

The Generalizability of School Growth Scores Derived from Student Growth Percentiles for Use in School Accountability and Principal Evaluation Systems

**Paper presented to the Annual Meeting of the
National Council on Measurement in Education
San Francisco, CA**

April 28, 2013

Andrea Lash
WestEd

Mary Peterson
WestEd

Richard Vineyard
Nevada Department of Education

Vanessa Barrat
WestEd

Loan Tran
WestEd

The Generalizability of School Growth Scores Derived from Student Growth Percentiles for Use in School Accountability and Principal Evaluation Systems

Under direction of the Nevada State Legislature, the Nevada Department of Education plans to include measures of student growth in mathematics and reading achievement in the state's school accountability system and in the statewide system for educator evaluation. The methodology selected for measuring achievement growth is the Student Percentile Model developed by Damian Betebenner (2009). While this method has received widespread attention by states and districts across the nation, there have been few reports of research into the psychometric properties of the resulting growth scores. This paper summarizes a study of the stability of school growth scores over years. It was conducted at the request of the Nevada Department of Education with data from the state's elementary and middle schools. For mathematics and reading, it reports generalizability coefficients and standard errors of measurement of the growth scores and explores implications of these estimates for the identification of low- and high-performing schools.

The Nevada Policy Context

The 2009 Nevada State Legislature mandated the adoption of a statewide growth model of student achievement for school accountability purposes. Nevada Revised Statute Chapter 386.650 outlined the intent of adopting a growth model to track student progress from year to year to determine whether a school has made progress in the achievement of its pupils. The model was to be built from data available from the current statewide assessment system that administers criterion-referenced tests of reading and mathematics in grades 3 through 8.

The Nevada Department of Education (NDE) established five criteria for selecting a growth model. The model should be a valid and reliable metric that is communicable and technically sound; provide new information about growth that was currently masked by the Nevada accountability system; be supported by the state's assessment data system to be conditioned on past performance; be compatible with test scales that are not vertically aligned; and define legitimate criteria by which to judge how much growth is expected of schools (Davidson, 2010). The NDE selected a growth model based on student growth percentiles and the work of Damian Betebenner (2009, 2011) as best meeting these criteria. Of particular concern was the need to identify a means to measure growth that could be applied to tests that

were not vertically equated. At the time, value-added models for measuring effectiveness of schools were not recommended for such tests.

The 2011 Nevada State Legislature expanded the use of growth model data to educator evaluation. Nevada Revised Statute Chapters 391.3125 and 391.3127 required the creation of a comprehensive system for evaluating educator effectiveness. According to the new laws, at least 50% of an educator's evaluation must be based on student achievement data, of which one data element is student growth. Current plans are to use the school-level measure of growth in the evaluation of principals and in the evaluation of teachers who teach grades or subject areas that are not tested in the statewide testing program.

Nevada's policies and applications of growth model data in high-stakes decisions have moved much more rapidly than the research to support them. This situation may be true in other states as well. Across the country, the student growth percentile methodology is now in various stages of use—ranging from preliminary investigation to full-scale adoption—in approximately 20 states (New Jersey Department of Education, 2012). Yet, little information has been published or shared about properties of the scores from this model (cf. Goldschmidt, Choi, & Beaudoin, 2012).

Recognizing that the growth model they had selected was quite new and had not been studied extensively, NDE sought research on the student growth percentile scores. Earlier descriptive studies examined how students' growth scores varied by student demographic features (Tran and Crane, 2010) and how school-level growth scores correlated with other indicators of school performance (Crane, Tran, and Lash, 2010). The study reported here examined the stability of school-level scores over years.

The Nevada Growth Model

In the Nevada Growth Model, a student's score is a percentile score that identifies how well the student performed on a test relative to other students in the state at the same grade level with the same prior achievement level. Students' prior achievement is evaluated by scores on tests completed in (up to 5) previous years. A student's growth score, called the student growth percentile (SGP), is the student's percentile rank in the conditional distribution of the current year's achievement of same-grade students in the state, conditioned on prior achievement. An SGP of 40 in 2012 for a sixth grader indicates that the student had an achievement test score in

2012 that was equal to or higher than the test scores of 40 percent of sixth-graders in the state whose prior-years' test scores were similar to those of the student. SGPs are computed for students in grades 4 through 8¹. In each grade, the median across students in the state is an SGP of 50.

The school-level growth measure is the median of the growth scores for the students in the school. In schools with a growth score below 50, the typical student in the school has scored lower than the typical student in the state who started out the year with similar achievement levels.

As a descriptive measure, the school growth score may help to summarize the performance of a particular group of students in a particular year. If it is to be used as a basis of decision making in a school accountability system or in a principal evaluation system, then it should reflect an enduring trait of a school and principal that can be used to infer their effectiveness. One basic feature then is that the score should remain stable from one assessment to the next. Nevada state and district administrators have asked how reliable the growth measure is over time. As a reference point, how does the stability of this new measure compare with the stability of the measure of achievement status that has been used in the previous school accountability system: the percentage of students in a school who are judged to be proficient on the statewide test? How does stability change when growth is estimated by averaging annual growth scores from multiple years? This study addressed these questions in an attempt to help Nevada officials make informed decisions about the use of the growth scores to evaluate schools and principals. Given the number of states and districts around the country that are considering use of this same methodology in accountability and educator evaluation systems, the research should have value outside of Nevada as well. Therefore, NDE senior administrators have provided permission to disseminate the results of this study.

Analysis Methods & Data Examined

We conducted a one-facet generalizability study (Brennan, 2001; Shavelson & Webb, 1991) to address the research questions. The single facet in the design is years; it is crossed with schools. The appendix provides a brief description of concepts from Generalizability theory (G-

¹ Grade 4 is the earliest grade in which growth scores are computed because their computation requires at least one previous year of test scores and statewide testing in Nevada begins in grade 3. Grade 8 is the last grade because it is the last grade for annual statewide testing in Nevada.

theory) in the context of this design. For four school years beginning in 2007/08, NDE provided SGPs in reading and in mathematics, linked to school identifiers, for all Nevada students in grades 4 through 8 (who had the necessary test scores). From the student-level data sets, we computed the median SGP score for each school, and, using the software GENOVA (Brennan, 2001), estimated variance components in each subject for a one-facet, crossed design. NDE also provided the test scores in mathematics and reading for the pupils with SGP scores which allowed us to also complete the g-study analysis on the status measure of school achievement.

The dataset provided by NDE included data for students from 494 schools in 2008, 512 in 2009, 518 in 2010 and 521 in 2011. Following NDE policy, we excluded from computation of school scores students who had not been enrolled at the school the full school year, and we included schools in the analysis only if they had, for all four study years, at least 25 students with SGP scores. Of the schools in the data set, 472 had four years of data, and 426 of those had 25 or more students with SGP scores each of the four years.

