SREE

2040 Sheridan Rd
Evanston, IL 60208

Society for Research on Educational Effectiveness
Advancing Education Research

www.sree.org
202.495.0920

Abstract Title Page

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

Authors and Affiliations:

Jaekyung Lee
University at Buffalo, SUNY
Jeremy Finn
University at Buffalo, SUNY
Xiaoyan Liu
George Mason University

Abstract Body

Background / Context:

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; 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.

Purpose / Objective / Research Question / Focus of Study:

Extending our prior research on K-12 academic growth trajectories 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?" (see Figure 1). The rationale for time-indexed assessment of effect sizes comes from the well-established pattern of curvilinear 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) and 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. Time-indexed effect size estimation may also provide new insights into post-treatment follow-up evaluation of treatment effects. After treatment termination, a time-indexed effect size may not diminish as much as a conventional effect size if the growth rate of the control group also slows down over the same period.

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. We take an approach to the validation of this new index by employing (1) interpretive arguments (i.e., specification of proposed interpretations and uses of the index) and (2) validity arguments (i.e., evaluation of the interpretive arguments based on evidence) (see Kane, 2006). First, we provide a framework for calculations and interpretations of a time-indexed effect size based on two different designs of educational research/evaluation: pretest-posttest or repeated measures designs and posttest only designs. Second, we present methodological steps for developing longitudinal norms of growth and converting *d* into *d´*. Third, as one element of the supporting validity evidence, we demonstrate how to interpret and use *d´* through applications of the time-indexed effect size metric to well-known research examples. The results of *d* and *d´* for the same studies are compared and cross-validated. Last, we discuss threats to validity, caveats, and ameliorative strategies for valid interpretations and uses of the time-indexed effect size.

Population / Participants / Subjects:

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).

Significance / Novelty of study:

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 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.

Statistical, Measurement, or Econometric Model:

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 (1) 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). 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 norms.

Usefulness / Applicability of Method:

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)[1]. 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). 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.

Conclusions:

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 (Author, 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. In the paper, we will discuss potential threats and technical issues for validating and applying these ideas.

[1] 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´*.

Appendices

Appendix A. References

Author (2011)

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.

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.

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.

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.

Appendix B. Tables and Figures

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		
		d			*d*	
	small	medium	large	small	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

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_l)

	Reading					Math				
Grades	White $gl\text{-}_{w}$	Black $gl\text{-}_{b}$	Hispanic $gl\text{-}_{h}$	Asian/ Pacific Islander $gl\text{-}_{ap}$	American Indian/ Alaska Native $gl\text{-}_{aa}$	White $gl\text{-}_{w}$	Black $gl\text{-}_{b}$	Hispanic $gl\text{-}_{h}$	Asian/ Pacific Islander $gl\text{-}_{ap}$	American Indian/ Alaska Native $gl\text{-}_{aa}$
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

Figure 1

Illustration of a time-indexed effect size (*d′* = T2 – T1) for experimental research with pretest-posttest of achievement (Y) between experimental group (E) and control (C) group

