessons with low reliability, (2) nine lessons with low reliability, (3) three lessons with high reliability, and (4) from nine lessons with high reliability. In total, we create three VA and five OBS correlated with TQ at high, middling, or low levels for each teacher, with one of these VA and one of these OBS being generated without error.

Data Analysis:
When creating the simulated data for each of the 4500 iterations, we first begin by randomly generating a TQ score ($\mu = 0, \sigma_{TQ} = 1$) for each of teacher. We then randomly generate VA ($\mu = 0, \sigma = \sigma_{VA}$) and OBS ($\mu = 0, \sigma = \sigma_{OBS}$) for each teacher, correlated with TQ at either a high ($\rho_{TQ,x} = 0.75$), middling ($\rho_{TQ,x} = 0.50$), or low ($\rho_{TQ,x} = 0.25$) level. We then correlate the randomly generated VA and OBS to explore the impact that varying the correlation of each to TQ would have on the correlation between the two proxies for TQ.

We then investigate the impact of error on correlations between VA and OBS by creating each score with error. For each teacher, we assigned 'classrooms' of 40 students, each with an 'achievement' score, $ACHIEVE_{st}$, created from the following equation:

$$ACHIEVE_{st} = \psi_s + \text{VA} \qquad (1)$$

The outcome, $ACHIEVE_{st}$, for each student *s* with teacher *t,* is a function of teacher *t*'s VA from above, and a randomly generated student effect ψ_s ($\mu = 0, \sigma = 3.33 * \sigma_{VA}$)[***]. We then calculate VA with error from either 20 or 40 of the students in the classroom using the following multilevel equation, where students are nested within teachers:

$$ACHIEVE_{st} = \text{VA}' + \varepsilon_{st} \qquad (2)$$

VA' represents the teacher's VA with error, generated from either 20 or 40 students.

For each teacher, we also create nine 'lesson observation' scores, LES_{lt}, with either low- or high-levels of reliability:

$$LES_{lt} = \lambda_l + \text{OBS} \qquad (3)$$

The outcome, LES_{lt}, for each lesson *l* taught by teacher *t,* is a function of teacher *t*'s OBS from above, and a randomly generated lesson effect λ_l, with a mean of zero and either a high relative

[§§] Future analyses will also add a second dimension of TQ and similarly vary its correlation to VA and OBS.
[***] We arrived at a student effect with this standard deviation based on common values in the literature. Typically, teacher effects on achievement for math are between 0.10 to 0.20, and student effects are around 0.50. The formula above assumes a student effect that is approximately 3.33 times the teacher effect (i.e., $\sigma_{VA} \approx 0.15$ if $\sigma = 0.50$).

value for the low reliability condition ($\sigma = 1.72 * \sigma_{OBS}$)[†††] or a low relative value for the high reliability condition ($\sigma = 1 * \sigma_{OBS}$). We then calculate OBS with error from either three or nine of the lessons using the following multilevel equation, where lessons are nested within teachers:

$$LES_{lt} = \text{OBS}' + \varepsilon_{lt} \quad (4)$$

OBS' represents the teacher's OBS with error, generated from either 3 or 9 lessons, and with either low- or high-levels of reliability. Overall, we recover three VA and five OBS from our simulation. We run correlational analyses on these scores to determine how each variation in score generation impacts correlations between VA and OBS.

Findings / Results:

(Please insert Figure 1 here). Figure 1 shows the results from our investigation into how varying the correlation of VA or OBS to TQ impacts the observed VA and OBS correlation. We find, as expected, that weaker correlations of VA or OBS to TQ result in weaker observed correlations between VA and OBS. Even with both proxies relating to TQ at a high level ($\rho_{TQ,x} = 0.75$), however, we still see that the average correlation between VA and OBS is approximately 0.55. In fact, ignoring all other factors that influence correlations, Figure 1 supports the arguments that weak observed correlations in literature may arise from the possibility that both VA and OBS are poor proxies for TQ, or that each are proxies for different dimensions of TQ. This finding is depicted by the distribution of correlations between VA and OBS in the subgraphs "Low-Low", "Low-Mid", and "Low-High", which demonstrate correlations that fall mainly within the range of observed correlations from other empirical studies (i.e., $0 < \left|\rho_{VA,OBS}\right| < 0.30$).