An alternative design that was considered was a two-facet design with students as a random facet nested within the school-by-year interaction. Students would be nested in the interaction because a student would be assigned to only one school each year. (Following NDE policy, students who transfer between schools are not included in the school assessment.) For this study, student-level data would need to be input to the analysis. The G-study analysis of this design, however, would incorrectly estimate school-level scores as the mean of the SGP's for students in the school. In the one-facet design, school-level scores are entered directly and thus Nevada's school-level score – the median SGP – could be studied. The trade-off between these two designs is that with the single-facet design it is not possible to consider score variability associated with the student sample and thus it is not possible to examine how the number of students included in a school's growth estimate might affect the stability of the school score. To provide some information about the effect of school enrollment on stability, the population of schools in the study was divided into thirds on the basis of the number of students with SGP scores who were enrolled in the school and the single-facet analysis was conducted within each of the three sub-populations as well as for the total school population.

Findings

The G-study results for mathematics for the school-level growth score and the school-level status score are presented in Tables 1 and 2, respectively. Results for reading for the school-level growth and status scores are presented in Tables 3 and 4, respectively. Each table summarizes analyses for the full school sample and for the three sub-samples that differ in student enrollment. The tables provide the estimates of variance components derived from the G-study analysis along with two indicators of the stability of scores: generalizability coefficients and standard errors of measurement. We discuss each of these in turn.

Variance Components

The G-study provides estimates of the three components of variance that are identified in the single-facet design of this study. The first component, for schools, is associated with differences among schools and is comparable to the variability of true scores in Classical Test Theory (CTT). Another component, labeled residual in the tables, contains the variability associated with the interaction of schools and years along with all other sources of random errors that have not been accounted for in the design. The variability associated with the interaction are due to random sources of error that cause school scores to change over years in ways that are not systematic across schools. The residual component is comparable to the error component of CTT.

A third component, for years, is associated with variability in the average scores for years. For example, if the scoring process was changed one year and the change introduced an error that added 5 points to each school's score, that error would contribute to the component of variance associated with years. In CTT, with the assumption of parallel test forms, there is no comparable variance component. This component is expected to be close to zero for scores derived from the Nevada Growth Model because the score scale is centered each year so that the median score for students statewide, in each grade is 50. As Table 1 and 3 show, this component is zero for growth scores for the total school sample and close to zero for each of the sub-samples. That is not the case, however, for the status score (percentage of students who were proficient) as Table 2 and 4 show. Some changes over time in the testing situation led to differences in annual average scores on the status measure. Modifications to the tests or to curricula, for example, could produce score changes over years.

Generalizability Coefficient

The generalizability coefficient, like the reliability coefficient, is the proportion of variance in scores that may be attributed to differences among schools. It is a ratio of the school component of variance to a total that is the sum of the school and error components of variance. For the Nevada assessment systems, the error component includes the year and residual components.² Tables 1 through 4 provide the coefficient for scores computed in a single year and for aggregates of annual scores. That is, for estimates of school growth and status that are averages of 2, 3, or 4 annual scores. Thus, if the stability of a single-year score is not sufficient, policymakers might wait and evaluate a school's growth, for example, by averaging the growth scores derived from two separate years.

Generalizability coefficient presented in Tables 1 and 2 for mathematics scores are displayed in Figure 1. These are the coefficients for the full sample of schools. The coefficients for the growth score are lower than those for the status score. For scores that summarize growth and status in a single year, the generalizability coefficients were estimated to be .43 and .77, respectively. The estimate for the growth score is very close to the test-retest reliability coefficient of .46 that Goldschmidt, Choi, & Beaudoin (2012, p 47) reported in an examination of school scores derived from the SGP model. Goldschmidt et al computed test-retest reliability coefficients separately for elementary and middle-school and found both estimates to be .46. Figure 1 displays how the generalizability coefficient from the Nevada data changes as more information is added to the estimate of a school's performance. As the number of annual scores in the estimate increases, so does the coefficient. Four years of annual growth scores would be needed to achieve stability near the level of the stability of the single-year estimate for status.

² Within G-theory it is possible to define the errors important to a test situation and combine their components in different ways for different situations. In test situations where the goal is to assess schools relative to one another, for example, to identify the top ten percent of the schools, then systematic errors due to the year of testing that affect each school in the same way would not change the ordering of schools or the identification of the top ten percent. In that situation, the component of variance attributed to year would not be a source of error that one would need to consider in computing the generalizability coefficient or the standard error of measurement (SEM). However, if the school scores were to be used instead to make absolute decisions as to whether a school was above or below a certain cut score, for example, then the systematic errors due to years could affect the decisions about schools and should be included in the computation of coefficient and SEM. Since Nevada's accountability and educator evaluation systems will assess scores against an absolute criterion, we include the component for years in our estimates of error variance.

Generalizability coefficients presented in Tables 3 and 4 for reading scores are displayed in Figure 2 for the full school sample. As it was for mathematics, for reading the coefficients for growth are smaller than the coefficients for status. For scores that summarize achievement for a single year, the coefficient for growth was .38 and for status it was .80. Again, the estimate for growth is similar to the test-retest coefficients Goldschmidt et.al. report for reading. Their estimates were .32 for elementary schools and .33 for middle schools (Goldschmidt, Choi, and Beaudoin , 2012, p 47). As Figure 2 shows, the generalizability coefficients increase when more information is included in the estimate of school achievement for both growth and status. If four years of annual growth scores are averaged to estimate school growth, then the generalizability coefficient would be expected to be .70.

School size matters to the stability of growth measures. This can be seen in Tables 1 and 3 in the results for the three sub-samples of schools. Higher coefficients were found for samples composed of larger schools.

The Standard Error of Measurement & Misclassification Rates

The patterns in the findings for the standard error of measurement (SEM) track the findings presented for the generalizability coefficient, as expected. The SEM improves (decreases) with increases in the number of annual scores that are averaged to estimate a school score. And the SEM improves with increases in the size of the school (Tables 1 through 4).

With the SEM it is possible to begin to examine how stability of school growth scores may affect decisions when the scores are used to classify schools or educators into performance groups, as they will be used in the Nevada accountability system and educator evaluation system. As an example, Figure 3 shows a theoretical distribution of observed growth scores in mathematics, based on a single-year estimate, for schools having a “true” score of 50³. Observed scores differ from 50 because of measurement errors. Assuming a normal distribution of errors with standard deviation equal to the SEM, it is possible to compute the percentage of observed scores that would fall below the cut score that identifies “low performing” schools and the percentage that would fall above the score that identifies “high performing” schools. For this

³ An SGP of 50 is the median growth score for the student population in the state. It also is close to the annual mean for schools which was between 50.3 and 50.8 for the four years in this study.

example, we use cut scores of 40 and 60 which NDE has considered.⁴ The figure shows that for schools with typical growth of 50, 10% of the schools would be expected to be judged to be low performing, having observed scores lower than 40 due to measurement errors. Another 10% would be judged to be high performing because their observed scores would be greater than 60 due to measurement errors. The rate of misclassification is reduced by adding more years of data, as shown in Figure 4.

