Effect Sizes for Quantifying Student and School Growth in Achievement: In Search of Practical Significance

James Soland
Assistant Professor of Quantitative Methods
University of Virginia

Yeow Meng Thum
Senior Research Fellow
NWEA

Background and Context:

With challenges of determining practical significance in empirical studies, effect sizes are often used because they are not dependent on a particular significance criterion, sample size, or scale. One application of effect sizes that has received little attention concerns effect sizes for estimates of students' academic growth (Bloom et al., 2008; Hill et al., 2008; Lipsey et al., 2012). There are several potential limitations of the effect sizes for academic growth currently used. First, many studies use effect sizes that make strong assumptions about the nature of the growth occurring, including an implicit assumption that within-student scores are uncorrelated over time. Second, when inferences are being made about growth in the aggregate at, say, the school level, growth is often standardized in units of student-level gains (Thum & Hauser, 2015). Given that the variance in aggregate growth is likely to be smaller than at the individual level, one could imagine the consequence is effect sizes that are smaller when the denominator is student-level variance. Third, many growth studies do not have empirical benchmarks (estimates of growth in the population) as a yardstick against which to compare sample growth.

Purpose and Research Questions

Our study investigates the implications of using common effect sizes for growth then provides empirical benchmarks for growth. We use a national sample of schools whose students took the Measures of Academic Progress (MAP) Growth (a cross-grade computerized adaptive achievement test with a vertical scale) from 2011-15. The predicted joint distribution of student scores was obtained from a novel multilevel model that can simultaneously estimate between and within year growth for students and schools. Using this approach, benchmarks for effect sizes for growth for any between- or within-year period of time in the sample can be produced that account for days of instruction, initial achievement level, and summer loss. This data-model combination means we were able to examine the practical consequences of taking different approaches to producing effect sizes for test score gains by investigating two research questions:

1. How different are effect size estimates of growth in student achievement when standardized relative to a distribution of gains rather than a distribution of scores from one timepoint (or pooled SDs from two timepoints)?
2. How different are effect size estimates of school-level growth when standardized relative to distributions of student- versus school-level gains?

In our study, we estimate effect sizes using different approaches, then compare them quantitatively and interpretively while providing national benchmarks for growth.

Data and Sample

We use MAP Growth results from six age-cohorts of students who took the MAP Growth mathematics and reading tests. Up to three years of longitudinal test scores are used for each student in a -cohort, with a maximum of nine scores spanning three grade levels (see Table 1). To achieve meaningful generalizability on MAP performance among the US population of public schools, we derived school-level post-stratification weights. The NCES school characteristics included measures of school poverty, racial make-up, type (e.g., charter), grades served, and location. Details on the weights and weighting procedure are found in Authors (2015).

Research Design and Analysis

We began by specifying various effect sizes used in most research on growth then comparing them to effect sizes that make less strong assumptions about growth. We examined the test score for grade g, y_g, and the variance of that test score for a given sample and grade, s_g^2. One commonly used effect size standardizes the mean difference relative to the variance at Time 1 (we use grades three and four as an example):

$$\frac{y_4 - y_3}{s_3} \quad (1a)$$

Alternatively, one could produce the following effect size:

$$\frac{y_4 - y_3}{\sqrt{\frac{s_4^2 + s_3^2}{2}}} \quad (1b)$$

This effect size can be interpreted as the mean test score difference between 3rd and 4th grade students at Time 1 scaled relative to the pooled SD of the scores in those grades. These effect sizes have been used extensively in the literature (for example, Hill et al., 2008).

An oft overlooked problem with Hill et al.'s (2008) effect size is that it ignores the correlation between within-student test scores over time. In Equations 1(a) and 1(b), the implicit null hypothesis is that no growth is occurring. By contrast, Equation 2 below accounts for that correlation:

$$\frac{y_4 - y_3}{\sqrt{s_4^2 + s_3^2 - 2cov(y_4, y_3)}} \quad (2)$$

Here, $cov(y_4, y_3)$ is the covariance of the test scores between Grade 4 at Time 2 and Grade 3 at Time 1. Without including this term, the assumption is that scores are uncorrelated over time (Gibbons, Hedeker, & Davis, 1993; Zimmerman, 1997). Research suggests, however, that this assumption is not tenable (Authors, 2019). Further, the effect size in Equation 1 is likely to be

smaller than the effect size in Equation 2 because the denominator for the latter will be smaller when the correlation between within-cohort test scores across time points is nonzero.