(Please insert Figure 2 here). (Please insert Figure 3 here.) Figures 2 and 3 show the correlations between VA and OBS when adding error to the teacher scores. For Figure 2, we specifically observe correlations when adding error to VA and OBS that are correlated to TQ at the same level (e.g., when both are highly correlated to TQ). We see that when both measures serve as poor proxies for TQ, very little difference exists in the amount of correlational attenuation between different combinations of scores, even when using scores with less error (i.e., more lessons used to generate OBS, more students to generate VA, or high reliability OBS scores). Second, matching intuition, correlations between VA generated from more students and OBS generated from more videos with higher reliabilities demonstrate less attenuation. Finally, the factor that appears to have the largest impact on correlational attenuation is the reliability of the observation scores; for the Mid-Mid and High-High conditions, four of the five lowest attenuated correlations arose between VA and OBS with higher reliability (i.e., scores with the "-H" suffix).

Even in the best case scenario (i.e., VA from generated from 40 students and OBS generated from nine lessons with high reliability), error attenuates correlations between VA and OBS for the "High-High" condition a little under 0.10; considering the typical classroom size, and the fiscal and temporal cost of performing teacher observations, this suggests that error has a non-negligible impact on observed correlations. In the worst case for the "High-High" condition, correlations are attenuated almost up to 0.30. If average correlations in this condition between

[†††] Our low reliability condition assumes that 25% of teacher observation scores are due to true differences between teachers, as opposed to error due to differences between construct irrelevant sources like lessons or raters. Our high reliability condition assumes that 50% of teacher observation scores are due to true differences. These values approximate those found in a generalizability study conducted on a math-specific observation instrument (Hill, Charalambous, & Kraft, 2012).

VA and OBS are approximately 0.55 without error, as we see above, the average observed correlation with error will be approximately 0.25. Thus, even if VA and OBS are both fairly *good* proxies of TQ, reported weak correlations between the metrics may be caused by error.

Figure 3 shows the average amount of attenuation for correlations between VA and OBS when adding error to scores generated when the correlation of VA and TQ is high and the correlation of OBS and TQ is low ("High-Low"), and when the correlation of OBS and TQ is high and the correlation of VA and TQ is low ("Low-High"). We find that the pattern of attenuation is similar across both conditions, despite differing levels of construct-irrelevant error (i.e., error from the student effect on VA, or error from the lesson effect on OBS) between the two metrics. Again we see that the factor that appears to have the largest impact on correlational attenuation is the reliability of the observation scores, with four of the five lowest attenuated correlations arising between VA and OBS with higher reliability. For the complete set of correlations between every combination of VA and OBS, please see Tables 1-9 in the appendix.

Conclusions:

Value-added and classroom observations are key components of new teacher evaluation systems, and are widely used as proxies for teacher quality. That research has mainly shown the measures to be weakly correlated, however, questions their validity as proxies. Some research has proposed that the weak correlations result from the fact that value-added and observations measure different underlying teacher traits, or that such empirical findings are products of error. In our paper, we use simulated data to systematically analyze three different possibilities for low correlations: both value-added and observation scores may be poor proxies of teacher quality, both measures might be good proxies for different dimensions of teacher quality, and error.

We find that the first two explanations can result in the correlations between value-added and observations that are seen in extant literature; when one measure strongly correlates to teacher quality and the other measure weakly correlates, or when both measures weakly correlate, correlations between value-added and observations are low.

This result contradicts theory and intuition, however; that teacher quality should not be somehow related to one or both of a teacher's instructional quality and his or her impact on student achievement seems unlikely. We considered instead the impact of error in scores on the correlation between the two metrics. Our analysis found that error could result in highly attenuated correlations, even when both value-added and observation scores served as good proxies for teacher quality. Furthermore, the factor that contributed most strongly to counteracting the attenuation of correlations due to error appeared to be the underlying level of reliability of observation scores. From this we conclude that for researchers worried about error impacting their analyses of different measures of teacher quality, and policymakers who are worried about constructing a coherent, concordant teacher evaluation system, more weight should perhaps be placed on developing an inherently reliable observation instrument, instead of investing in more expensive components of evaluation systems (i.e., observations). Future research should explore the different explanations for weak correlations between value-added and observation scores using real data, and also on a wider set of parameters.