Figures 3 &4 have been helpful in communicating with state-level administrators and policymakers about the effect of measurement errors on the accuracy of school classifications. But the conditional error rates shown in these figures are for one score level only. Different rates of misclassification would be expected for different true scores. An estimate of the overall or unconditional error rate would be more helpful for decision makers.

Estimating unconditional error rates requires information about the distribution of true scores. For this example, we estimated the true score for each school using a regression equation, known as Kelley’s formula (Hubert and Wainer, 2013), that takes into account the reliability of the measure. The formula is

$$\hat{T}_s = \bar{X} + r_{xx'}(X_s - \bar{X});$$

Where \hat{T}_s is the estimated true score for school s , \bar{X} is the mean of school scores, $r_{xx'}$ is the estimated reliability of the score, estimated in this example by the generalizability coefficient, and X_s is the observed score for school s .⁵

We looked at two types of classification errors: identifying schools with true scores above the cut-score of 40 as low performing schools (false positives), and failing to identify as low performing schools below the cut score (false negatives). Table 5 provides the expected misclassification rates for mathematics and for reading when judgments are based on the score

⁴ Cut scores have not been identified for the educator effectiveness system. When this research was underway, they also had not been identified for the school accountability framework. Now the new accountability framework categorizes schools into 5 groups, rather than 3, and has cut scores that differ by subject and grade range. However, the scores of 40 and 60 are still informative examples. They are within a few points of the scores NDE selected to distinguish the two lowest and two highest groups of schools from middle group of schools in the accountability system. .

⁵ For this example, we took for the observed score the average of the four estimates of a school’s growth rather than selecting one of the four estimates. The reliability estimate used in this example was .75, the generalizability coefficient estimated from the G-study for the average of 4 annual scores.

from a single year's growth estimate and also when they are based on averages of 2, 3, or 4 annual estimates. The table also shows the error rates for identification of high performing schools.

For mathematics, the proportion of schools in the population that would be misclassified when identifying low performing schools was estimated to be .14 when the judgment is based on a single year's estimate of school growth. When the judgment is based on averages of two or more annual estimates of growth, the proportion of schools that would be misclassified would be expected to drop to .09, .06, and .05 for averages based on 2,3, and 4 years, respectively.

The proportion of schools that are not low performing that would be expected to be identified incorrectly as low performing (false positives) is .13 when a single year's estimate of growth is used. The proportion of low performing schools that are not identified (false negative) is expected to be .37. While the error rate for false negatives is higher than the error rate for false positives, its effect on the overall error rate across all schools is small because only 4% of schools had estimated true scores below 40, the cut score for low performing schools. The error rates for both types of errors decline when the judgments are based on the average of two or more annual growth scores.

For reading, the error rates overall, and for false positives, are slightly smaller than they were for mathematics, reflecting the smaller SEM in reading as compared with mathematics. While the error rates for false negatives were large, as they were for mathematics, they had little effect on the overall error rates because only 1.6% of the schools had estimated true scores below 40 and 1.2% had estimated true scores above 60. Most errors of classification, then, would be false positives.

Dissemination of the Study and Next Steps.

The need to understand the stability of the values produced by the Nevada Growth Model is critical to the state's ability to use the scores fairly in the evaluation of the performance of schools and soon in the evaluation of the effectiveness of educators. At their request, early findings⁶ were presented to administrators from NDE, Clark County School District, which serves 75% of the students in the state, and Washoe County School District, which serves 15% of

⁶ All findings have been presented except for the estimates of unconditional misclassification rates. These are new analyses completed for this paper which will be presented to NDE and Nevada policymakers in the future.

the students in the state. We also presented early findings to the Nevada State Legislative Committee on Education.

It is difficult to track whether or not the research has been considered in decisions regarding the design of the state accountability and educator evaluation systems. We do know that the question of score stability is a topic of conversation in planning meetings. State administrators have reported that the research is a primary reason that the new accountability framework will consider multiple years of data (C. Crothers, personal communication, December, 2011) and the state's application for flexibility under the provisions of the Elementary and Secondary Education Act notes that research into the stability of scores will be used to develop the final decision process. Current plans for the statewide educator evaluation system require at least three years of scores for educators prior to any actions being taken relative to probationary status, promotion, or dismissal.

Thus, there are reasons to believe that the findings of the generalizability study have made their way into the debates about the best way to use test information in high-stakes decisions for Nevada schools and educators. Yet, this one study addresses only a small question about the use of the growth scores in accountability and educator evaluation systems. We examined errors of measurement associated with the median growth percentile for schools, but these systems are creating other measures from the growth score which may differ from the median growth percentile in terms of stability. Examining the properties of the complex decision rules that are being developed to sort and classify schools and educators will be important if we are to understand the generalizability of the outcomes. And more important yet are questions of validity that have not been examined for the SGP growth model as it is used in these high-stakes situations.

An example of the complexity of score generation can be seen in the new system for school accountability in Nevada. In the summer of 2012, Nevada received a conditional approval of its request for additional flexibility under the provisions of the Elementary and Secondary Education Act. As part of the request, the state proposed the development of a new system for the evaluation of schools: the Nevada School Performance Framework (NSPF), which is heavily dependent on measures of student growth in the calculation of the individual school performance index scores. The NSPF uses individual SGP values to produce a median growth percentile (MGP) for a school, and also to evaluate individual student growth toward

proficiency. The growth to a target or Adequate Growth Percentile (AGP) evaluates a student's growth trajectory to achieve, or maintain, a status measure of proficient or "Meets Standards" within three years or by the eighth grade. The NSPF also rewards schools for being able to reduce the gaps in AGP performance between specific identified subpopulation groups (IEP, ELL, and FRL) and the highest performing groups. For elementary and middle schools in the state, the combination of growth measures (MGP, AGP, and Gap reduction) accounts for 60% of a school's index score. Growth measures are less prominent for high schools under the NSPF.

In addition to using growth measures for school accountability, the new state educator evaluation system, the Nevada Educator Performance Framework, uses SGP and gap reduction measures as major components in the Student Outcomes domain, which makes up 50% of the score that will be used to judge educators. Within the Student Outcomes domain, the school median growth percentile in combination with a school wide measure of gap reduction (based on AGP) accounts for 90% of the score – median growth percentile is 70% and gap reduction is 20%. Only 10% of the Student Outcomes score is based on measures of proficiency/status.

Our generalizability study of the median growth percentile is only a first step in examining the generalizability of the scores that policymakers will use in high-stakes decisions about schools and principals. The composite measures for school accountability and educator effectiveness that have been proposed so far in Nevada include two or more components based on the school growth measure. Each component requires a transformation of or further computation with the median growth percentile that can introduce new errors or guard against some errors. It is not clear how the estimates of the stability of the median growth percentile and the expected error rates of decisions based on it might apply to the newly developed composite measures.

