Time-indexed Effect Size for P-12 Reading and Math Program Evaluation

Jaekyung Lee

University at Buffalo, SUNY

Jeremy Finn

University at Buffalo, SUNY

Xiaoyan Liu

George Mason University

Paper presented at the Society for Research on Educational Effectiveness (SREE) spring 2012 conference, Washington, DC on March 9, 2012. This research was supported by research grant # R305D090021 from the U.S. Department of Education. The views expressed herein are solely those of the authors.

Abstract

Our current capacity to understand or provide a context for interpreting the size of an effect in education program evaluation is limited. To address the problem, this study proposes a time-indexed effect size metric (d') to estimate how long it would take for an "untreated" control group to reach the treatment group outcome in terms familiar to educators—years/months of schooling. This study extends prior work on K-12 academic growth norms to preschool level. Through applications of the time-indexed effect size to selected examples of experimental or quasi-experimental research (e.g., What Works Clearinghouse reviews), the study demonstrates that it could provide a developmentally appropriate context for interpretations of educational program effects at different levels of schooling. It is a step toward bridging the gap between educational research and practice.

While there has been much discussion of the role and function of effect sizes in social and behavioral research, there is general agreement that effect sizes are valuable tools to help evaluate the magnitude of a difference or relationship, particularly, whether a statistically significant difference is a difference of practical concern (see Cohen, 1994; Cooper & Hedges, 1994; Kirk, 1996; Schmidt, 1996; Thompson, 1996; Wilkinson & APA Task Force on Statistical Inference, 1999). Accordingly, effect size reporting has now become a de facto requirement for publication. Researchers are asked to provide readers with information to assess the magnitude of the observed effect or relationship as the basis of judgments about practical or clinical significance in conjunction with statistical significance testing (APA, 2001; Knapp & Sawilowsky, 2001; Thompson, 2001).

However, it is still challenging for practitioners to understand or translate a metric representing a standardized group mean difference on a more familiar yardstick such as years/months of schooling. While educational treatment effects have been sometimes reported in terms of grade-equivalent (GE) units (Finn et al., 2001; Gormley et al., 2005; Seltzer, Frank & Bryk, 1994), conventional GEs have many limitations due to their reliance on test-specific publisher's proprietary norms derived from aggregated cross-sectional data and restricted to K-12 (Peterson, Kolen, & Hoover, 1989; Schulz & Nicewander, 1997). In light of these problems, this study develops new national norms of academic growth based on longitudinal national datasets in P-12 reading and math, and applies a time-indexed effect size metric with those new norms to education program evaluation.

Conceptual Framework

Extending our prior research on K-12 academic growth trajectories (Lee, Finn, & Liu, 2011) to the preschool level, the present research attempts to address the question: "How much time is needed for students in the control group to catch up with students in the treatment group?" The rationale for time-indexed assessment of effect sizes comes from the well-established pattern of curvilinear or piecewise linear academic growth patterns over the entire course of child development and education (see Beggs & Hieronymus, 1968; CTB/McGraw-Hill, 1997, 2003; Harcourt, 2002, 2004; Lee, 2010; Lichten, 2004; McGrew & Woodcock, 2001). Studies also observed the likelihood of greater environmental effects or intervention effects at the earlier stage of development when the pace of academic growth is relatively faster (Bloom, 1964; Ramey & Ramey, 1998). Time-indexed effect size would enable educational researchers to more accurately assess effect sizes in the context of students' developmental stage or grade level when the intervention occurs.

This study contextualizes an effect-size-like index of educational treatment effects or any group mean differences in academic achievement by referencing time. The new effect size metric can enrich effect size interpretations while serving as a supplement (but not substitute) for conventional standardized effect size measures. Specifically, we introduce a new time-indexed effect size metric (d') based on the notion of time-varying academic growth trajectories in P-12 reading and math as evidenced through empirical analyses of U.S. national longitudinal datasets.