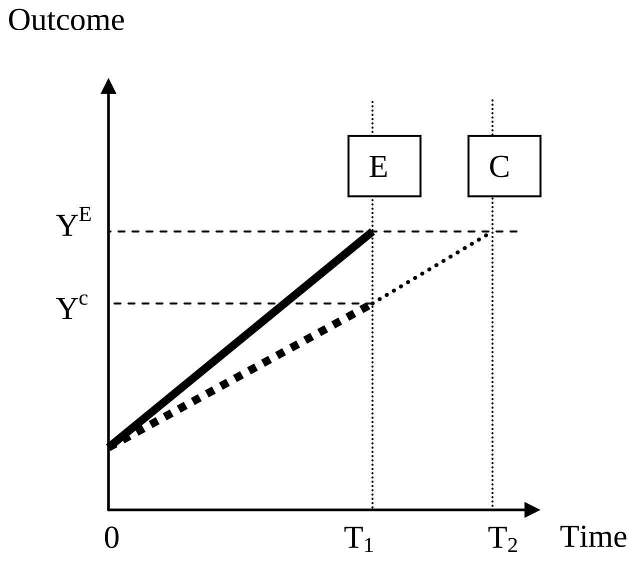
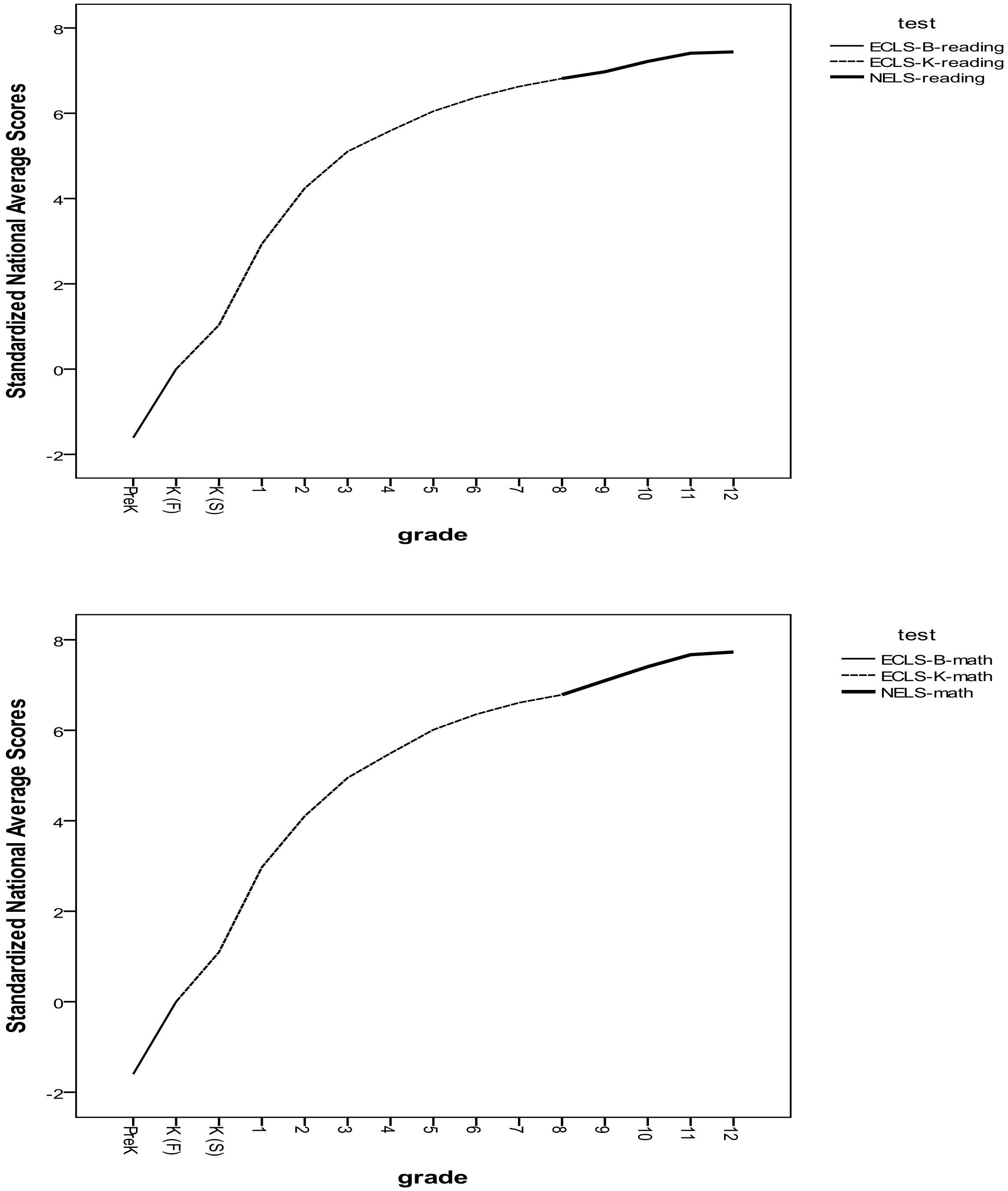


Figure 2

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)


Standardized National Average Scores
PreK
K (F)
K (S)
grade
test
—— ECLS-B-reading
---- ECLS-K-reading
━━ NELS-reading
Standardized National Average Scores
PreK
K (F)
K (S)
grade
test
—— ECLS-B-math
---- ECLS-K-math
━━ NELS-math