Planning for the new accountability and educator effectiveness systems has had to proceed rapidly in Nevada in order to meet deadlines imposed by state laws and federal requirements. So rapidly, that research has not kept up or been able to inform decisions about the creation of categorical scores and composite scores. There is no reason to think the situation is different in other states, many of which have also selected the Student Growth Percentile model to evaluate their schools and educators. NDE's willingness to share research conducted in Nevada schools is encouraging. If other states do the same, it may be possible to build more

quickly some understandings of the properties of the growth scores that have been proposed for use in high-stakes accountability and educator evaluation systems.

References

- Betebenner, D. W. (2009). Norm- and criterion-referenced student growth. *Educational Measurement: Issues and Practice*, 28(4) 42-51.
- Betebenner, D. W. (2011). *A technical overview of the student growth percentile methodology: Student growth percentiles and percentile growth projection/trajectories*. Dover, NH: National Center for the Improvement of Educational Assessment.
- Brennan, R. L. (2001). *Generalizability theory*. New York: Springer.
- Crane, E., Tran, L. & Lash, A. (2010). *Assessing School Performance with Nevada's Growth Model*. Technical Assistance Memo. San Francisco, CA: REL West at WestEd.
- Davidson, A (2010). *Update on development of the Nevada Growth Model of Achievement* (2010). Presentation to the NPEP Technical Advisory Committee in Las Vegas, Nevada, by the Nevada Department of Education.
- Davidson, A., Ozdemir, S., & Harris, J. (2010). *An approach to criterion-referenced growth modeling: One state's application of student growth percentiles*. Paper presented at the annual meeting of the American Educational Research Association, Denver, CO.
- Goldschmidt, P. Choi, K., & Beaudoin, J.P. (2012). *Growth model comparison study: Practical implications of alternative models for evaluating school performance*. A paper commissioned by the Technical Issues in Large-Scale Assessment State collaborative on Assessment and Student Standards, Council of Chief State School Officers. Washington, DC: Council of Chief State School Officers.
- New Jersey Department of Education. (2012). *An illustration of SGP adoption in the United States*. Trenton, NJ. Retrieved online May 2012 from http://www.state.nj.us/education/njsmart/performance/SGP_Adoption.pdf
- Shavelson, R. J., & Webb, N. M. (1991). *Generalizability theory: A primer*. Newbury Park, CA: Sage Publications.
- Tran L. and Crane, E. (2010). *Subgroup Comparisons of Nevada's Student Growth Percentile*. Technical Assistance Memo. San Francisco, Ca: REL West at WestEd.
- Hubert, L., & Wainer, H. (2013). *A statistical guide for the ethically perplexed*. Boca Raton, FL: Taylor & Francis Group/CRC Press

Table 1. School Growth (Median SGP) in Mathematics: Variance Components, Standard Errors of Measurement, and Generalizability Coefficients by School Size and for All Schools Pooled

Estimates Derived from the Generalizability Study	Schools Grouped by Size			All Schools pooled (N=426)
	Less than 261 students (N=142)	261 to 363 Students (N=142)	More than 363 students (N=142)	
Variance Component for:				
School	59.45	39.37	46.05	49.40
Year	0	0.11	0.56	0
Residual	90.84	59.99	41.43	64.27
Total	150.29	99.47	88.04	113.67
Standard Error of Measurement for:				
The score from 1 year	9.53	7.75	6.48	8.02
The average of scores from 2 years	6.74	5.48	4.58	5.67
The average of scores from 3 years	5.50	4.48	3.74	4.63
The average of scores from 4 years	4.77	3.88	3.24	4.01
Generalizability Coefficient for:				
The score from 1 year	0.40	0.40	0.52	0.43
The average of scores from 2 years	0.57	0.57	0.69	0.61
The average of scores from 3 years	0.66	0.66	0.77	0.70
The average of scores from 4 years	0.72	0.72	0.81	0.75

This table reads: For the analysis of all schools pooled, the component of variance in the scores attributed to differences between schools was 49.4, the component attributed to error and other unexplained sources was 64.27, and the total was 113.67. Using these values, the standard error of measurement for the score from a single year was 8.02 (the square root of the residual component), and the generalizability coefficient for a single year's score was 0.43 (the proportion of total variance that is attributed to schools). Estimates of the standard error for n years are derived by taking the square root of [(year component /n)+ (residual variance component/n)]; the generalizability coefficient for n years is: school variance component/[(school variance)+(year variance/n)+(residual variance component/n)].

Note: These analyses are based on scores from 426 schools that had SGP values estimated for at least 25 students each of four years: 2008-2011

Source: Authors' analysis of data provided by the Nevada Department of Education

Table 2. School Proficiency (Percent of Students who are Proficient) in Mathematics: Variance Components, Standard Errors of Measurement, and Generalizability Coefficients by School Size and for All Schools Pooled

Estimates Derived from the Generalizability Study	Schools Grouped by Size			All Schools pooled (N=426)
	Less than 261 students (N=142)	261 to 363 Students (N=142)	More than 363 students (N=142)	
Variance Component for:				
School	94.09	100.13	152.12	119.54
Year	9.31	8.93	10.62	9.21
Residual	43.87	21.37	13.28	26.58
Total	147.27	130.42	176.02	155.33
Standard Error of Measurement for:				
The score from 1 year	7.29	5.5	4.89	5.98
The average of scores from 2 years	5.16	3.89	3.46	4.23
The average of scores from 3 years	4.21	3.18	2.82	3.45
The average of scores from 4 years	3.65	2.75	2.44	2.99
Generalizability Coefficient for:				
The score from 1 year	0.64	0.77	0.86	0.77
The average of scores from 2 years	0.78	0.87	0.93	0.87
The average of scores from 3 years	0.84	0.91	0.95	0.91
The average of scores from 4 years	0.88	0.93	0.96	0.93

This table reads: For the analysis of all schools pooled, the component of variance in the scores attributed to differences between schools was 119.54, the component attributed to differences in years was 9.21, the component attributed to error and other unexplained sources was 26.58, and the total was 155.33. Using these values, the standard error of measurement for the score from a single year was 5.98 (the square root of the sum of the component for year and the component for residual), and the generalizability coefficient for a single year's score was 0.77 (the proportion of total variance that is attributed to schools). Estimates of the standard error for n years are derived by taking the square root of $[(\text{year component} / n) + (\text{residual variance component} / n)]$; the generalizability coefficient for n years is: $\text{school variance component} / [(\text{school variance}) + (\text{year component} / n) + (\text{residual component} / n)]$.