Figure 1 illustrates the concept and measurement of time-indexed effect size based on hypothetical linear patterns of academic growth for an experimental group (E) and a control group (C) in a particular grade. Assuming that both groups have the same average pretest scores, Y^E and Y^c represent the average posttest scores of the outcome variable Y for the experimental group and control group respectively.[1] Unlike a conventional effect size measure that focuses on the group difference on the Y (outcome variable) axis, we switch the focus to the X (time) axis. The time-indexed effect size (d') is the extra time (in school years/months) needed for the control group to reach Y^E, the outcome that the experimental group has reached at the end of treatment (see Figure 1):

$d' = T_2 - T_1$

where T_2 = time needed (in school years/months) for the control group to reach Y^E from baseline (time zero);

T_1 = time spent (in school years/months) for the experimental group to reach Y^E or for the control group to reach Y^c from the baseline (time zero)

Indeed, existing national norms from test publishers can provide general reference points since the tests not only have been widely used in many school districts across the nation, but are also derived from nationally-representative norming samples with vertical scales of achievement; the norms usually cover every grade from K to 12 with test administrations in both fall and spring. Prior research attempted to use such test norms to establish grade-referenced benchmarks for effect size interpretations in core subjects (Bloom, Hill, Black & Lipsey, 2008). Although the test publisher data provide useful references of academic growth for all grades in many subjects, those norms derived from cross-sectional snapshot data from multiple cohorts may not accurately represent true longitudinal growth by confounding cohort effects and grade effects. Further, test publisher data is aggregated, and lacks information on student subgroup differences in growth

[1] In Figure 1, it is assumed that the two groups' pretest scores are equal as shown by the same baseline position at Time zero as a result of random assignment and/or matching. However, if the two groups' average pretest scores turn out to be different, then their average posttest scores can be adjusted based on the initial status difference in calculating time indexed-effect size.

norms. This prevents researchers from using matching or other adjustment methods that would take into account possible differences between their study sample and national norming sample. This study addresses those problems by using longitudinal datasets to create growth norms for P-12 and disaggregating the results by subgroups.

Methods

In this study, we constructed national norms of academic growth for P-12 reading and math achievement through the analysis and synthesis of existing longitudinal datasets (see Figure 2). Test publisher norms are based on seasonal testing schedules that can provide gains from fall to spring within same school years and then gains (or losses) from spring to fall between adjacent school years. In contrast, national longitudinal data usually are based on annual or biennial (or even longer time span) testing schedules that only provide gains between adjacent or remote school years. This study capitalizes on three separate national longitudinal datasets, the Early Childhood Longitudinal Study-Birth Cohort (ECLS-B), the Early Childhood Longitudinal Study-Kindergarten Cohort (ECLS-K) and the National Education Longitudinal Study of 1988 (NELS:88) to construct our own national norms of academic growth. These National Center for Education Statistics (NCES) datasets provide information on a child's academic growth along with background characteristics of the child, family, and school. The ECLS-B followed academic growth trajectories from preschool (age 4) to Kindergarten. The ECLS-K followed academic growth trajectories from Kindergarten to grade 8. The NELS tracked individual students' academic growth from grade 8 to grade 12.

Longitudinal analyses of the ECLS-B, ECLS-K and NELS databases were carried out with data weighted by appropriate panel weights. Analysis of a weighted sample provides results that are representative of the population from which participants were drawn. For ECLS-B, the analytical sample was restricted to children born in 2001 whose math knowledge/skills were assessed at both age 4 and Kindergarten; they entered Kindergarten for the first time during either 2006 or 2007 (N= 6,051). For ECLS-K, typical students refer to those who spent one year in kindergarten, and who entered grade 1 the following year and grade 3 two years later, etc. (N=5,959); students who were repeating kindergarten in 1998, or who were not in Kindergarten, grade 1, grade 3, grade 5, and grade 8 at the time of each spring follow-up assessment, were not included in the analysis. Likewise, the NELS sample used for this study was comprised of only students who were in grade 8 for the first time in the fall of 1988, and who were in grade 10 in the spring of 1990 and in grade 12 in the spring of 1992 (N=10,879).

Examination of the growth curve was carried out using the IRT estimated number right scores for reading and math in the respective surveys. We created national norms of academic growth by computing g, standardized measures of reading and math achievement gain scores (in pooled standard deviation units) between successive grades. Because the assessments do not cover all grades, gains were computed only between successive waves of assessments available in the datasets (i.e., preK-fall K in ECLS-B; fall K-spring K, K-grade 1, grades 1-3, grades 3-5, grades 5-8 in ECLS-K; grades 8-10 and grades 10-12 in NELS). We used equation below to compute *g* values with descriptive statistics of academic growth for all students as well as by subgroups as classified by key background variables (gender, race/ethnicity, poverty, parent education, school type and location).

$$g = \sum_t \left[\frac{\left(\bar{Y}_{t+1} - \bar{Y}_t\right)}{\sqrt{\left(s^2{}_{t+1} + s^2{}_t\right)/2}} \right]$$

where $\bar{Y}_t$ = mean of test score at time t; $s^2{}_t$ = variance of test score at time t

