Assessing the Consistency of Average Student Growth Using a Split Classroom Approach

Neal Kingston
University of Kansas

Fei Gu
McGill University

Wenhao Wang
Hongling Lao
University of Kansas

Introduction

The United States Department of Education is promoting the use of student growth on summative assessments as part of teacher evaluation. As in any measurement model there are multiple sources of error variance that in this case will detract from our ability to assess the true average student growth attributable to a teacher.

In this study, using data from a state testing program, the reliability of average student growth attributable to teachers was assessed in two ways: within year and across years. Within year all children are taught by a teacher in the same classroom(s) in a given year. Their current year test scores are compared with their previous year scores (when they typically had different teachers). Each class is divided into random halves and the two average student growth scores are calculated for each teacher and those scores are correlated across teachers with the same class size (class size will affect the reliability of average class growth similar to how test length affects test score reliability). We call this the split-class method. In order to calculate cross-year correlations, an average student growth score was calculated for each teacher based on all students' growth scores of each teacher from 2010 to 2011 and then again from 2011 to 2012, and the two cross-year growth scores of all teachers are correlated within a similar range of class sizes.

The sources of error variance are different with these two designs. In the first design, within each classroom there is no variation in how the class is taught. In a cross-years design the teacher or school might make changes to instruction or curriculum from one year to another. Similarly there is no differential variation in teacher health, personal circumstances, or attitudes in the split-class design, but might be in the cross-years design. In addition, the makeup of the classroom might affect a teacher's ability to facilitate learning from one year to the next in the cross-year design.

In this study we will assess the variance due to within-year and cross-year factors using three different common growth models.

Review of Literature on a Subset of Student Growth Models

Simple Gain Model

Operations of the simple gain model are straightforward. Growth is defined as the difference between a student's current and prior score (Auty, Bielawski, Deeter, et al., 2008). It is also called difference scores or gain scores, regarding the differences between prior (pretest) and current (posttest) scores. The actual growth can be further used in growth-to-standard models to determine whether a student's growth is adequate by comparing to target growth (Betebenner, 2009). Furthermore, a group growth can be estimated by aggregating individual difference scores at the teacher, school, or district level.

Since the growth is the difference between current and prior scores, the method requires a vertical scale for meaningful interpretation of the gain. Vertical scaling is the process to place scores onto a common scale for tests that measure the same or similar constructs if the tests do not have a common scale at the beginning. For example, this requirement has led to the adaptation of normalized scores (z scores) within grade level in some cases. Goldschmidt, Choi, and Beaudoin (2012) pointed out that this practice assumes that achievement standards across grades are vertically moderated. However, this vertical scaling practice should be adopted with caution. In a study by Tong and Kolen (2007), they compared 11 scaling methods with both real and simulated data. It showed that the 11 scaling methods were able to retain the general characteristics using simulated data when the assumptions are met. However, for the real data, the 11 methods produced vertical scales that showed decelerating growth from lower to higher grades. For the within-grade variability, different scaling methods produced different results. For instance, the Thurstone method produced enlarging variability over grades, whereas the IRT method produced fluctuating or decreasing variability over grades.

The advantage of this method is that its calculation is simple and transparent. It provides a direct estimate of student growth (Auty, Bielaswski, Deeter, et al., 2008). However, gain scores have been criticized as biased and inherently unreliable from theoretical considerations, as well as empirical studies designed to investigate the reliability of measured gains (e.g., Cronbach & Furby, 1970; Lord & Novick, 1968; Lord, 1956, 1963). On the other hand, Rogosa and Willett (1983) demonstrated that the reliability of the simple gain model is respectable when the individual differences in true changes are big enough, which supports Zimmerman and Williams (1982) claims that "gain scores in research can be highly reliable." The reliability of the difference score can be greater than that of the pretest and posttest scores when interperson variability in true change is large (Willett, 1988). Moreover, Willett (1988) also argued that even if the difference scores were always unreliable, this would not necessarily be a problem for the measurement of within-person change. Williams and Zimmerman (1996) tried to examine the reliability and validity of simple gain model from a statistical perspective within the framework of classical test theory. Specifically, Williams and Zimmerman (1996) admitted that many difference scores are unreliable. In practice, however, the reliability of a test score is determined by a number of different factors (e.g., the test construction procedure, the nature of the instrument), and "in this respect a difference between scores is similar".

In terms of validity, Williams and Zimmerman (1996, p.11) argued that the validity of difference scores is higher than formerly believed and "the existence of valid difference scores cannot be ruled out by statistical arguments alone."

This model is flawed by several disadvantages. First, it ignores the role of the school context (Burstein, 1980). Students with the same background tend to cluster in the same school. The clustering effect would potentially bias the aggregated estimates of school effects (Raudenbush & Willms, 1995). Particularly, estimates are biased when the intraclass correlation between the students and the schools is greater than zero (Aitkin & Longford, 1986). Second, this model ignores the teacher effect as well. McCaffrey and his colleagues (2009) showed that 50 percent variations of the student scores were explained by the teacher effect in elementary schools, whereas 70 percent variations by the teacher effect in middle schools. Finally, this model ignores the student difference in starting points. Thus, the method is criticized as "growth to nowhere" (Goldschmidt, Choi, & Beaudoin, 2012).

Student Growth Percentiles (Colorado Growth Model)

The student growth percentiles (SGP) model is a normative quantification of individual student growth, proposed by Betebenner (2008a). It has been adopted by 12 states, while 13 other states show some interest (Betebenner, 2010a). It requires external criteria to decide whether student growth percentiles as "adequate" or "enough" to reach desired achievement standards. This model describes how typical a student's growth is by comparing his/her current achievement to his/her *academic peers* with the same previous assessment score. It estimates the probability of observing a student's current achievement conditioned on their prior achievement.

In the SGP model, students are compared with their academic peers (who have the same prior scores) only, regardless of their actual prior scores. It is a conditional status based on students' prior scores. If the student's current score exceeds the scores of most of their academic peers, they have done well in a normative sense, at a high percentile under that conditional distribution (Betebenner, 2011). Similarly, two students with the same percentile score for the current year might not have the same absolute amount of growth if they had different prior test scores. A student's current year score is situated normatively as a student's growth percentile at time t taking into account student performance at time 1, 2, ..., up to t-1. Because the SGP model requires a large amount of data to generate sufficient coverage across the percentiles, Betebenner (2009, 2010b) also developed method to smooth the conditional distribution when sample size is not big enough.

In reality, students are nested within schools such that group level aggregation is involved. Betebenner (2008b) recommended the use of the median as a "typical" student to represent the growth of all students at the school. Due to the ordinal nature of percentile ranks, means are inappropriate to use because it assumes an interval scale underlying the averaged unites. However, Castellano and Ho (under review) argued that strict equal-interval properties are rare and the inferences and properties of means may be useful even when scales are quasi-interval. By contrasting the median and mean SGP models with two real statewide data, they found four percentile ranks dissimilarity for a school's ranking between these two aggregation functions, at worst with 30 percentile ranks difference.