Note: These analyses are based on scores from 426 schools that had SGP values estimated for at least 25 students each of four years: 2008-2011

Source: Authors' analysis of data provided by the Nevada Department of Education

Table 3. School Growth (Median SGP) in Reading: Variance Components, Standard Errors of Measurement, and Generalizability Coefficients by School Size and for All Schools Pooled

Estimates Derived from the Generalizability Study	Schools Grouped by Size			All Schools pooled (N=426)
	Less than 261 students (N=142)	261 to 363 Students (N=142)	More than 363 students (N=142)	
Variance Component for:				
School	35.36	27.08	27.14	30.24
Year	0	0	0.20	0
Residual	71.85	39.98	34.87	48.84
Total	107.21	67.06	62.21	79.08
Standard Error of Measurement for:				
The score from 1 year	8.48	6.32	5.92	6.99
The average of scores from 2 years	5.99	4.47	4.19	4.94
The average of scores from 3 years	4.89	3.65	3.42	4.03
The average of scores from 4 years	4.24	3.16	2.96	3.49
Generalizability Coefficient for:				
The score from 1 year	0.33	0.40	0.44	0.38
The average of scores from 2 years	0.50	0.57	0.61	0.55
The average of scores from 3 years	0.60	0.67	0.70	0.65
The average of scores from 4 years	0.66	0.73	0.76	0.71

This table reads: For the analysis of all schools pooled, the component of variance in the scores attributed to differences between schools was 30.24, the component attributed to error and other unexplained sources was 48.84, and the total was 79.08. Using these values, the standard error of measurement for the score from a single year was 6.99 (the square root of the residual component), and the generalizability coefficient for a single year's score was 0.38 (the proportion of total variance that is attributed to schools). Estimates of the standard error for n years are derived by taking the square root of $[(\text{year component} / n) + (\text{residual variance component} / n)]$; the generalizability coefficient for n years is:

$\text{school variance component} / [(\text{school variance}) + (\text{year variance} / n) + (\text{residual variance component} / n)]$.

Note: These analyses are based on scores from 426 schools that had SGP values estimated for at least 25 students each of four years: 2008-2011

Source: Authors' analysis of data provided by the Nevada Department of Education

Table 4. School Proficiency (Percent of Students who are Proficient) in Reading: Variance Components, Standard Errors of Measurement, and Generalizability Coefficients by School Size and for All Schools Pooled

Estimates Derived from the Generalizability Study	Schools Grouped by Size			All Schools pooled (N=426)
	Less than 261 students (N=142)	261 to 363 Students (N=142)	More than 363 students (N=142)	
Variance Component for:				
School	126.15	164.76	148.84	153.78
Year	4.90	5.95	17.14	6.35
Residual	40.48	16.52	31.27	32.40
Total	171.53	187.23	197.25	192.53
Standard Error of Measurement for:				
The score from 1 year	6.74	4.74	6.96	6.22
The average of scores from 2 years	4.76	3.35	4.92	4.40
The average of scores from 3 years	3.89	2.74	4.02	3.59
The average of scores from 4 years	3.37	2.37	3.48	3.11
Generalizability Coefficient for:				
The score from 1 year	0.73	0.88	0.75	0.80
The average of scores from 2 years	0.85	0.94	0.86	0.89
The average of scores from 3 years	0.89	0.96	0.90	0.92
The average of scores from 4 years	0.92	0.97	0.92	0.94

This table reads: For the analysis of all schools pooled, the component of variance in the scores attributed to differences between schools was 153.78, the component attributed to differences in years was 6.35, the component attributed to error and other unexplained sources was 32.40, and the total was 192.53. Using these values, the standard error of measurement for the score from a single year was 6.22 (the square root of the sum of the component for year and the component for residual), and the generalizability coefficient for a single year's score was 0.80 (the proportion of total variance that is attributed to schools). Estimates of the standard error for n years are derived by taking the square root of $[(\text{year component}/n) + (\text{residual variance component}/n)]$; the generalizability coefficient for n years is: $\text{school variance component}/[(\text{school variance})+(\text{year component}/n) + (\text{residual component}/n)]$.

Note: These analyses are based on scores from 426 schools that had SGP values estimated for at least 25 students each of four years: 2008-2011

Source: Authors' analysis of data provided by the Nevada Department of Education

Table 5. Estimated Error Rates in Classifying Schools into Low Performance and High Performance Categories on the Basis of Growth Scores in Mathematics and in Reading

Classification Errors	Number of Annual Growth Scores Averaged to Obtain Final Growth Estimate			
	1	2	3	4
Mathematics				
Proportion of all schools misclassified when judging if schools are low performing (growth < 40)	.14	.09	.06	.05
1. Proportion of schools with growth > 40 that are incorrectly identified as low performing	.13	.07	.05	.04
2. Proportion of low performing schools that are not identified	.37	.33	.31	.29
Proportion of all schools misclassified when judging if schools are high performing (growth > 60):	.16	.11	.08	.07
1. Proportion of schools with growth < 60 that are incorrectly identified as high performing	.14	.09	.07	.06
2. Proportion of high performing schools that are not identified.	.37	.33	.31	.29
Reading				
Proportion of all schools misclassified when judging if schools are low performing (growth < 40)	.11	.06	.05	.04
1. Proportion of schools with growth > 40 that are incorrectly identified as low performing	.10	.06	.04	.03
2. Proportion of low performing schools that are not identified	.38	.34	.31	.28
Proportion of all schools misclassified when judging if schools are high performing (growth > 60):	.12	.07	.04	.03
1. Proportion of schools with growth < 60 that are incorrectly identified as high performing	.11	.06	.04	.03
2. Proportion of high performing schools that are not identified.	.38	.33	.30	.27

Note. For mathematics, the percentage of schools estimated to have true scores below the cut score of 40 is 4 percent; the percentage estimated to have true scores above the cut score of 60 is 6 percent. In Reading, the estimated percentages are 1.64 percent below 40 and 1.2 percent above 60. The analysis is based on scores from 426 schools.

This table reads: “The proportion of schools that are expected to be misclassified when identifying low performing schools is .14 when a single year’s growth score is used to classify schools. When the judgment is based on an average of 2 annual growth scores, the proportion is .09. When it is based on an average of 3 annual scores the proportion of misclassifications is .06 and when based on an average of 4 annual scores the proportion is .04.”

Source: Authors’ analysis of data provided by the Nevada Department of Education.