Then annual growth rates were estimated by dividing standardized test score gains by elapsed time in months between successive waves of assessments, and multiplying by 10 to obtain the full school year gain. These final *g* values (estimated standardized gains per school year) are shown in Table 1, where interpolation method was used to estimate gains for missing grades (grades 2, 6, 7, 9, 11). The g values were used as a denominator to convert *d* (standardized group mean differences) into *d′* (years/months of schooling) in corresponding subjects and grades, using the formula:

$$d' = \frac{d}{g}$$

For quick reference, we constructed a table of conversions (see Table 2). Three common benchmark values of Cohen's *d* (0.2 for small effect, 0.5 for medium effect and 0.8 for large effect) were converted into years/months of schooling by dividing *d* values by corresponding *g* values in Table 1. We followed the same steps to construct the conversion table for demographic subgroups based on their national longitudinal growth subgroup norms: the subgroup variables include race. Online conversion program has been developed, and it requires researchers' inputting effect size information on subject (reading, math), grade (P-12), category (subgroups), and *d* (see Figure 3).

Results

According to the conversion table for reading (Table 2), the effect size for a reading program with *d*=0.2 (i.e., 20% of one standard deviation) in pre-K (age 4) and Kindergarten would be equivalent to two months ($d' = 0.2$) and one month of schooling ($d' = 0.1$) respectively. The same "small" effect turns into the longer time of schooling at upper grades: the effect size of

.2 would become worth four months ($d' = 0.4$) in grade 4, one year in grade 8 ($d' = 1.0$), and three years plus four months ($d' = 3.4$) in grade 12. For a math program with a small effect (d=0.2), the time-indexed effect size would vary from two months ($d' = 0.2$) in pre-K, one month ($d' = 0.1$) in Kindergarten, three months ($d' = 0.3$) in grade 4, nine months ($d' = 0.9$) in grade 8, and one year plus three months ($d' = 1.3$) in grade 12. For both reading and math growth norms, the time-indexed effect size tends to increase gradually over the course of schooling until grade 12.

We applied time-indexed effect size formula to selected examples of curricular interventions in P-12 that provided information on intent-to-treat (ITT) effect sizes and met evidence standards by What Works Clearinghouse (WWC)[2]. Table 4 summarizes the results.

For preschool 4-year old cohort, the evaluation of Head Start impact showed significant effect with average $d = 0.20$ in language/literacy and insignificant effect (effect size was not reported) in math (Puma et al., 2010). Using the conversion formula, this program effect on reading is equivalent to approximately two additional months of learning in that preschool year ($d' = 0.20/1.06 = 0.19$). For second-grade students, the evaluation of elementary school math curricula showed that *Saxon Math* schools scored 0.17 standard deviations higher than *Scott Foresman-Addison Wesley Mathematics* schools (Agodini et al., 2010). This program effect is roughly equivalent to one month of school learning plus one-third of another month ($d' = 0.17/1.27 = 0.13$).

[2] According to "WWC QUICK REVIEW PROTOCOL (VERSION 2.0)" the rating of *Meets WWC evidence standards* applies to well-executed randomized controlled trials, regression discontinuity studies, and single-case studies. There were some WWC-reviewed interventions that did not target specific grades or break down results by grades. For example, the evaluation of Washington DC scholarship opportunity program for K-12 students (voucher for private schools) shows insignificant effects with average $d = .11$ in reading and $d = .02$ in math (Wolf et al., 2010). Because the aggregated K-12 results were not reported separately for different grades, we could not translate d into d'.

For a five-year longitudinal study of Spanish-speaking Kindergarten students, the comparison of the transitional bilingual education group with the structured English immersion group showed that SEI group performed better than TBE group in reading but the gap became smaller and changed from significant to insignificant by the end of grade 4 (d = .54 in K; d = .42 in grade 1; d = .20 in grade 2; d = .16 in grade 3; d = .25 in grade 4) (Slavin et al., 2010). When these program effects are translated into school time units, it turns out that the SEI advantage of reading gain does not diminish as much over time due to increasingly slower pace of learning at the upper grades (d' = .33 in K; d' = .24 in grade 1; d' = .16 in grade 2; d' = .20 in grade 3; d' = .46 in grade 4). An evaluation study of supplemental literacy classes for struggling ninth-grade readers (Corrin et al., 2009) found that the effect size for reading comprehension was 0.08, equivalent to about three months of schooling (d' = 0.08/0.26 = 0.31).