Another effect size for growth directly related to Equation 2 is the following:

$$\frac{y_4 - y_3}{SD(y_4 - y_3)} \quad (3)$$

In this equation, we divide the gain between third and fourth grade by the standard deviation of that gain. One can show that this effect size is equivalent to the one in Equation 2 given $var(Y - X) = var(Y) + var(X) - 2cov(Y, X)$.

All four effect sizes can also be adapted to aggregate level inferences by replacing student-level means and standard deviations with school-level equivalents. Table 2 summarizes these effect sizes.

Results

Tables 3-4 show estimates of student growth (math and reading) assuming a generic gain of 5 RIT (points on the MAP Growth scale) using the effect sizes previously described. Effect sizes are often twice as large when the within-student correlation is accounted for in the effect size equation. Tables 4-5 show the same results, but at the school level and therefore standardized relative to school rather than student gains. Effect sizes for school-level growth are much larger when the effect size denominator is school-level rather than student-level variances.

Conclusion

Our findings indicate that many effect sizes used to quantify and compare student growth make strong assumptions about the nature of that growth. Further, effect size estimates differ substantively dependent on the effect size used. Additional consideration should be given to how to establish best practices as a field.

References

Authors (2015).

Authors (2019).

Bloom, H. S., Hill, C. J., Black, A. R., & Lipsey, M. W. (2008). Performance trajectories and performance gaps as achievement effect-size benchmarks for educational interventions. *Journal of Research on Educational Effectiveness, 1*(4), 289–328.

Gibbons, R. D., Hedeker, D. R., & Davis, J. M. (1993). Estimation of effect size from a series of experiments involving paired comparisons. *Journal of Educational Statistics, 18*(3), 271–279.

Hill, C. J., Bloom, H. S., Black, A. R., & Lipsey, M. W. (2008). Empirical benchmarks for interpreting effect sizes in research. *Child Development Perspectives, 2*(3), 172–177.

Zimmerman, D. W. (1997). Teacher's corner: A note on interpretation of the paired-samples t test. *Journal of Educational and Behavioral Statistics, 22*(3), 349–360.

Table 1

Means and SDs of Achievement for the Sample by Term, Sample Sizes by Year

		Mathematics			Reading		
Grade	Statistic	Fall	Winter	Spring	Fall	Winter	Spring
3	Mean	190.4	198.21	203.4	188.29	195.6	198.62
	School SD	5.88	6.12	6.46	6.7	6.59	6.63
	Student SD	13.1	13.29	13.81	15.85	15.14	15.1
	Ns (schools, students, test scores)	1456, 135458, 609075			1457, 134519, 606602		
4	Mean	201.94	208.71	213.49	198.16	203.6	205.92
	School SD	6.11	6.55	7.03	6.49	6.43	6.48
	Student SD	13.76	14.27	14.97	15.53	14.96	14.92
	Ns (schools, students, test scores)	1443, 130077, 592305			1456, 134361, 615399		
5	Mean	211.44	217.23	221.36	205.68	209.83	211.79
	School SD	7	7.57	8.16	6.27	6.23	6.31
	Student SD	14.68	15.33	16.18	15.13	14.65	14.72
	Ns (schools, students, test scores)	1440, 148818, 518378			1448, 148564, 526716		
6	Mean	217.62	222.06	225.32	210.99	214.21	215.75
	School SD	7.19	7.61	8.05	6.31	6.31	6.44
	Student SD	15.53	16	16.71	14.94	14.53	14.66
	Ns (schools, students, test scores)	1451, 165541, 546568			1452, 162887, 554445		
7	Mean	222.65	226.12	228.59	214.45	216.91	218.16
	School SD	7.71	8.06	8.43	6.57	6.57	6.69
	Student SD	16.59	17.07	17.72	15.31	14.98	15.14
	Ns (schools, students, test scores)	1415, 190705, 781050			1418, 194033, 814818		
8	Mean	226.3	229.15	230.93	217.24	219.09	220.07
	School SD	8.8	9.21	9.62	7.47	7.39	7.48
	Student SD	17.85	18.31	19.11	15.72	15.37	15.73
	Ns (schools, students, test scores)	1377, 199759, 619604			1396, 206667, 664327		