The first advantage of the SGP model is that scores across years are not required to be vertically scaled, even though contiguous prior test scores are generally required for the SGP model (Goldschmidt et al., 2012). On the other hand, Castellano and Ho (under review) argued that a vertical scale may be required to make growth inferences. The second advantage is that it is more robust to outliers than OLS regression (Betebenner, 2011), despite the fact that they may also be affected by outliers, and sometimes estimating extreme conditional quantiles is required. Third, SGPs are described to be invariant to monotonic transformations of the test scales, supported by Briggs and Betebenner's (2009) study of the scale invariance of SGPs at the aggregate level.

Finally, other advantages include that the normative interpretation of student growth is easy for stakeholders to understand, and it is easy to aggregate individual data to higher units (e.g., teachers and schools). However, more properties of SGPs are to be further explored, such as sensitivity of SGPs to spline parameterization, bias and invariance under various sample sizes and covariate inclusion decisions, as suggested by Castellano and Ho (under review).

ANCOVA Model

The analysis of covariance (ANCOVA) model is designed to separate the effects of confounding variables from the interested treatment effect on the dependent variables. The covariate adjusted model can show the posttest difference among students who had the same pretest score (Goldschmidt, Choi, & Beaudoin, 2012). This model does not provide results in terms of growth as that in simple gain model. Instead, it intends to address explicitly the current student achievement accounting for differing prior achievement (Wright, 2008), and establish associations between students' average conditional status and classroom/school membership (Castellano & Ho, under review).

Student's current achievement is affected by many factors (e.g., ability, socioeconomic status [SES], motivation), in addition to teacher and school effect. ANCOVA is thus adopted to separate the effects of different covariates (e.g., SES, ability which is indicated by prior achievement) on the variable of interest (e.g., teacher, school effect). It is worth noting that the estimation procedure in random effect models assumes that there is no correlation between group-level effects and student prior scores (Castellano & Ho, under review). Although this assumption is usually violated in practice, random-effect models are often used in the value-added model (VAM) (Kim & Fees, 2006). In fact, ANCOVA is a popular tool for the VAM, which assesses how much students have learned during a time frame instead of how much they know at a specific point (status model). It is obvious there is a distinction between the intention/function of ANCOVA and the purpose of VAM. Wright (2008) explicitly pointed out the danger of using ANCOVA in VAM.

One advantage of the ANCOVA model is that it does not require a vertical scale as simple gain model does. It is more robust to either vertical or non-vertical scales (Wright, 2008). In addition, it estimates individual achievement and group level effect (i.e., teacher-effect) simultaneously.

If the covariate is measured with error, the ANCOVA model is likely to produce biased results (McCaffrey et al., 2004). More specifically, Wright (2008) pointed out that the negative consequences include (1) the estimated slope is biased toward zero; (2) the estimated teacher effects are biased toward the value that is estimated in a status model instead of a VAM; (3) the estimated teacher effects would be highly correlated with students' socioeconomic status. Generally, to ameliorate the bias in estimates, including multiple prior assessment scores (the same subject and other subjects as well) into the model is an effective method (Wright, 2008).

Methods

This study compares the consistency of ranking teachers using growth scores obtained from the three different growth models just described. First, teachers' growth scores in 2011 and 2012 were calculated, and the correlation between teachers' growth scores in these two years was used as a measure of the between-year consistency. Second, a teacher's growth score based on half of his/her students (randomly selected) was calculated such that two growth scores can be obtained from the two halves for the same teacher. In other words, there are two growth scores for each teacher within every year. Then, the correlation can be calculated between the two growth scores across teachers. This correlation, called the half-class correlation, was used to investigate the within-year reliability. Moreover, the number of students is varying across teachers, which may affect the teacher's growth score and thus the between- and within-year correlations. Herein, the between- and within-year correlations will be calculated using the teacher's growth scores from) all available teachers, and 2) different subsets of teachers who have a pre-specified number of students. In this section, the assessments, sample, and three models used for calculating growth scores will be introduced.

The Assessments

Summative assessments from 2009 to 2012 from one state assessment program were used to calculate the average student growth scores for teachers of 7th grade mathematics and English language arts. All items in these assessments were multiple-choice items with four options. The summative assessment scores across years are not vertically scaled. The simple gain model and the student percentile growth model use the summative assessment scores from two consecutive years, whereas the ANCOVA model use summative assessment scores from three consecutive years. Particularly, for a teacher who taught grade 7 math in the 2011-2012 academic year, his or her students' end-of-year summative math assessment scores from grade 7 (2012), 6 (2011), and 5 (2010) were used in the ANCOVA model to calculate this teacher's growth scores in 2012; and for a teacher who taught grade 7 reading in the 2010-2011 academic year, his or her students' end-of-year summative reading assessment scores from grade 7 (2011), 6 (2010), and 5 (2009) were used in the ANCOVA model to calculate this teacher's growth scores in 2011.

The Sample

Students who took end-of-year summative assessments each of three consecutive years (2009, 2010, and 2011 or 2010, 2011, and 2012) and whose teacher information is present were included in the sample. Table 1 shows the number of students and teachers in each grade and subject sample for the 2011-2012 and 2010-2011 academic year growth score calculation.

Table 1. Total number of students and teachers for each subject and year

Test and Year	Students	Teachers
Math 2012	31,202	920
Math 2011	30,985	915
Reading 2012	31,369	1,005
Reading 2011	31,309	1,098

The Models

Three models are considered in this paper. The first model is the ***simple gain*** model**.** It calculates the student growth score by differencing the end-of-year assessment scores from two consecutive years. The students who took two end-of-year summative math or reading assessments (either from 2011 and 2012 or from 2010 and 2011) are included. Since the end-of-year summative assessment score is not vertically scaled across years, the scores are first normalized (transformed to z-scores by student ranking) to overcome the non-comparability shortcoming of this model. The student simple gain growth score is:

$$Gain_i = Z(Y)_{it} - Z(Y)_{i(t-1)},$$

where $Z(Y)_{it}$ is the normalized assessment score for student *i* at time t (t = 2012, 2011). Then, the teacher growth score is calculated by averaging the growth scores of all his or her students.

The second model is the ***student growth percentile*** model. Since only two end-of-year summative assessment scores are needed, the students who took two end-of-year summative math or reading assessments (either from 2011 and 2012 or from 2010 and 2011) are used in this model. To calculate the growth percentile of a student, his or her end-of-year summative assessment in the previous year is chosen as a conditioning variable. Percentile ranks are calculated for current year scores for the group of students who had the same scores the previous year. Finally, a teacher's growth score is the median of the growth percentiles of all his or her students. Vertical scaling is not necessary for this model.

The third model is the **analysis of covariance** (ANCOVA).The covariates in this model are the two previous end-of-year summative assessments, controlling for the student's previous ability. Students who took all three end-of-year summative math or reading assessments (either from 2010, 2011 and 2012 or from 2009, 2010 and 2011) are included in the analyses for this model. The ANCOVA model is formulated as

$$Y_{it} = \beta_0 + \beta_1 Y_{i(t-1)} + \beta_2 Y_{i(t-2)} + \boldsymbol{\beta} TeacherID + \varepsilon_i,$$

where Y_{it}, $Y_{i(t-1)}$, and $Y_{i(t-1)}$ are summative assessment scores for student *i* at time t (t = 2012, 2011); β_0 is the intercept representing the growth score of the reference teacher[1]

[1] The reference teacher is determined by his or her order in the teachers sorted by the teacher ID.

when two previous assessment are 0; β_1 and β_2 are the coefficients for the two covariates; *Teacher ID* is categorical variable and $\boldsymbol{\beta}$ is the coefficient vector including the coefficients associated with teacher categories in the *TeacherID* vector; ε_i is the random error for student *i*. A teacher's growth score is calculated as the sum of the intercept and his or her teacher category coefficient.