An advantage of using our disaggregated norms by subgroups (e.g., racial breakdown as shown in Table 3) is that it allows for differentiation of program effects based on subgroup-specific growth rates. For example, the evaluation of Head Start impact on 4-year old cohort's basic reading skills during Kindergarten showed significantly more favorable impact on Blacks than on Whites (Puma et al., 2010). The program effect was d = .40 for Blacks vs. d = -.19 for Whites, and they are equivalent to d' = .40/1.67 = 0.24 for Blacks vs. d' = -.19/1.52 = -0.13 for Whites respectively; the Black-White gap in program benefit for reading amounts to 3-4 months.

For Project STAR class size experiment (Finn et al., 2001), Table 4 demonstrates how the effects of small classes changed from K to grade 3 as a result of the choice of different effect size metrics. Conventional effect size d shows that ion reading, the small-class advantage declined in each subsequent year (except for slight rebound in grade 3 reading). In contrast, time-indexed effect size d' show that the small class effect in reading remained fairly stable or sometimes increased at the higher grades. In math, for example, we estimate that it would take students in larger classes about one and half months to catch up to the performance of students in smaller classes across grades K-3 (d' = .15 in K; d' = .16 in grade 1; d' = .15 in grade 2; d' = .16 in grade 3).

Conclusion

Our current capacity to understand or provide a context for interpreting the size of an effect in education program evaluation is limited. To address the problem, we proposed a time-indexed effect size metric to estimate how long it would take for an “untreated” control group to reach the treatment group outcome in terms familiar to educators—years/months of schooling. This study extends prior work on K-12 academic growth norms (Lee, Finn & Liu, 2011) to preschool level with ECLS-B data. Applications of the time-indexed effect size d' to the selected examples of prior research demonstrate that it could provide a more developmentally appropriate context for interpretations of educational program effects at different levels of schooling. It is a step toward bridging the gap between educational research and practice.

Applying our new national longitudinal growth norms to time-indexed effect size calculation for specific research findings requires caveats as it requires checking two major assumptions. If researchers apply national norms to estimate typical growth under the control group situation and compare it with their own study sample results, they should check construct equivalence between norm data and study data; that is, how well the specific test used to measure the effect of the intervention aligns with the test used to develop national norms. Further, there should be a reasonable match between the study sample and the norm group.

References

Agodini, R., Harris, B., Thomas, M., Murphy, R., Gallagher, L., & Pendleton, A. (2010). *Achievement effects of four early elementary school math curricula: Findings for first and second graders* (NCEE 2011-4001). Washington, DC: National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, U.S. Department of Education.

American Psychological Association. (2001). *Publication manual of the American Psychological Association* (5th ed.). Washington, DC: Author.

Beggs, D. L., & Hieronymus, A. N. (1968). Uniformity of growth in the basic skills throughout the school year and during the summer. *Journal of Educational Measurement, 5*, 91-97.

Bloom, (1964). *Stability and Change in Human Characteristics.* New York, John Wiley & Sons.

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*, 289-328.

Cohen, J. (1994). The earth is round ($p < .05$). *American Psychologist, 4*9, 997-1003.

Cooper, H. & Hedges, L. V. (Eds.). (1994) *The handbook of research synthesis.* New York: Russell Sage Foundation.

Corrin, W., Somers, M.-A., Kemple, J., Nelson, E., & Sepanik, S. (2009). *The Enhanced Reading Opportunities study: Findings from the second year of implementation* (NCEE 2009-4036). Washington, DC: National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, U.S. Department of Education.

CTB/McGraw-Hill (1997). Technical Bulletin 1. Monterey, CA: Author.

CTB/McGraw-Hill. (2003). *TerraNova 2nd Edition CAT Technical Report.* Monterey, CA: Author.

Finn, J. D., Gerber, S. B., Achilles, C. M., & Boyd-Zaharias, J. (2001). The enduring effects of small classes. *Teachers College Record*, *103*(2), 145-183.

Gormley, W.T., Gayer, T., Phillips, D., & Dawson, B. (2005). The effects of universal pre-K on cognitive development. *Developmental Psychology, 41*(6), 872-884.

Harcourt Educational Measurement. (2002). *Metropolitan8 Form V Technical Manual* (Metropolitan Achievement Tests 8th edition). San Antonio, TX: Author.