Appendices

Not included in page count.

Appendix A. References

References are to be in APA version 6 format.

Goldhaber, D. D., Brewer, D. J., & Anderson, D. J. (1999). A three-way error components analysis of educational productivity. *Education Economics*, *7*(3), 199-208.

Hanushek, E. (1971). Teacher characteristics and gains in student achievement: Estimation using micro data. *The American Economic Review*, 280-288.

Hanushek, E. A. (1992). The Trade-Off between Child Quantity and Quality. *Journal of Political Economy*, *100*(1), 84–117.

Hill, H. C., Charalambous, C. Y., & Kraft, M. A. (2012). When rater reliability is not enough teacher observation systems and a case for the generalizability study. *Educational Researcher*, *41*(2), 56-64.

Kane, T. J., & Staiger, D. O. (2012). Gathering feedback for teachers: Combining high-quality observations with student surveys and achievement gains. *Policy and practice brief prepared for the Bill and Melinda Gates Foundation*.

Murnane, R. (1981). Interpreting the evidence on school effectiveness. *The Teachers College Record*, *83*(1), 19-35.

Nye, B., Konstantopoulos, S., & Hedges, L. V. (2004). How large are teacher effects?. *Educational evaluation and policy analysis*, *26*(3), 237-257.

Appendix B. Tables and Figures

Not included in page count.