Modeling the Relationship between Correlations and Number of Students in Sample

With the relatively small number of teachers whose average student scores were compared, correlations are not estimated with great stability. To reduce the noise associated with this issue, correlations were Fisher-Z transformed and then regressed on the natural log of class size. This logarithmic relationship was chosen to reflect the ceiling effect of the regression of correlations of average growth with class size. While correlations of average growth are expected to increase with class size (since the means are estimated more accurately in larger classes), but are limited to 1.0, this relationship seemed appropriate, though other functional forms might also be reasonable.

Correlations based on very small samples are highly variable. For example, for any class size that has only two teachers, correlations will be either positive one or negative one. Moreover, since the very low and very high class sizes are less common, these cases will have extreme influence when estimating regression coefficients. Therefore, for the split-class method class sizes with fewer than 10 teachers were omitted from the modeling of the relationship between class size and correlation.

For the cross-year method too many cases had fewer than 10 teachers so a different approach was used. Correlations were averaged for 5 ranges of sample sizes and the midpoint of the class sizes was used in the regression.

Results of these analyses can be found in Appendix E.

Using the regression equation for each model, predicted Fisher-Z transformed correlations were estimated for sample sizes of 10 to 100 and transformed back to the correlation metric.

Results

Appendices A and B present the correlations for the split-class method for Mathematics and Reading, respectively. There are two entries (in subsequent rows) for most half-class size: one from 2011 and one from 2012.

Appendices C and D present similar information for the cross-years method.

Appendix E show the regressing of Fisher-z transformed correlations for the three student growth models based on the split-class and cross-year methods for both mathematics and reading. Cross-year regressions are based on grouped class-size data and the middle of the range of class sizes is listed.

Table 2 presents the estimated correlations for different mathematics and reading class sizes based on the split-class data. For mathematics, correlations are about the same for the three growth models, with the ANCOVA model performing as well or slightly better than the other two methods at each class size. For reading the improvement with the ANCOVA model was greater. Since the ANCOVA model uses more data than the other two models its superior performance is not surprising.

Table 2. Predicted Reliability of Average Growth Scores at Different Sample Sizes based on the Split-Class Method

	Mathematics			Reading		
Number of Students	Simple Gain Scores	Student Growth Percentiles	ANCOVA	Simple Gain Scores	Student Growth Percentiles	ANCOVA
10	0.72	0.74	0.78	0.52	0.52	0.58
15	0.79	0.80	0.83	0.59	0.58	0.65
20	0.83	0.83	0.86	0.63	0.62	0.70
25	0.85	0.85	0.88	0.66	0.64	0.73
30	0.87	0.86	0.89	0.68	0.67	0.75
40	0.90	0.89	0.91	0.71	0.70	0.78
50	0.91	0.90	0.92	0.74	0.72	0.81
60	0.92	0.91	0.93	0.76	0.74	0.82
70	0.93	0.92	0.94	0.77	0.76	0.84
80	0.94	0.92	0.94	0.78	0.77	0.85
90	0.94	0.93	0.95	0.79	0.78	0.86
100	0.95	0.93	0.95	0.80	0.79	0.86

There is no clear-cut rule for how reliable a measure should be, but for most professionally developed assessments used to make important decisions about students reliability estimates are in the .90-.95 range. And even with reliabilities in that range professional testing standards say that multiple measures should be used for important decisions. For mathematics a split-class correlation of .90 is attained when average student growth scores are based on 40 or more students. For reading a .90 correlation is not reached even with 100 students.

Table 3 presents the estimated correlations for different mathematics and reading class sizes based on the cross-years data. For mathematics, the student growth percentiles method and the ANCOVA method performed about the same and outperformed the simple gain score method (especially for small class sizes). For reading the ANCOVA model performed better than the student growth percentiles method which in turn performed better than the simple gain scores method. It is particularly important to note that all methods showed much lower reliability using the cross years method. No method at any reported sample size had a reliability near .90. If the method used for predicting reliability based on sample sizes holds for significantly larger samples, it would require a teacher to have data from 11,464 students to achieve a reliability of .90 using the ANCOVA method. Clearly the additional variability associated with the same teacher's student average growth across years is a significant factor.

Table 3. Predicted Reliability of Average Growth Scores at Different Sample Sizes based on the Cross-Year Method

	Mathematics			Reading		
Number of Students	Simple Gain Scores	Student Growth Percentiles	ANCOVA	Simple Gain Scores	Student Growth Percentiles	ANCOVA
10	0.19	0.39	0.36	0.22	0.33	0.38
15	0.28	0.45	0.43	0.27	0.36	0.43
20	0.34	0.49	0.47	0.30	0.38	0.47
25	0.39	0.51	0.50	0.33	0.40	0.49
30	0.43	0.54	0.53	0.35	0.42	0.51
40	0.48	0.57	0.57	0.38	0.44	0.54
50	0.52	0.60	0.60	0.41	0.45	0.57
60	0.55	0.61	0.62	0.43	0.47	0.58
70	0.57	0.63	0.64	0.44	0.48	0.60
80	0.59	0.64	0.65	0.46	0.49	0.61
90	0.61	0.66	0.66	0.47	0.50	0.62
100	0.63	0.67	0.67	0.48	0.50	0.63

Discussion

Determining the fairness of a teacher evaluation system based in part or exclusively on the performance of that teacher's students is complex. Within-year variance estimated by the split-class method is largely due to variability in individual students, but may also be due in part to a particular teacher's ability to connect to some but not all students. Cross year variance estimated by the cross-years method includes the variance due to within year influences, but also includes other sources of variability in student growth, some of which might not be reasonably under teacher control.

Once can partition the variance associated with teacher evaluation scores based on student growth into true variance and multiple sources of error variance. The squared reliability tells us the proportion of true variance. The remaining variance is error associated with one or more factors. Using the ANCOVA growth model and a class size of 100 (the equivalent of a grade 7 mathematics teacher teaching four 25 student sections), the split-class correlation of .95 and cross-years correlation of .67 indicate 45% of the variability of teacher evaluation scores (assuming they are based on only student average growth) would be due to long-term teacher quality, 45% would be due to cross-year variability in student performance, and 10% would be due to the sample of students. A similar analysis for Reading raises even greater concerns, with 40% of the variability of teacher evaluation scores due to long-term teacher quality, 34% due to cross-year variability in student performance, and 26% due to the sample of students.

The second of the three sources of variability in the previous paragraph, cross-year variance, might or might not be under teacher control, but even if it is there are policy considerations. With a cross-year correlation of .7, which is higher than we see for any of the reading models for a teacher whose evaluation score is based on 100 students, 40% of the teachers who are evaluated as being in the top-quarter of all teachers one year will be evaluated as being outside the top-quarter the subsequent year. Moreover, 2% of those teachers who were in the top-quarter one year will be in the bottom quarter the next. This type of year-to-year variation in teacher ratings may raise questions of credibility for any such system.