Figure 1. Generalizability Coefficients for School Status and Growth Scores in Mathematics

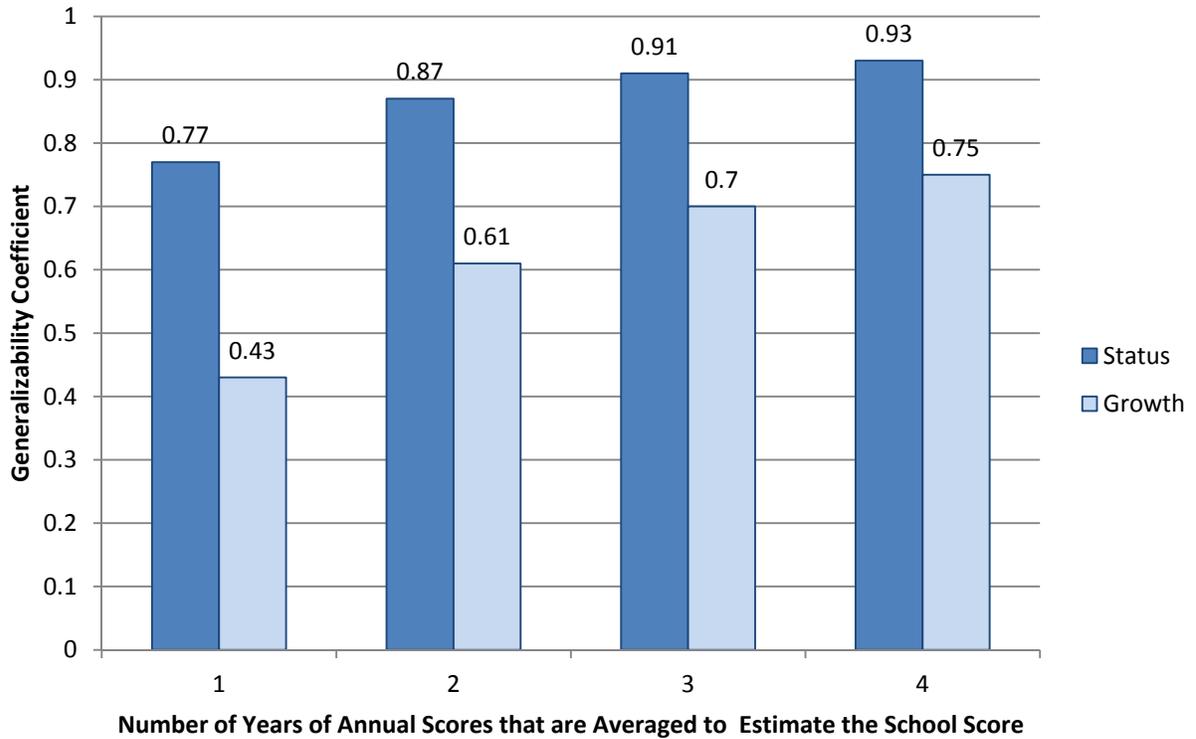


Figure 2. Generalizability Coefficients for School Status and Growth Scores in Reading

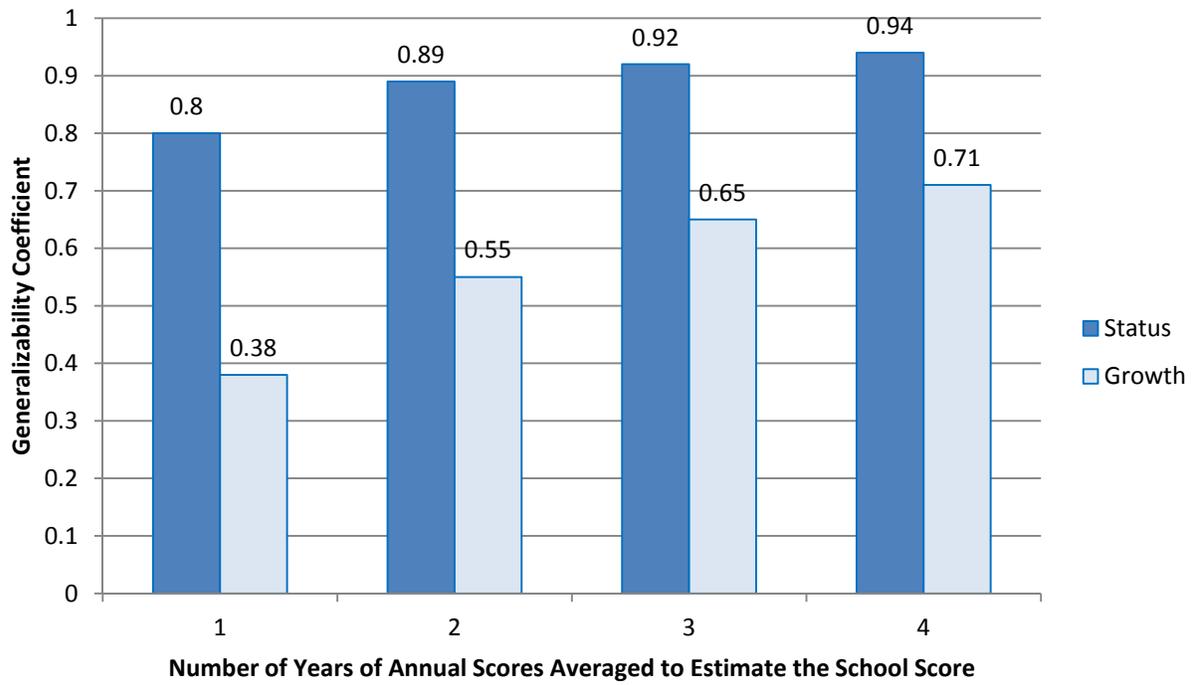


Figure 3. Classification Rates into Low Performing (<40), Typical, and High Performing (>60) School Categories Based on Growth Estimated from a Single Year for a School with a True Growth of 50.