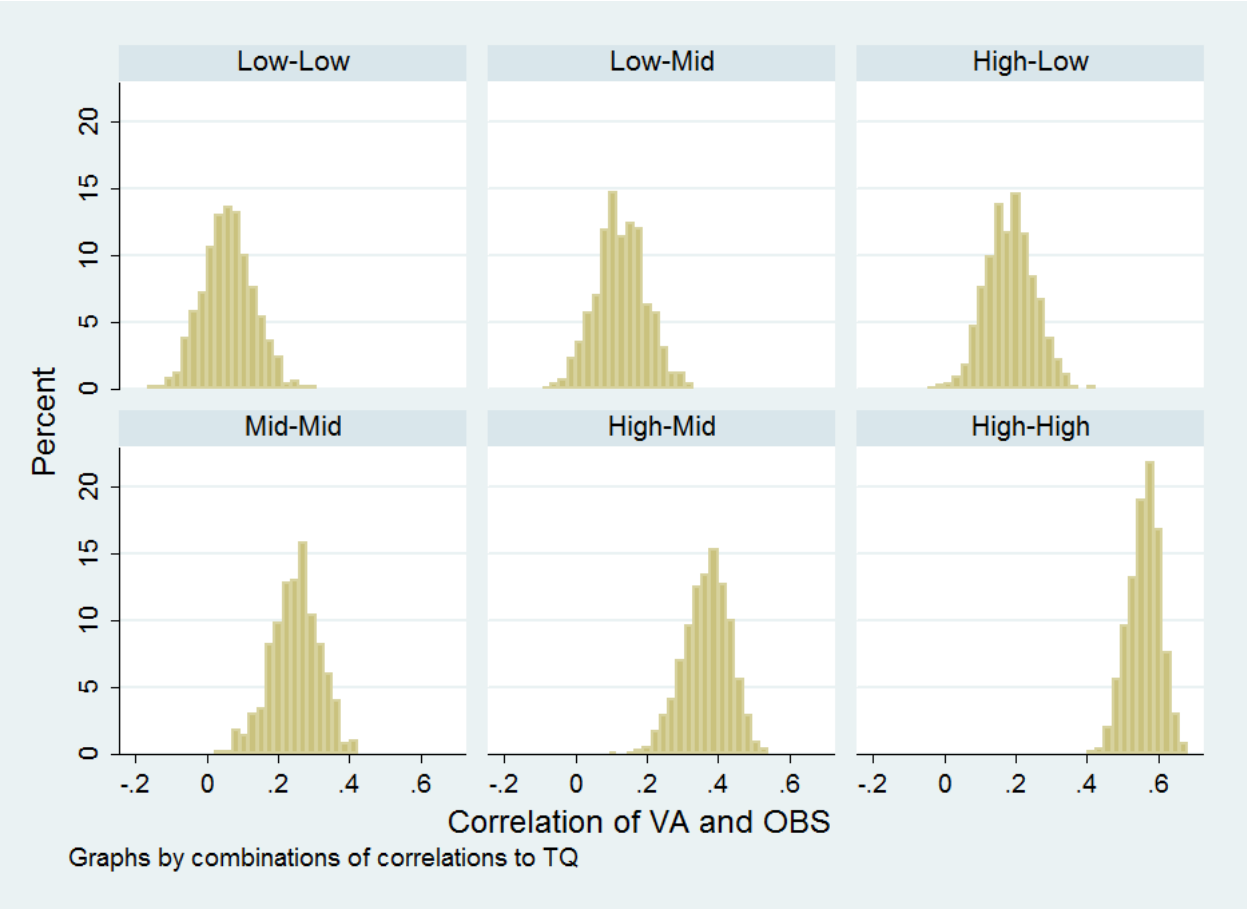

Figure 1. Histogram of correlations between value-added and observation scores, grouped by correlations of each measure to teacher quality.