References

Aitkin, M., & Longford, N. (1986). Statistical modeling issues in school effectiveness studies. *Journal of the Royal Statistical Society*, *149*, 1-43.

Auty, W., Bielawski, P., Deeter, T., Hirata, G., Hovanetz-Lassila, C., Rheim, J., Goldschmidt, P., O'Malley, K., Blank, R., and Williams, A. (2008). *Implementer's guide to growth models.* Washington, DC: Council of Chief State School Officers.

Betebenner, D. W. (2008a). Toward a normative understanding of student growth. In Ryan, K. E. & Shepard, L. A. (Eds.), *The future of test-based educational accountability* (pp. 155-170). New York, NY: Taylor & Francis.

Betebenner, D. W. (2008b). *A primer on student growth percentiles.* Retrieved from the Georgia Department of Education website: http://www.doe.k12.ga.us/

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

Betebenner, D. W. (2010a). *New directions for student growth models.* The National Center for the Improvement of Educational Assessment. Retrieved from http://www.ksde.org/LinkClick.aspx?fileticket=UssiNoSZks8%3D&tabid=4421&mid=10564

Betebenner, D. W. (2010b). SGP: Student growth percentile and percentile growth projection/trajectory functions [R package version 0.0-6].

Betebenner, D. W. (2011). A technical overview of the student growth percentile methodology: student growth percentiles and percentile growth projections/trajectories. The National Center for the Improvement of Educational Assessment. Retrieved from http://www.nj.gov/education/njsmart/performance/SGP_Technical_Overview.pdf

Briggs, D., & Betebenner, D. W. (2009, April). The invariance of measurement of growth and effectiveness to scale transformation. (Paper presented at the 2009 NCME Annual Conference, San Diego, CA.)

Burstein, L. (1980). The analysis of multi-level data in educational research and evaluation. *Review of Research in Education*, *4*, 158-233.

Castellano, K. E., & Ho, A. D. (Submitted). Simple choices among aggregate-level conditional status metrics: From median Student Growth Percentiles to value-added models.

Cronbach, L. J., & Furby, L. (1970). How we should measure change—or should we? *Psychological Bulletin*, *74*, 68-80.

Goldschmidt, P., Choi, K., Beuadoin, J.P. (2012). *Growth model comparison study: Practical implications of alternative models for evaluating school performance.* Washington, DC: Council of Chief State School Officers.

Kim, J.-S., & Frees, E. W. (2006). Omitted variables in multilevel models. *Psychometrika*, *71*, 659-690.

Lord, F. M. (1956). The measurement of growth. *Educational and Psychological Measurement*, 16, 421-437.

Lord, F. M. (1963). Elementary models for measuring change. In C. W. Harris (Ed.), *Problems in measuring change.* Madison, Wisconsin: University of Wisconsin Press.

Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores.* Reading MA: Addison-Wesley.

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., Sass, T. R., Lockwood, J. R., & Mihaly, K. (2009). The intertemporal stability of teacher effect estimates. *Education Finance and Policy*, *4*, 572–606.

Raudenbush, S., & Willms, D. (1995). The estimation of school effects. *Journal of Educational and Behavioral Statistics*, *20*, 307-335.

Rogosa, D. R., & Willett, J. B. (1983). Demonstrating the reliability of the difference score in the measurement of change. *Journal of Educational Measurement*, *20*, 4-???.

Tong, T., & Kolen, M. J. (2007). Comparisons of methodologies and results in vertical scaling for educational achievement tests. *Applied Measurement in Education*, *20*, 227-253.

Williams, R. H., & Zimmerman, D. W. (1996). Are simple gain scores obsolete? *Applied Psychological Measurement*, *20*, 1-???.

Wright, S. P. (2008). Estimating Educational Effects using Analysis of Covariance with Measurement Error. Paper presented at CREATE/NEI Conference, Wilmington, NC, October 2008. Accessed 06-25-2013 at http://www.createconference.org/documents/archive/2008/2008wright.pdf

Zimmerman, D. W., & Williams, R. H. (1982). Gain scores in research can be highly reliable. *Journal of Educational Measurement*, *19*, 149-154.

Appendix A

Split-Class Correlations For Average Mathematics Growth Scores by Half-Class Size

Simple Gain			SGP			ANCOVA		
Half Class Size	cor	n Teachers	Half Class Size	cor	n Teachers	Half Class Size	cor	n Teachers
1	0.339113	64	1	0.475773	64	1	0.508953	62
1	0.195387	86	1	0.363855	86	1	0.372632	83
2	0.455659	59	2	0.544483	59	2	0.510783	55
2	0.087718	65	2	0.328371	65	2	0.326894	63
3	0.389302	43	3	0.442627	43	3	0.463003	44
3	0.453304	48	3	0.486514	48	3	0.378873	46
4	0.361358	33	4	0.417864	33	4	0.719218	39
4	0.418394	40	4	0.399086	40	4	0.563795	40
5	0.42114	25	5	0.62947	25	5	0.500444	25
5	0.728085	28	5	0.513958	28	5	0.750803	30
6	0.594116	27	6	0.757515	27	6	0.744392	27
6	0.628976	34	6	0.626073	34	6	0.756348	43
7	0.45789	34	7	0.421674	34	7	0.781449	30
7	0.730692	38	7	0.819201	38	7	0.793961	36
8	0.628056	21	8	0.795379	21	8	0.790574	21
8	0.580084	22	8	0.670125	22	8	0.693941	25
9	0.769378	26	9	0.595875	26	9	0.758283	22
9	0.532623	29	9	0.713643	29	9	0.681061	30
10	0.315362	24	10	0.379842	24	10	0.816973	19
10	0.674661	25	10	0.805257	25	10	0.684035	25
11	0.687371	26	11	0.665976	26	11	0.718486	17
11	0.592313	27	11	0.75142	27	11	0.65102	32
12	0.885396	14	12	0.794073	14	12	0.824638	13
12	0.90037	16	12	0.952036	16	12	0.788731	20
13	0.661121	13	13	0.663226	13	13	0.71576	18
13	0.578136	14	13	0.757764	14	13	0.813146	19
14	0.722263	22	14	0.805988	22	14	0.90407	15
14	0.887443	25	14	0.858179	25	14	0.813264	21
15	0.742327	10	15	0.91876	10	15	0.785366	15
15	0.746064	13	15	0.73051	13	15	0.802291	15
16	0.82939	11	16	0.905083	11	16	0.877726	14
16	0.832145	17	16	0.897214	17	17	0.863816	16
17	0.825667	12	17	0.698405	12	18	0.83327	13
18	0.572497	10	18	0.671782	10	18	0.650495	15
18	0.668278	15	18	0.657131	15	19	0.844496	14
19	0.829002	13	19	0.840816	13	20	0.897263	13
20	0.912474	11	20	0.879332	11	21	0.692818	13
20	0.863225	13	20	0.913925	13	26	0.411237	10
22	0.773203	10	22	0.639895	10	26	0.862775	11
23	0.913772	12	23	0.702459	12	28	0.844493	11
27	0.77312	14	27	0.718984	14	29	0.948217	10
28	0.916515	11	28	0.859875	11	29	0.923283	13
29	0.901764	10	29	0.721028	10	30	0.932704	15
29	0.895475	11	29	0.804288	11	33	0.897613	12
31	0.950711	10	31	0.78435	10	34	0.932013	10
31	0.816028	11	31	0.915603	11	34	0.942848	11
32	0.774132	11	32	0.743047	11	35	0.887133	11
32	0.929113	12	32	0.935492	12	35	0.968638	12
34	0.843765	11	34	0.90761	11	37	0.951218	10
34	0.93967	16	34	0.914456	16	37	0.889711	11
35	0.902211	13	35	0.882958	13	40	0.945776	10
37	0.877104	13	37	0.808206	13	44	0.73812	10
38	0.920468	10	38	0.928527	10	47	0.942995	10
40	0.89764	10	40	0.914783	10			
42	0.963546	10	42	0.960662	10			
48	0.747217	12	48	0.89949	12			