Harcourt (2004). *Stanford Achievement Test 10th Edition Technical Data Report*. San Antonio, TX: Author.

Hirschhorn, D. B. (1993). A longitudinal study of students completing four years of UCSMP mathematics. *Journal for Research in Mathematics Education*, *24*(2), 136–158.

Kane, M. T. (2006). Validation. In R. L. Brennan (Ed.), *Educational Measurement* (4th ed., pp. 17–64). Westport, CT: Praeger Publishers.

Kirk, R. E. (1996). Practical significance: A concept whose time has come. *Educational and Psychological Measurement, 56*, 746-759.

Knapp, T. R., & Sawilowsky, S. S. (2001). Strong arguments: Rejoinder to Thompson. *The Journal of Experimental Education*, *70,* 94-95.

Lee, J. (2010). Tripartite Growth Trajectories of Reading and Math Achievement: Tracking National Academic Progress at Primary, Middle and High School Levels. *American Educational Research Journal, 47*(4), 800-832.

Lee, J., Finn, J. & Liu, XY. (2011, April). Developing Time-Indexed Effect Size Metrics for K-12 Reading and Math Educational Evaluation. Paper presented at the annual meeting of American Educational Research Association (AERA), New Orleans, Louisiana.

Lichten, W**.** (2004). On the Law of Intelligence. *Developmental Review, 24*(3), 252-288.

McGrew, K. S. & Woodcock, R. W. (2001). *Woodcock-Johnson III Technical Manual.* Itasca, IL: Riverside Publishing.

Petersen, N.S., Kolen, M.J., & Hoover, H.D. (1989). Scaling, norming, and equating. In R.L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 221–262). New York: Macmillan.

Puma, M., et al. (2010). Head Start impact study: Final report. Washington, DC: Administration for Children and Families, U.S. Department of Health and Human Services.

Ramey, C.T., & Ramey, S.L (1998). Early intervention and early experience. American Psychologist, 53(2), 109-120

Riccio, J., Dechausay, N., Greenberg, D., Miller, C., Rucks, Z., & Verma, N. (2010). *Toward reduced poverty across generations: Early findings from New York City's conditional cash transfer program.* New York, NY: MDRC.

Schmidt, F. (1996). Statistical significance testing and cumulative knowledge in psychology: Implications for the training of researchers. *Psychological Methods, 1*, 115-129.

Schulz, E.M. and Nicewander, W.A., (1997). Grade equivalent and IRT representations of growth. *Journal of Educational Measurement,* ***34***, 315–331.

Seltzer, M. H., Frank, K.A., & Bryk, A.S. (1994). The Metric Matters: The Sensitivity of Conclusions Concerning Growth in Student Achievement to Choice of Metric. *Educational Evaluation and Policy Analysis, 16*(1), 41-49.

Slavin, R. E., Madden, N., Calderon, M., Chamberlain, A., & Hennessy, M. (2010). *Reading and language outcomes of a five-year randomized evaluation of transitional bilingual education.* Baltimore, MD: Johns Hopkins University.

Thompson, B. (2001). Significance, effect sizes, stepwise methods, and other issues: Strong arguments move the field. *Journal of Experimental Education, 70*, 80-93.

Thompson, B. (1996). AERA editorial policies regarding statistical significance testing: Three suggested reforms. *Educational Researcher, 2*5(2), 26-30.

Wilkinson, L. & American Psychological Association Task Force on Statistical Inference. (1999). Statistical methods in psychology journals: Guidelines and explanation. *American Psychologist, 54*, 594-604. [reprint available through the APA Home Page: http://www.apa.org/journals/amp/amp548594.html]

Wolf, P., Gutmann, B., Puma, M., Kisida, B., Rizzo, L., Eissa, N., & Carr, M. (2010). *Evaluation of the DC Opportunity Scholarship Program: Final report* (NCEE 2010-4018). Washington, DC: National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, U.S. Department of Education.