Table 2

Taxonomy of Effect Sizes for Growth in Achievement

Decisions			Research Question
Account Within-Student Correlation?	No. of Groups Compared	Appropriate Equation	How large is the gain in the sample/population relative to:
No	1	Eq. 1(a) or 1(b)	The SD at one timepoint or the pooled SDs of pre- and post-test scores?
Yes	1	Eq. 2	The pooled SDs of pre- and post-test scores accounting for a pre/post correlation?
Yes	1	Eq. 3	The SD of the pre- and post-test score gain?

Note. All of these effect sizes can be produced at the school-level as well by replacing the student-level subscripts in Equations 1-6 with school-level subscripts.

Table 3

Comparison of Different Effect Sizes for Mean Student Gains in Math

Grade	Point in Time		Gain		5 RIT Gain		
	Mean	SD	Mean	SD	Gain/SD Time1	Gain/Pooled SD	Gain/(SD of the Gain)
					Eq. 1(a)	Eq. 1(b)	Eq. 3
3	191.59	13.62	11.81	6.76	X	X	X
4	203.54	14.12	9.95	6.63	0.37	0.36	0.75
5	212.74	15.77	8.62	7.01	0.35	0.33	0.71
6	220.44	17.35	4.88	7.15	0.32	0.30	0.70
7	223.68	17.78	4.92	6.59	0.29	0.28	0.76
8	226.92	18.55	4.00	7.77	0.28	0.28	0.64

Table 4

Comparison of Different Effect Sizes for Mean Student Gains in Reading

Grade	Point in Time		Gain		5 RIT Gain		
	Mean	SD	Mean	SD	Gain/SD Time1	Gain/Pooled SD	Gain/(SD of the Gain)
					Eq. 1(a)	Eq. 1(b)	Eq. 3
3	188.15	16.25	10.47	7.25	X	X	X
4	198.39	15.97	7.54	6.83	0.31	0.31	0.73
5	205.83	15.57	5.96	7.19	0.31	0.32	0.70
6	211.55	15.92	4.20	7.48	0.32	0.32	0.67
7	214.33	16.19	3.83	7.05	0.31	0.31	0.71
8	216.99	15.89	3.09	8.30	0.31	0.31	0.60

Table 5

School Level Comparison of Different Effect Sizes for Mean Student Gains in Math

Grade	Point in Time		Gain		5 RIT Gain		
	Mean	SD	Mean	SD	Gain/SD Time 1	Gain/Pooled SD	Gain/(SD of the Gain)
					Eq. 1(a)	Eq. 1(b)	Eq. 3
3	191.59	6.29	11.81	2.74	X	X	X
4	203.54	6.48	9.95	2.44	0.79	0.78	2.05
5	212.74	7.78	8.62	2.85	0.77	0.70	1.75
6	220.44	8.11	4.88	2.53	0.64	0.63	1.98
7	223.68	8.50	4.92	2.20	0.62	0.60	2.27
8	226.92	9.18	4.00	2.50	0.59	0.57	2.00

Table 6

School Level Comparison of Different Effect Sizes for Mean Student Gains in Reading

Grade	Point in Time		Gain		5 RIT Gain		
	Mean	SD	Mean	SD	Gain/SD Time 1	Gain/Pooled SD	Gain/(SD of the Gain)
					Eq. 1(a)	Eq. 1(b)	Eq. 3
3	188.15	6.85	10.47	1.95	X	X	X
4	198.39	6.80	7.54	1.74	0.73	0.73	2.87
5	205.83	6.63	5.96	1.80	0.74	0.74	2.78
6	211.55	6.80	4.20	2.20	0.75	0.74	2.27
7	214.33	7.16	3.83	2.09	0.74	0.72	2.39
8	216.99	7.36	3.09	2.24	0.70	0.69	2.23