Appendix **A**

Split-Class Correlations For Average Reading Growth Scores by Half-Class Size

Simple Gain			SGP			ANCOVA		
Half Class Size	cor	n Teachers	Half Class Size	cor	n Teachers	Half Class Size	cor	n Teachers
1	0.2767239	68	1	0.3100873	68	1	0.1692362	71
1	0.1282331	74	1	0.1906169	74	1	0.2428428	75
2	0.1418799	72	2	0.2141611	72	2	0.1221135	67
2	0.1735344	78	2	0.1785536	78	2	0.2508316	81
3	0.0268208	51	3	0.0722081	51	3	0.3796685	55
3	0.4514864	70	3	0.4161112	70	3	0.4651588	69
4	0.0919196	43	4	0.2648617	43	4	0.431321	48
4	0.5243141	50	4	0.4438747	50	4	0.4934049	51
5	0.3426105	40	5	0.4916137	40	5	0.6267075	38
5	0.4323448	43	5	0.3684835	43	5	0.2152187	42
6	0.3962734	31	6	0.726611	31	6	0.678785	36
6	0.5862775	41	6	0.567062	41	6	0.5997453	38
7	0.4045954	43	7	0.3809203	43	7	0.563706	39
7	0.5484988	44	7	0.6443191	44	7	0.6855766	45
8	0.3071249	25	8	0.4403604	25	8	0.2401871	38
8	0.5510242	44	8	0.3884332	44	8	0.4102778	45
9	0.5161323	28	9	0.5606502	28	9	0.7215542	23
9	0.7266411	36	9	0.5582989	36	9	0.2001155	25
10	0.5779355	26	10	0.3572946	26	10	0.5603148	24
10	0.0862462	32	10	0.3893955	32	10	0.3803035	32
11	-0.0309687	17	11	-0.1638414	17	11	0.5063456	15
11	0.397281	39	11	0.4124925	39	11	0.7180875	35
12	0.6390735	19	12	0.5252861	19	12	0.6493802	20
12	0.297408	22	12	0.6078869	22	12	0.586737	20
13	0.134635	18	13	0.4675414	18	13	-0.0401319	18
13	0.5644485	20	13	0.6871423	20	13	0.6153722	23
14	0.6112886	17	14	0.5886777	17	14	0.6024613	16
14	0.6578578	23	14	0.5937647	23	14	0.2907707	25
15	0.540653	14	15	0.4128881	14	15	0.498775	18
15	0.742135	16	15	0.5040965	16	16	0.6195927	10
16	0.768037	11	16	0.7288378	11	16	0.5152456	17
16	0.6935151	12	16	0.5662325	12	17	0.4270875	14
17	0.673474	10	17	0.6229878	10	18	0.3490528	15
17	0.8108423	13	17	0.5595857	13	18	0.6849591	22
18	0.4739297	18	18	0.645611	18	19	0.2700322	10
18	0.7066286	21	18	0.4165556	21	19	0.8364278	16
19	-0.1498238	11	19	0.2959357	11	21	0.5019906	21
19	0.5491448	13	19	0.724269	13	22	0.8286744	14
20	0.6637273	13	20	0.5962299	13	23	0.4491789	13
21	0.6523922	16	21	0.640959	16	23	0.8218857	13

Simple Gain			SGP			ANCOVA		
Half Class Size	cor	n Teachers	Half Class Size	cor	n Teachers	Half Class Size	cor	n Teachers
22	0.5119068	11	22	0.112913	11	24	0.529419	13
22	0.7451812	12	22	0.0842531	12	24	0.5682022	13
23	0.8235906	11	23	0.7022201	11	25	0.8057968	12
23	0.8544422	13	23	0.8073352	13	25	0.852086	14
24	0.800937	10	24	0.5571612	10	27	0.7663287	11
24	0.4645088	14	24	0.5850263	14	27	0.9014987	14
25	0.8082708	11	25	0.7455123	11	29	0.7129364	10
26	0.7456461	10	26	0.6263238	10	29	0.649876	13
26	0.184114	11	26	0.1332966	11	30	0.8760438	13
28	0.7840854	13	28	0.5015236	13	30	0.6850655	19
29	-0.0550664	10	29	-0.0395282	10	31	0.5095243	12
30	0.7130167	15	30	0.8407854	15	31	0.7446156	12
30	0.7428977	18	30	0.7158166	18	32	0.8654749	11
31	0.7744675	12	31	0.8545027	12	32	0.9216012	13
31	0.6815702	13	31	0.41585	13	33	0.903582	10
32	0.7259725	10	32	0.9181092	10	34	0.5481196	10
32	0.7033432	14	32	0.8528998	14	34	0.8689311	14
33	0.7012265	11	33	0.6565636	11	35	0.9120826	12
33	0.2371069	16	33	0.5494308	16	36	0.6201541	10
34	0.8735283	10	34	0.8522526	10	36	0.8810278	10
35	0.8524041	10	35	0.7062912	10	37	0.8398629	13
35	0.4793164	12	35	0.4339617	12	37	0.8443255	15
36	0.549105	13	36	0.2793141	13	39	0.7700116	12
36	0.5260243	14	36	0.7954917	14			
38	0.7973317	11	38	0.8102855	11			
38	0.9021404	11	38	0.4297282	11			
39	0.8177355	14	39	0.821403	14			
40	0.6709223	11	40	0.8896645	11			
42	0.542187	13	42	0.8394522	13			
44	0.7351228	10	44	0.8876877	10			