Table 1
National Longitudinal Data-based Norms of Academic Growth in P-12 Reading and Math: Standardized Achievement Gains per School Year (10 Months) by Subject and Grade

grades	(1) Reading Gains g_r	(2) Math Gains g_m
PreK	1.06	1.04
K	1.66	1.76
1	1.76	1.66
2	1.23	1.27
3	0.81	0.95
4	0.54	0.77
5	0.50	0.73
6	0.35	0.44
7	0.27	0.33
8	0.20	0.22
9	0.26	0.47
10	0.40	0.47
11	0.37	0.67
12	0.06	0.15

Table 2

Time-indexed effect sizes based on the national norms of academic growth in P-12 reading and math: conversion of *d* (standardized group mean differences) to *d´* (years/months of schooling)

	Reading			Math		
	small	*d* medium	large	small	*d* medium	large
grades	0.2	0.5	0.8	0.2	0.5	0.8
PreK	0.2	0.5	0.8	0.2	0.5	0.8
K	0.1	0.3	0.5	0.1	0.3	0.5
1	0.1	0.3	0.5	0.1	0.3	0.5
2	0.2	0.4	0.6	0.2	0.4	0.6
3	0.2	0.6	1.0	0.2	0.5	0.8
4	0.4	0.9	1.5	0.3	0.7	1.0
5	0.4	1.0	1.6	0.3	0.7	1.1
6	0.6	1.4	2.3	0.5	1.1	1.8
7	0.8	1.9	3.0	0.6	1.5	2.4
8	1.0	2.5	4.0	0.9	2.2	3.6
9	0.8	1.9	3.1	0.4	1.1	1.7
10	0.5	1.3	2.0	0.4	1.1	1.7
11	0.5	1.3	2.1	0.3	0.8	1.2
12	3.4	8.4	13.5	1.3	3.3	5.3

Note.

For d prime value, number in the ones place refers to year and number in the tenths place refers to month. For example, d' = 1.2 means that the effect is worth one year and two months of schooling time.

Table 3

National Longitudinal Data-based Norms of Academic Growth in P-12 Reading and Math by Race/Ethnicity: Standardized Achievement Gains per School Year (*g*)

	Reading					Math				
Grades	White $g_{r\text{-}w}$	Black $g_{r\text{-}b}$	Hispanic $g_{r\text{-}h}$	Asian/ Pacific Islander $g_{r\text{-}ap}$	American Indian/ Alaska Native $g_{r\text{-}aa}$	White $g_{m\text{-}w}$	Black $g_{m\text{-}b}$	Hispanic $g_{m\text{-}h}$	Asian/ Pacific Islander $g_{m\text{-}ap}$	American Indian/ Alaska Native $g_{m\text{-}aa}$
PreK	1.00	1.02	1.14	1.28	0.93	1.01	0.97	1.08	1.09	1.00
K	1.67	1.52	1.73	1.80	1.75	1.84	1.45	1.74	1.78	1.96
1	1.82	1.54	1.64	1.90	1.62	1.73	1.43	1.61	1.58	1.35
2	1.27	1.12	1.22	1.14	0.98	1.30	1.10	1.27	1.40	1.23
3	0.84	0.74	0.81	0.75	0.65	0.97	0.82	0.95	1.05	0.92
4	0.55	0.48	0.55	0.52	0.74	0.77	0.70	0.80	0.88	0.83
5	0.51	0.45	0.51	0.48	0.7	0.74	0.67	0.76	0.84	0.80
6	0.35	0.3	0.36	0.38	0.36	0.42	0.51	0.45	0.41	0.42
7	0.27	0.23	0.28	0.29	0.28	0.32	0.38	0.33	0.31	0.31
8	0.2	0.17	0.21	0.22	0.21	0.22	0.26	0.23	0.21	0.21
9	0.27	0.22	0.23	0.29	0.15	0.48	0.39	0.44	0.53	0.37
10	0.41	0.34	0.35	0.44	0.23	0.48	0.39	0.44	0.52	0.37
11	0.35	0.3	0.43	0.54	0.38	0.64	0.62	0.68	0.76	0.64
12	0.06	0.05	0.07	0.09	0.06	0.14	0.14	0.15	0.17	0.14

Table 4
Selected examples of academic intervention effects on reading/math achievement as shown in the units of standard deviation (*d*) and years/months of schooling (*d′*)