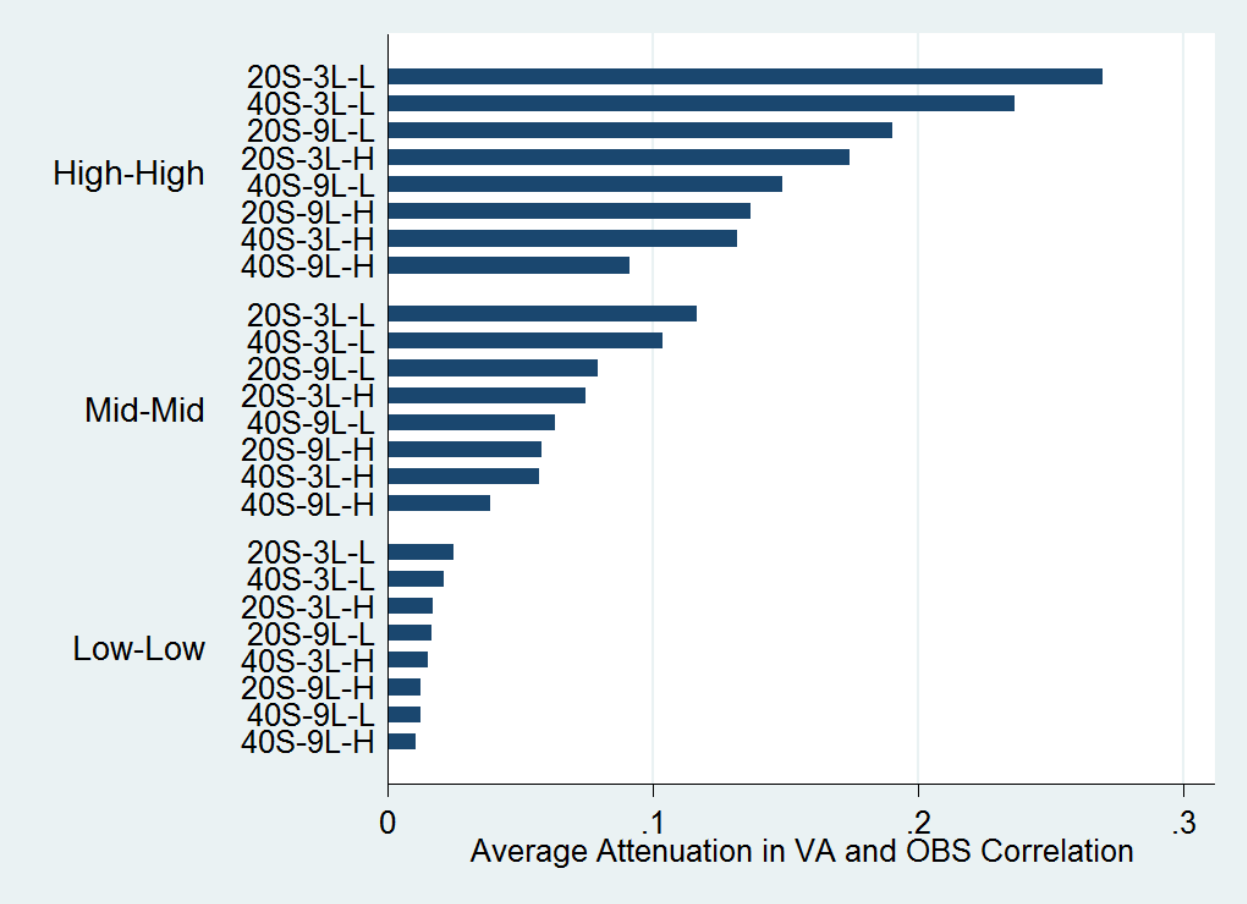

Figure 2. Average amount of attenuation to the correlation of value-added and observation scores when adding different amounts of error to scores that correlate at the same level to teacher quality. Categories should be interpreted as "#Students-#Lessons-High/Low Reliability".

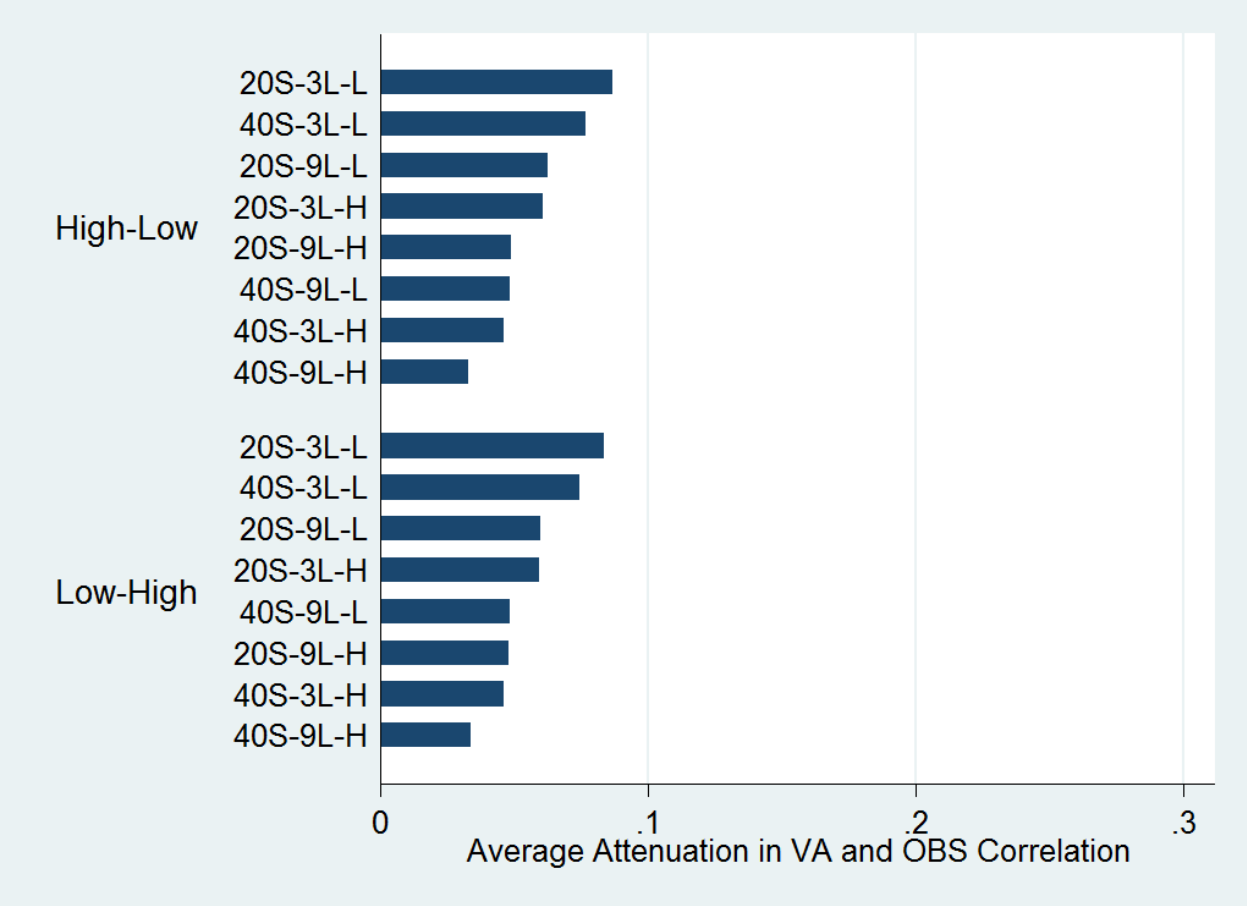

Figure 3. Average amount of attenuation to the correlation of value-added and observation scores when adding different amounts of error to scores that correlate at different levels to teacher quality. Categories should be interpreted as "#Students-#Lessons-High/Low Reliability".