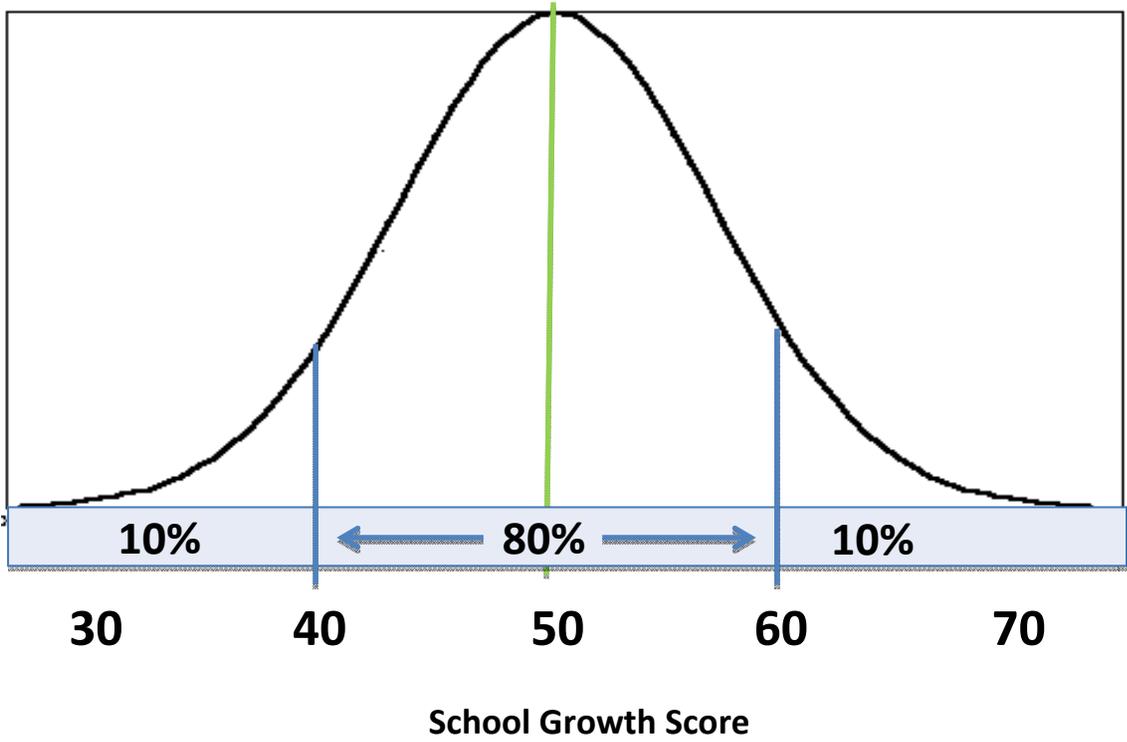
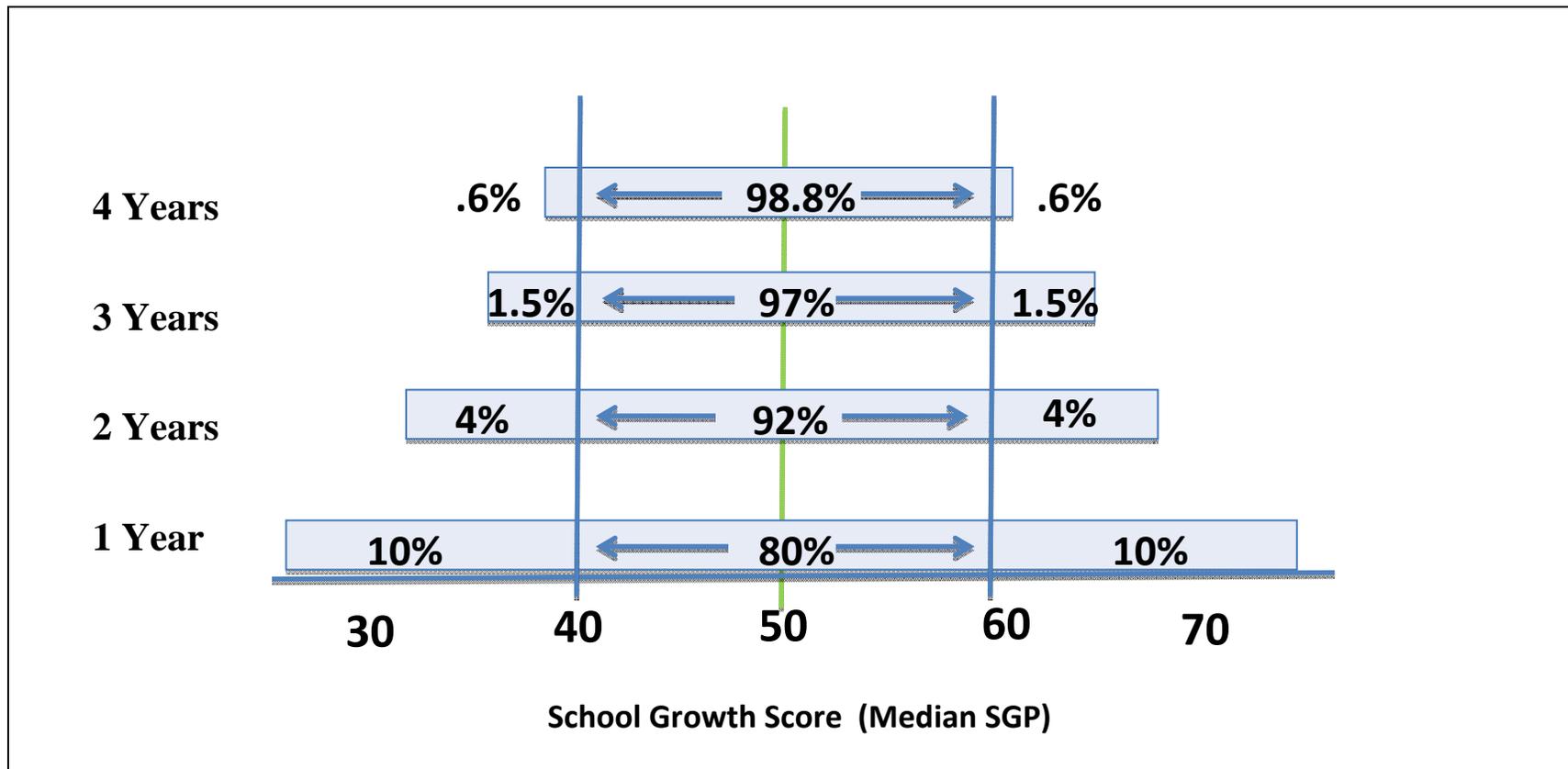


Figure 4. Classification Rates into School Performance Groups for Schools with True Growth Scores of 50 by the Number of Years of Annual Growth Scores Averaged to Estimate School Growth



Note. Low Performing Schools score <40 and high performing schools score > 60. All other schools would be classified as typical.

Appendix: A Brief Introduction to Generalizability Theory

When we interpret a measurement (e.g., a test score) taken on a particular day using a particular measurement method, it is rare that we limit our interpretation of the measurement to that specific day and measurement method. Instead, from that single measurement we infer conclusions about the individual to answer broader questions, such as: Has this student developed the mathematics knowledge needed for college-level courses? Is this teacher a capable teacher of mathematics? Is this principal an effective instructional leader? These are not questions about performance on a particular day or about performance assessed by a particular method of measurement. They are questions about an individual's enduring traits.

When we interpret scores we generalize from the particular score observed to a broader universe of possible scores that we *could* have observed—for example, those that could be achieved on different days using different measurement methods. Generalizability theory provides a conceptual framework that is useful in accounting for key features, called facets, of the universe of admissible observations, what Webb and Shavelson (2005) describe as “all possible observations that decision makers consider to be acceptable substitutes for the observation in hand.” Generalizability theory also provides a statistical method to evaluate the precision of our generalization and to examine how changing the way we sample measurements will alter the precision of our inferences.

As Brennan (2001) describes it, generalizability theory applies the framework of factorial designs from Analysis of Variance (ANOVA) to measurement problems. Just as the introduction of factorial designs to experimentation offered researchers the opportunity to study the individual and combined effects of multiple independent variables, the introduction of generalizability theory to the field of measurement provided researchers an opportunity to study the individual and combined effects of multiple sources of measurement error.