Subject	Age/Grade	Treatments (vs. Control Conditions)	Sources	Standardized effect size (*d*)	Time-indexed effect size (*d′*)
Reading	Preschool (4-year old)	Head Start (vs. no Head Start)	Puma et al. (2010)	0.20 for all 0.40 for Blacks (PreK) -0.19 for Whites (PreK)	0.19 for all 0.24 for Blacks (PreK) -0.13 for Whites (PreK)
	Grads K-4	Structured English immersion (vs. transitional bilingual education)	Slavin et al. (2010)	0.54 (K) 0.42 (G1) 0.20 (G2) 0.16 (G3) 0.25 (G4)	0.33 (K) 0.24 (G1) 0.16 (G2) 0.20 (G3) 0.46 (G4)
	Grades K-3	Project STAR small class size (vs. regular size)	Finn et al. (2001)	0.27 (K) 0.20 (G1) 0.17 (G2) 0.21 (G3)	0.16 (K) 0.12 (G1) 0.13 (G2) 0.26 (G3)
	Grade 9	Supplemental literacy class (vs. no supplemental class)	Corrin et al. (2009)	0.08 (G9)	0.31 (G9)
Math	Grade 2	Saxton Math (vs. Scott Foresman-Addison Wesley Math)	Agodini et al. (2010)	0.17 (G2)	0.13 (G2)
	Grades 10-11	UCSMP Math (vs. unspecified secondary math)	Hirschhorn (1993)	0.29 (G10) -0.62 (G11)	0.62 (G10) -0.93 (G11)
	Grades K-3	Project STAR small class size (vs. regular size)	Finn et al. (2001)	0.26 (K) 0.27 (G1) 0.19 (G2) 0.15 (G3)	0.15 (K) 0.16 (G1) 0.15 (G2) 0.16 (G3)

Note.
d′ is obtained by dividing *d* by *g* in corresponding subjects and grades from Table 1 for all students or Table 3 for racial subgroups.

Figure 1

Illustration of a time-indexed effect size for research with pretest-posttest measures of academic achievement (Y) for experimental (E) and control (C) groups

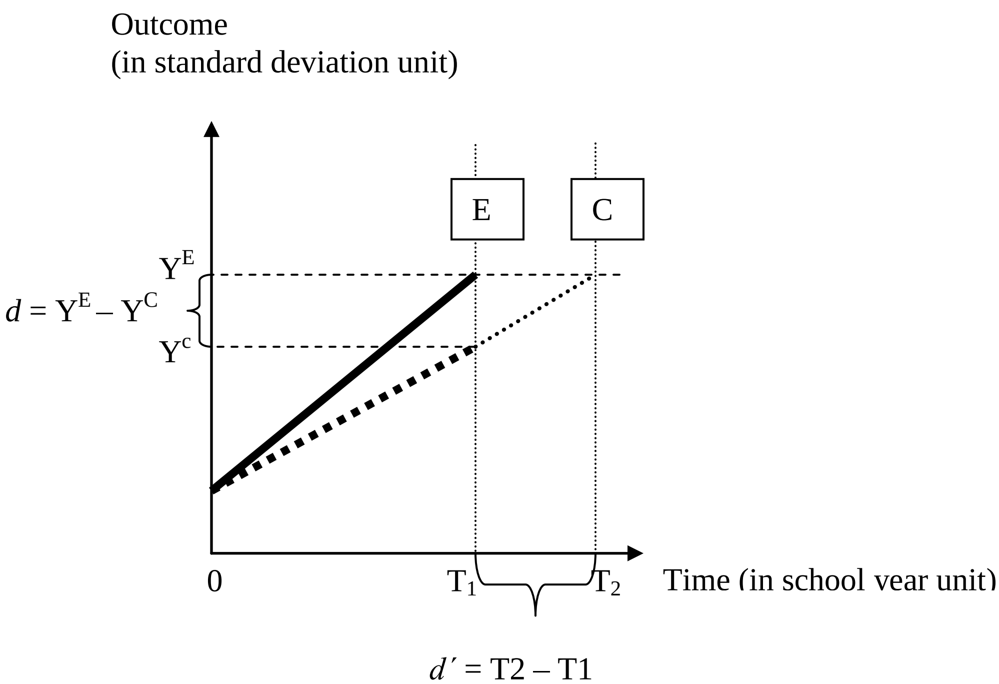
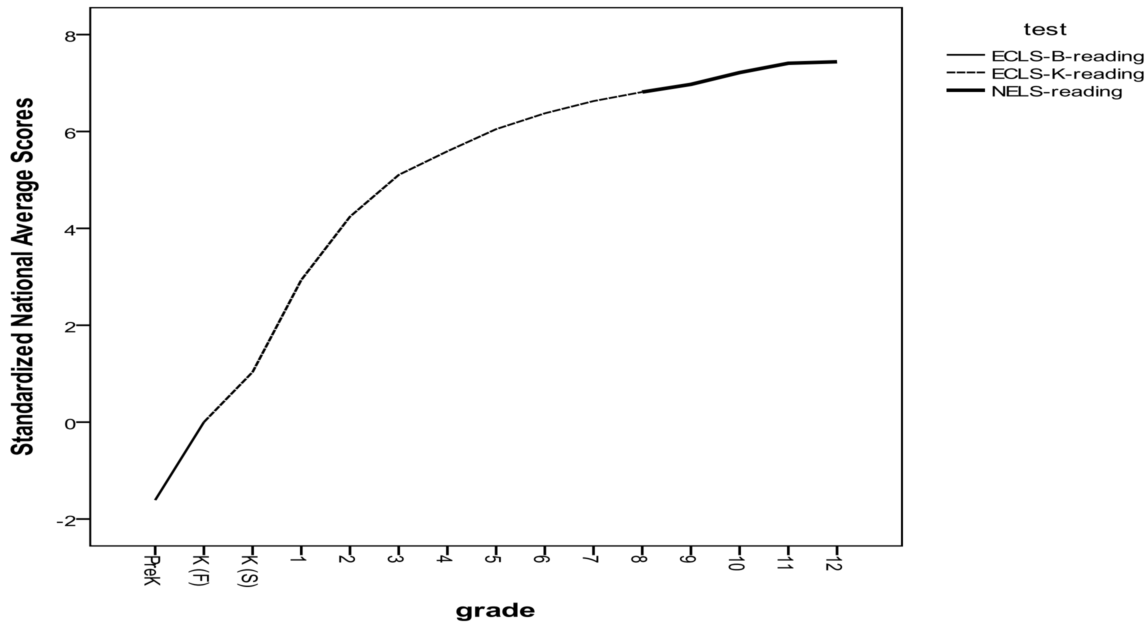