Table 1. Correlations of different value-added and observation scores, "High-High"

	VA Score		
OBS Score	No Error	20 Students	40 Students
No Error	0.56	0.45	0.50
Low Reliability			
3 Lessons	0.37	0.29	0.33
9 Lessons	0.47	0.37	0.41
High Reliability			
3 Lessons	0.49	0.39	0.43
9 Lessons	0.53	0.43	0.47

Table 2. Correlations of different value-added and observation scores, "Mid-Mid"

	VA Score		
OBS Score	No Error	20 Students	40 Students
No Error	0.19	0.15	0.16
Low Reliability			
3 Lessons	0.13	0.10	0.11
9 Lessons	0.16	0.13	0.14
High Reliability			
3 Lessons	0.16	0.13	0.14
9 Lessons	0.18	0.14	0.15

Table 3. Correlations of different value-added and observation scores, "Low-Low"

	VA Score		
OBS Score	No Error	20 Students	40 Students
No Error	0.06	0.05	0.06
Low Reliability			
3 Lessons	0.04	0.04	0.04
9 Lessons	0.05	0.05	0.05
High Reliability			
3 Lessons	0.05	0.04	0.05
9 Lessons	0.06	0.05	0.05

Table 4. Correlations of different value-added and observation scores, "High-Mid"

	VA Score		
OBS Score	No Error	20 Students	40 Students
No Error	0.37	0.29	0.33
Low Reliability			
3 Lessons	0.24	0.19	0.21
9 Lessons	0.31	0.24	0.27
High Reliability			
3 Lessons	0.32	0.25	0.28
9 Lessons	0.35	0.28	0.31

Table 5. Correlations of different value-added and observation scores, "Mid-High"

	VA Score		
OBS Score	No Error	20 Students	40 Students
No Error	0.37	0.30	0.33
Low Reliability			
3 Lessons	0.24	0.19	0.21
9 Lessons	0.31	0.25	0.27
High Reliability			
3 Lessons	0.32	0.26	0.28
9 Lessons	0.35	0.28	0.31

Table 6. Correlations of different value-added and observation scores, "Mid-Low"

	VA Score		
OBS Score	No Error	20 Students	40 Students
No Error	0.13	0.11	0.12
Low Reliability			
3 Lessons	0.08	0.07	0.07
9 Lessons	0.11	0.09	0.09
High Reliability			
3 Lessons	0.11	0.09	0.10
9 Lessons	0.12	0.10	0.11

Table 7. Correlations of different value-added and observation scores, "Low-Mid"

	VA Score		
OBS Score	No Error	20 Students	40 Students
No Error	0.13	0.10	0.11
Low Reliability			
3 Lessons	0.08	0.07	0.08
9 Lessons	0.11	0.09	0.10
High Reliability			
3 Lessons	0.11	0.09	0.10
9 Lessons	0.12	0.10	0.11

Table 8. Correlations of different value-added and observation scores, "High-Low"

	VA Score		
OBS Score	No Error	20 Students	40 Students
No Error	0.18	0.15	0.16
Low Reliability			
3 Lessons	0.13	0.10	0.11
9 Lessons	0.16	0.13	0.14
High Reliability			
3 Lessons	0.16	0.13	0.14
9 Lessons	0.18	0.14	0.16

Table 9. Correlations of different value-added and observation scores, "Low-High"

	VA Score		
OBS Score	No Error	20 Students	40 Students
No Error	0.18	0.15	0.16
Low Reliability			
3 Lessons	0.13	0.10	0.11
9 Lessons	0.16	0.12	0.14
High Reliability			
3 Lessons	0.16	0.13	0.14
9 Lessons	0.17	0.14	0.15