Appendix **C**

Cross-Year Correlations for Average Reading Growth Scores by Class Size

Simple Gain			SGP			ANCOVA		
Avg. Class Size	cor	n Teacher	Avg. Class Size	cor	n Teacher	Avg. Class Size	cor	n Teacher
1	0.905921	6	1	0.69792	6	1	0.054621	8
2	-0.41442	15	2	-0.20284	15	2	0.010682	12
3	-0.29587	14	3	-0.15257	14	3	0.086013	15
4	0.362707	22	4	0.398864	22	4	0.182582	19
5	0.405151	13	5	0.242489	13	5	0.63874	14
6	0.489579	16	6	0.55241	16	6	0.533171	16
7	0.834302	13	7	0.562306	13	7	0.796927	12
8	-0.13805	9	8	-0.21624	9	8	0.466509	14
9	0.255515	15	9	0.15351	15	9	0.026295	11
10	-0.19796	8	10	0.468703	8	10	0.63435	11
11	0.190625	12	11	0.503606	12	11	0.267944	11
12	0.568886	12	12	0.487065	12	12	0.495537	9
13	0.826029	9	13	0.754673	9	13	0.634385	6
14	-0.29263	9	14	0.556963	9	14	0.41639	14
15	0.2252	8	15	0.331873	8	15	0.654511	9
16	0.584763	9	16	0.772419	9	16	0.443051	12
17	0.30395	8	17	0.292771	8	17	-0.08793	6
18	-0.00537	10	18	0.393077	10	18	0.416472	10
19	-0.09575	9	19	0.444759	9	19	-0.65792	6
20	-0.05358	7	20	-0.3632	7	20	0.58309	7
21	0.6618	9	21	0.68677	9	21	0.808475	9
22	0.83442	5	22	0.742539	5	22	0.536372	5
23	0.587297	7	23	0.054326	7	23	0.86596	6
24	-0.30772	10	24	0.099254	10	24	0.319238	15
25	0.419006	13	25	0.322747	13	25	-0.09045	11
26	0.281818	8	26	0.188485	8	27	-0.07639	7
27	-0.66844	6	27	-0.62808	6	29	-0.0928	9
29	0.06337	7	29	0.347526	7	30	0.484965	6
30	0.059124	9	30	0.328962	9	35	0.352695	7
31	-0.26605	5	31	-0.24593	5	36	0.137928	8
33	0.500177	6	33	0.431511	6	39	0.912478	5
35	0.481387	5	35	-0.23444	5	41	-0.01491	8
36	-0.70033	7	36	0.257027	7	47	0.678223	8
39	0.073725	9	39	0.157576	9	48	0.534349	7
42	0.554113	5	42	0.472156	5	57	0.648201	8
48	0.682889	7	48	0.752314	7	58	0.682592	7
50	-0.77679	6	50	-0.17054	6	62	0.161729	5
57	0.643047	6	57	0.695224	6	64	0.650268	7
58	-0.47478	6	58	0.134029	6	65	0.804519	5

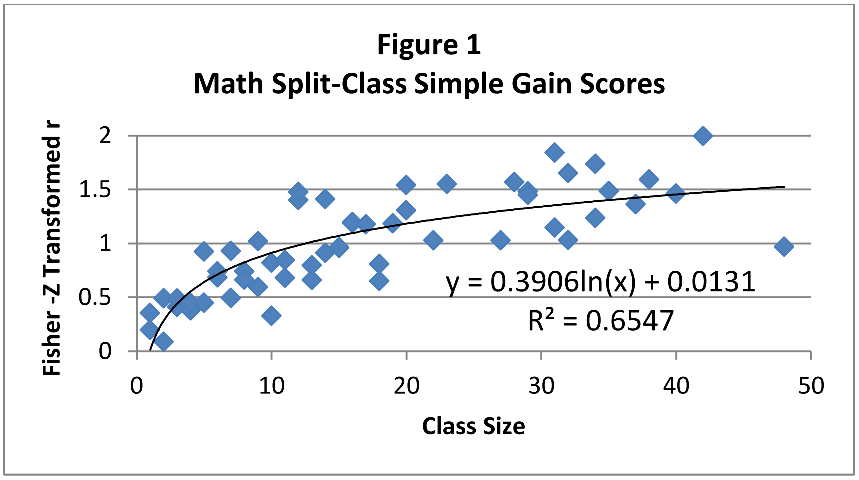
Simple Gain			SGP			ANCOVA		
Avg. Class Size	cor	n Teacher	Avg. Class Size	cor	n Teacher	Avg. Class Size	cor	n Teacher
60	0.835681	5	60	0.31407	5	66	0.442826	8
65	-0.16292	5	65	0.418368	5	67	0.728852	5
66	0.386766	6	66	-0.11066	6	69	0.035406	8
67	0.324562	5	67	-0.31007	5	70	0.766414	6
68	0.821166	7	68	0.632052	7	76	0.743402	9
70	0.101922	10	70	0.450757	10	78	0.479329	5
72	0.431179	5	72	-0.30512	5	79	0.556506	6
74	0.613328	6	74	0.850561	6	81	0.012887	6
77	0.727276	6	77	0.720766	6	83	-0.00565	6
80	0.761573	5	80	0.82916	5	94	0.730104	5
81	0.376463	6	81	0.443205	6			
84	0.179749	6	84	-0.204	6			
85	0.700276	5	85	0.946186	5			

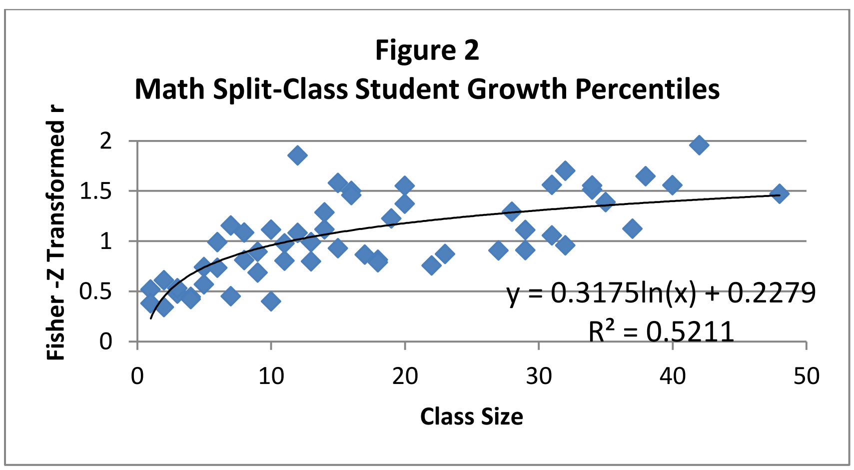
Appendix **D**

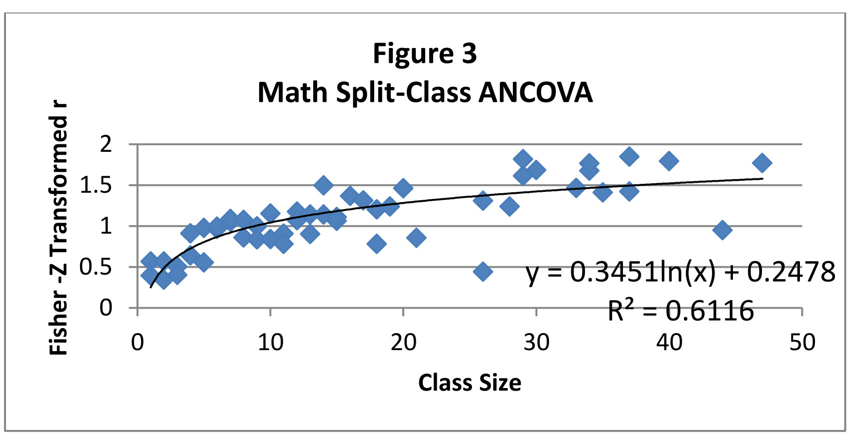
Cross-Year Correlations for Average Mathematics Growth Scores by Class Size