Figure 2

P-12 reading and math national average achievement trajectories (fall K as baseline) based on longitudinal datasets (ECLS-B for PreK-K, ECLS-K for K-8 and NELS for grades 8-12)



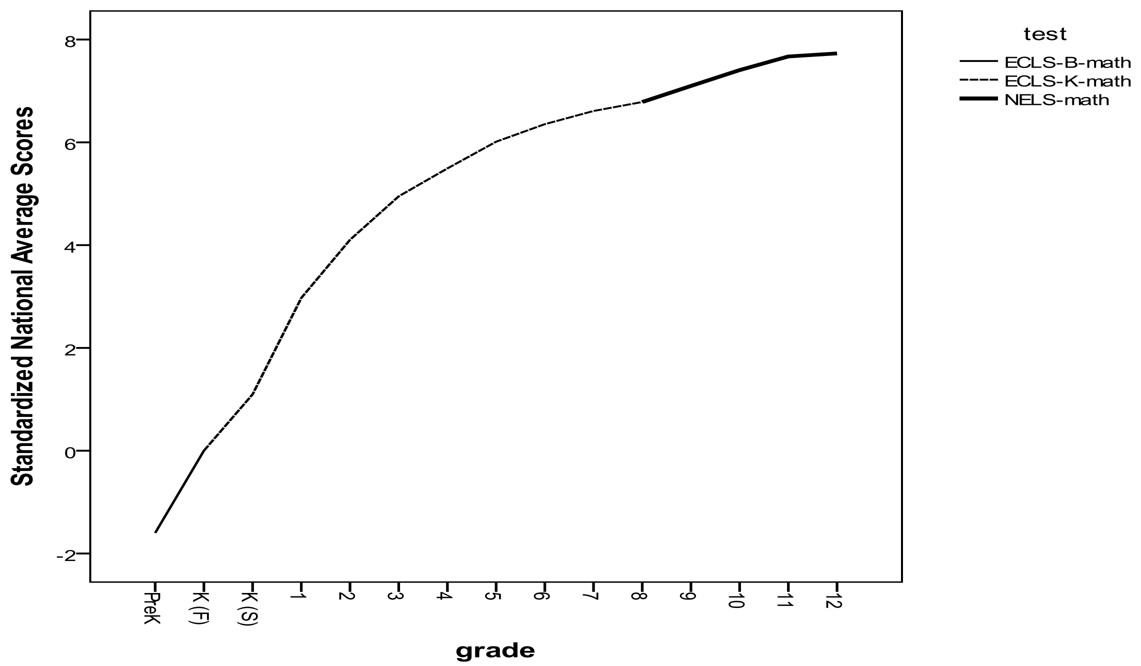


Figure 3

Online time-indexed effect size calculator screen shot from

http://gse.buffalo.edu/faculty/centers/ties/calculator