The Study Design

For the generalizability study (G-study) of the school effectiveness measures, we consider a simple design that includes only one facet, or source of error: the year in which a school is measured. The design is shown in table 2. Rows represent schools, columns represent years, and each cell has one observation, X_{sy} , which is the effectiveness measure observed for school s in year y . This is referred to as a crossed design because each school is observed each year.⁷ In the analysis of the Nevada Growth Model (NGM), X_{sy} would be the median SGP for students of school s in year y . In the analysis of the school status measure, X_{sy} would be the proportion of school s 's students who were proficient in the subject tested in year y .

⁷ In ANOVA, this would be a two-factor design with schools considered a factor of the design. In G-studies the people or objects measured are not considered a feature, or facet, of the design.

Applying the linear model as we do in ANOVA, X_{sy} may be represented as the sum of the expected values of effects associated with rows (schools), columns (years), their interaction (schools by years), and random errors. For the single-facet crossed design, the model is expressed for the G-study as:

$$X_{sy} = \mu + (\mu_s - \mu) + (\mu_y - \mu) + (X_{sy} - \mu_s - \mu_y + \mu); \quad [1]$$

In this equation, μ is the grand mean, the expectation of X_{sy} taken over all members of the school population and all years in the universe of observations; $(\mu_s - \mu)$ is the effect of school s , the deviation from the grand mean of the expected value of the school's score taken over all years in the universe of observation; $(\mu_y - \mu)$ is the effect of year y , the deviation from the grand mean of the expected value of the scores for year y taken over all schools in the population; and $(X_{sy} - \mu_s - \mu_y + \mu)$, is a residual effect or the portion of the score X_{sy} that is not explained by the other effects. The residual effect includes the effect of the interaction between schools and years as well as all other random sources of unexplained measurement errors.

Table 2. A Single-Facet Crossed Design

SCHOOL	YEAR						
	1	2	3	4	.	.	m
1	X_{11}	X_{12}	X_{13}	X_{14}	.	.	X_{1m}
2	X_{21}	X_{22}	X_{23}	X_{24}	.	.	X_{2m}
3	X_{31}	X_{32}	X_{33}	X_{34}	.	.	X_{3m}
4	X_{41}	X_{42}	X_{43}	X_{44}	.	.	X_{4m}
.
.
.
n	X_{n1}	X_{n2}	X_{n3}	X_{n4}	.	.	X_{nm}

The effects may vary. For example, the school effect, $(\mu_s - \mu)$, may vary across schools in the population. Each effect then has an associated variance, which is called a component of variance. For the effect of schools, years, and the residual, the variance components are, respectively, σ_s^2 ,

σ_y^2 , and $\sigma_{sy,e}^2$. The variance of X_{sy} , taken over all schools in the population and all years in the universe of observations, is the sum of the three variance components. Just as the variance of X_{sy} is the variance of a single school's score taken in a single year, variance components are also associated with one school and one year. This explanation will become important later as we use the variance components to estimate the effects of changing the number of years of data that enter a school's score.

Variance components are the parameters estimated in G-theory analyses. They can be estimated through application of Analysis of Variance (ANOVA) because they are the components of the expected mean squares in ANOVA. (For details about the use of ANOVA to estimate variance components see Brennan, 2001 or Cronbach, Gleser, Nanda, and Rajaratnam, 1972.) With variance components, it is possible to identify how variability in observed scores is expected to be affected by different facets of the universe and by sampling different numbers of observations from each facet. It also becomes possible to examine how the variability in scores would be affected by different types of designs for data collection. As a result, it is possible to consider how to maximize the information one wants from a score and minimize the impact of other sources of variability and, thus, to design data collection plans that provide dependable measurements. We review a few basic uses of the variance components here as they apply to the proposed study.

First, a key concept in G-theory is that, unlike in Classical Test Theory (CTT), there is not a single “true” score or “error score.” In CTT a person’s observed score is simply the sum of a “true” score and an “error” score. G-theory recognizes multiple sources of influence on scores. Which of those influences enters the “true” score and which enters the “error” score depends on the use of the score and how it is to be interpreted.

In the case of the simple one-facet crossed design, there are three effects, each having an associated variance component: school, year, and the residual. The residual component is equivalent to the error variance of CTT. The school component is equivalent to the true score variance of CTT. This leaves the year component, which does not have a comparable term in CTT because CTT assumes it to be zero (i.e., it assumes strictly parallel forms of tests or measurement methods, each form having the same mean value of scores). G-theory relaxes that assumption. For the school-level status measure (derived from student scores on the statewide achievement test), the variance component for year would be greater than zero if there had been a change in the statewide test that resulted in the test becoming easier, on average, than the previous years’ tests, for example. That type of change may be a source of error for some decisions but not for others. It would be a source of error when the scores of schools are used to make a decision that involves an “absolute” comparison, such as the comparison of a school’s score to a criterion score that schools must meet to be judged effective. A change in the average level of difficulty of the test from one year to the next could change a school’s position relative to the criterion, even in cases where the school’s effectiveness had not changed. In contrast, the change in test difficulty would not be a source of error for decisions involving “relative” comparisons, such as a decision to select the top 10 percent of the schools for awards. The rank order of schools would not be affected by a shift in test difficulty that adds a constant to each school’s score.

Thus, even for the simple single-facet design, we can express the SEM in two ways, depending on the nature of the decisions for which the scores will be used. For absolute decisions, the SEM includes both the year and residual variance components and is defined as:

$$SEM_{abs} = \sqrt{\frac{\sigma_y^2}{n_y} + \frac{\sigma_{sy,e}^2}{n_y}} , \quad [2]$$

Where n_y is the number of years of data that are averaged to obtain the school's scores. If we are using a score from a single year of data, then the SEM is simply based on the sum of the two variance components. The SEM is reduced (and reliability increased) if two or more years of data are sampled and averaged to obtain a school's score.

For relative decisions, the SEM includes only the residual component and is defined as:

$$SEM_{rel} = \sqrt{\frac{\sigma_{sy,e}^2}{n_y}} . \quad [3]$$

The generalizability coefficient (g-coefficient) is analogous to the reliability coefficient of CTT in that it is defined as a ratio of the variance among schools to the total variance. For absolute decisions this is⁸:

$$\rho_{abs}^2 = \frac{\sigma_s^2}{\sigma_s^2 + \frac{\sigma_y^2}{n_y} + \frac{\sigma_{sy,e}^2}{n_y}} \quad [4]$$

For relative decisions, the g-coefficient is:

$$\rho_{rel}^2 = \frac{\sigma_s^2}{\sigma_s^2 + \frac{\sigma_{sy,e}^2}{n_y}} \quad [5]$$

Note that equations 2 through 5 all contain the term n_y , the number of years of data that enter the school's score. By sampling more years, one would expect to reduce the SEMs and increase the reliability coefficients. Thus, once the components of variance have been estimated, the SEMs and g-coefficients can be estimated for situations that differ in the number of years of data that are averaged to obtain a school's score.

⁸ For absolute decisions, Brennan (2001) refers to this as an index of dependability rather than a g-coefficient, but for simplicity we refer to both the coefficients derived for absolute and relative decisions as g-coefficients.