Simple Gain			SGP			ANCOVA		
Avg. Class Size	cor	n Teacher	Avg. Class Size	cor	n Teacher	Avg. Class Size	cor	n Teacher
1	-0.09889	12	1	-0.1619	12	1	-0.18969	15
2	-0.19519	23	2	0.222424	23	2	0.282256	19
3	0.283353	12	3	0.418171	12	3	0.322792	12
4	-0.49018	19	4	0.296858	19	4	0.226165	22
5	0.523052	17	5	0.621733	17	5	0.621743	17
6	0.521118	15	6	0.638762	15	6	0.469609	13
7	0.776615	7	7	0.810584	7	7	0.510306	9
8	0.642508	14	8	0.465801	14	8	0.907457	9
9	0.306388	10	9	0.330499	10	9	0.670783	15
10	0.048326	13	10	0.25009	13	10	-0.32044	11
11	-0.23563	8	11	-0.07244	8	11	0.816658	9
12	-0.00311	8	12	0.236766	8	12	0.614633	7
13	0.2989	8	13	0.719701	8	13	0.411485	12
14	-0.02386	8	14	0.062692	8	14	0.785133	11
15	0.759708	13	15	0.718979	13	15	0.347492	7
16	0.493097	10	16	0.304358	10	16	0.730697	12
17	0.147914	11	17	0.406506	11	17	0.212968	10
18	-0.27484	11	18	0.217632	11	18	-0.39293	6
19	0.120324	5	19	0.054262	5	20	0.435207	9
20	0.609123	5	20	0.551082	5	21	0.488257	7
21	0.491912	5	21	0.259202	5	22	0.107687	5
22	-0.56902	6	22	0.142228	6	23	0.681155	7
23	0.143177	5	23	0.533092	5	24	0.829446	7
24	0.247733	8	24	0.266242	8	25	0.256635	9
25	0.381086	8	25	0.315388	8	28	0.274923	9
27	0.785692	6	27	0.773709	6	33	0.545404	6
29	0.470146	5	29	0.619819	5	34	0.335108	7
30	0.596149	6	30	0.549077	6	35	0.322463	5
32	0.593817	5	32	0.777536	5	37	-0.02196	5
34	0.320568	7	34	0.296309	7	39	0.192068	5
35	0.94797	5	35	0.323208	5	53	0.589652	5
37	0.582309	7	37	0.446237	7	54	0.870412	5
38	0.118092	5	38	0.698384	5	63	0.02838	5
56	0.716074	8	56	0.817863	8	65	0.713079	6
65	0.88784	5	65	0.617971	5	68	-0.21237	5
83	0.800742	6	83	0.784796	6	75	0.943419	5
84	0.815621	7	84	0.934659	7	81	0.880909	8
94	0.913336	7	94	0.946971	7	91	0.883965	5
						92	0.969579	5

Appendix E
Regressing Fisher-z Transformed Correlations on Class Size







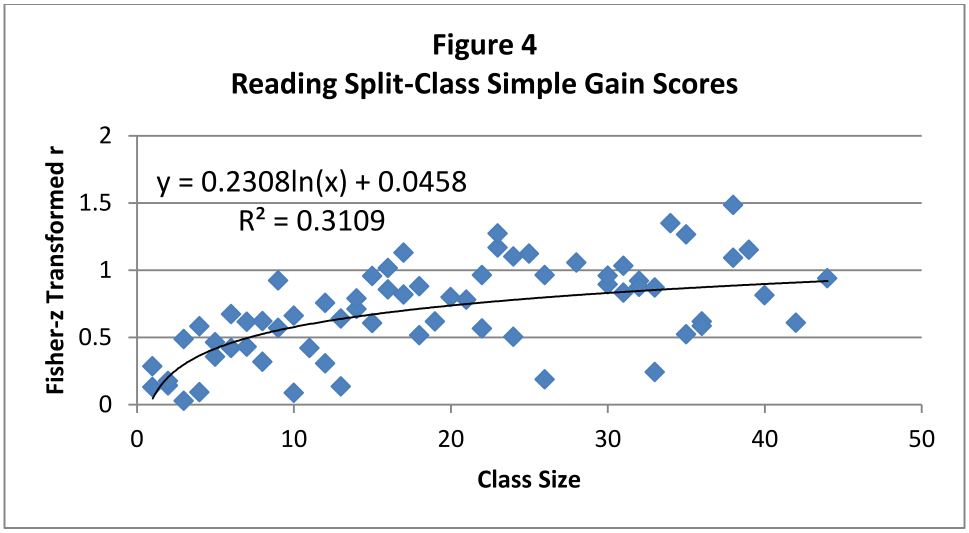


Figure 4
Reading Split-Class Simple Gain Scores
y = 0.2308ln(x) + 0.0458
R² = 0.3109
Fisher-z Transformed r
Class Size

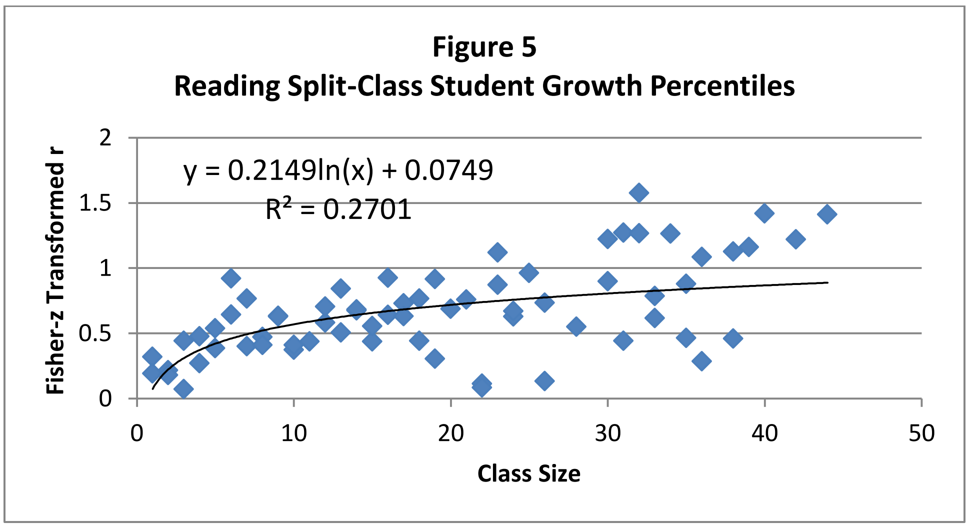


Figure 5
Reading Split-Class Student Growth Percentiles
y = 0.2149ln(x) + 0.0749
R² = 0.2701
Fisher-z Transformed r
Class Size

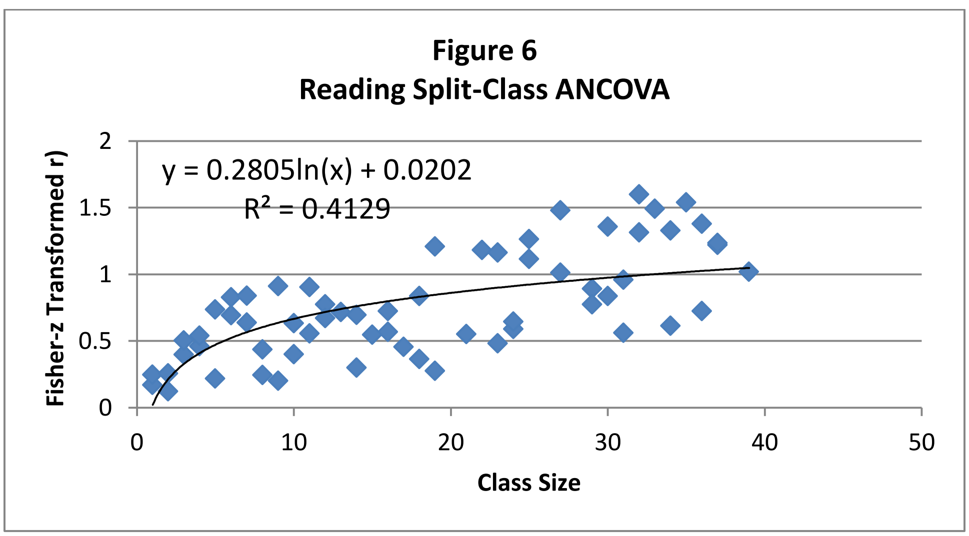


Figure 6
Reading Split-Class ANCOVA
y = 0.2805ln(x) + 0.0202
R² = 0.4129
Fisher-z Transformed r)
Class Size

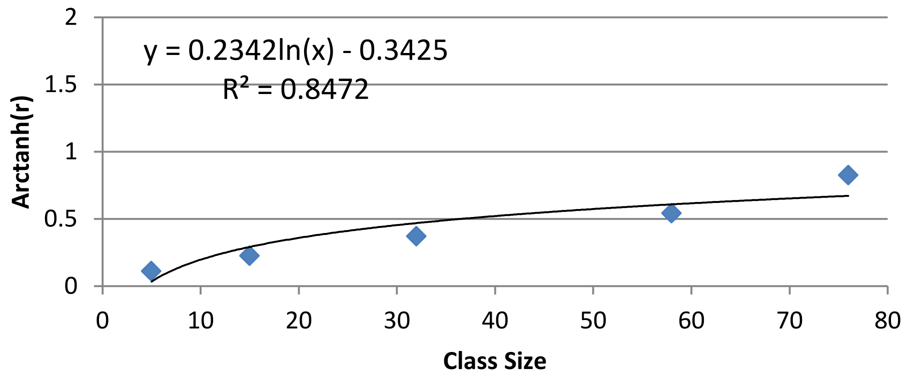


Figure 7
Math Cross-Years Simple Gain Scores

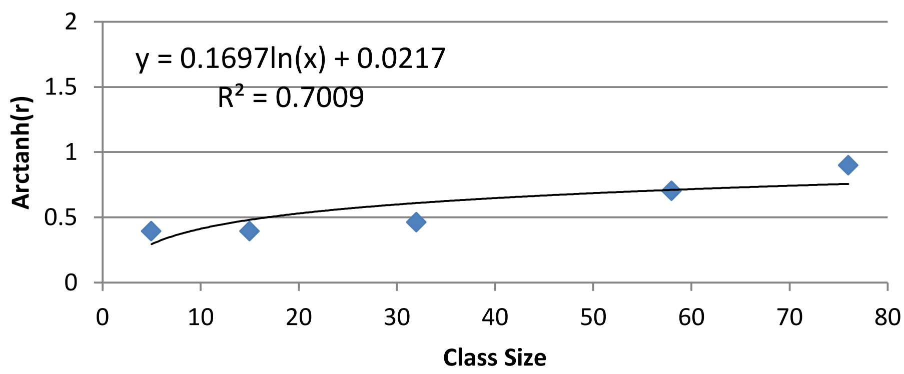


Figure 8
Math Cross-Years Student Growth Percentiles

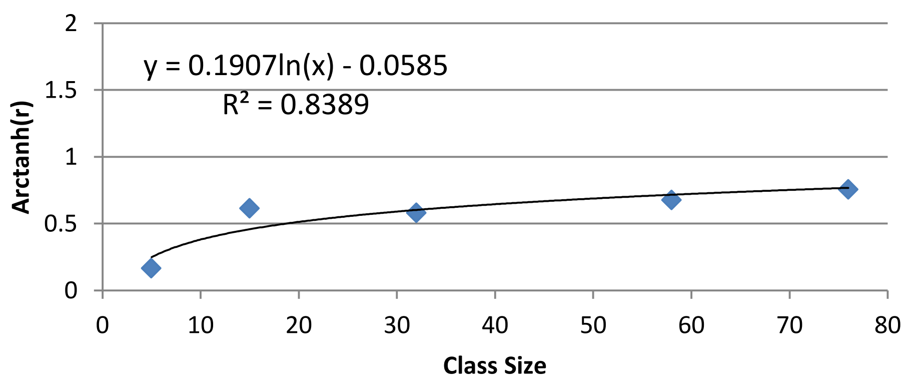


Figure 9
Math Cross-Years ANCOVA

Figure 10
Reading Cross-Years Simple Gain Scores

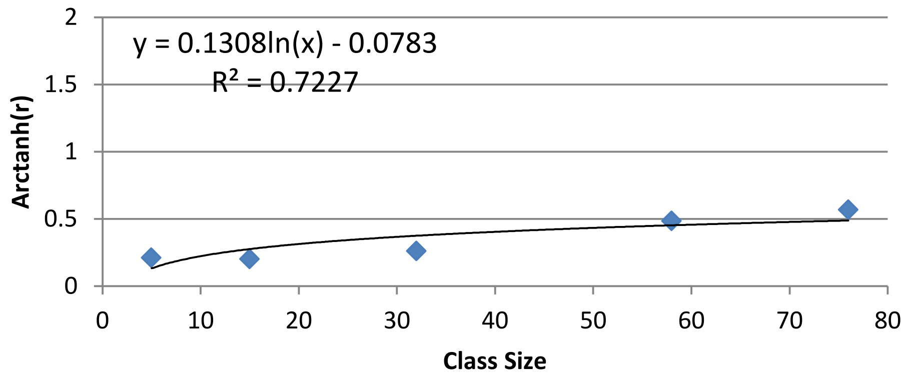


Figure 11
Reading Cross-Years Student Growth Percentiles

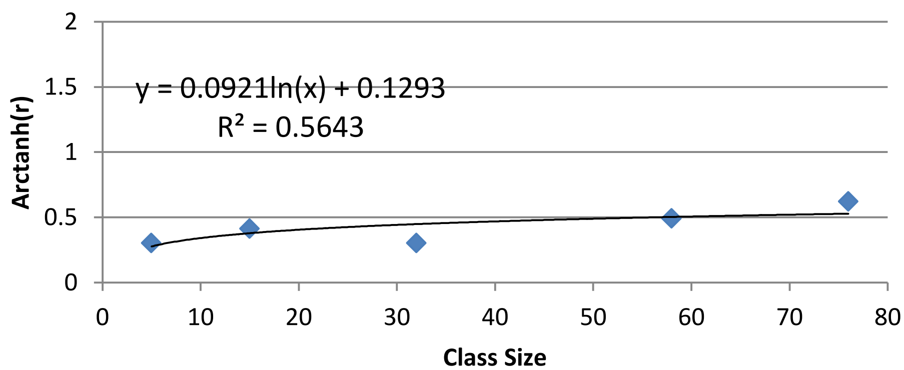


Figure 12
Reading Cross-Years ANCOVA