

Time-Indexed Effect Size Metric for K-12 Reading and Math Education Evaluation

Jaekyung Lee

Jeremy Finn

Xiaoyan Liu

State University of New York at Buffalo

This paper was presented at the 2011 annual meeting of American Educational Research Association, New Orleans, Louisiana in April, 2011.

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. Project information and online calculator (effect size conversion) are available at our project homepage: <http://gse.buffalo.edu/faculty/centers/ties>. Any correspondence of inquiries and feedback about this paper and online effect size calculator can be sent to the project PI, Dr. Jaekyung Lee by email (JL224@buffalo.edu).

Abstract

Through a synthesis of test publisher norms and national longitudinal datasets, this study provides new national norms of academic growth in K-12 reading and math that can be used to reinterpret conventional effect sizes in time units. We propose d' , 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. It serves as a supplement to conventional effect size metrics such as Cohen’s d by taking into account different amounts of time needed for learning at different age or grade levels. Through applications to Project STAR small class effects and NAEP racial achievement gaps, we demonstrate how to interpret and use d' . It is expected to provide a more developmentally appropriate context for interpreting the size of an effect, a step toward bridging the gap between educational research and practice.

Keywords: effect size, time-indexed effect size, d' , national norms, academic growth

The concept of effect size is ubiquitous within the scholarly research community. While the term effect size can have many operational definitions in scientific research, it is most commonly used to describe standardized measures of an effect's magnitude (e.g., correlation coefficient, Cohen's *d*, odds ratio, etc.). Standardized effect size measures are typically used when the metrics of those variables being studied do not have intrinsic meaning to the reader (e.g., a scale score on an achievement test) or when results from multiple studies using different scales are being considered for meta-analytic synthesis. 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).

Although unstandardized measures such as mean differences can serve as effect size measures, the use of arbitrary scales for measuring student achievement outcomes can have unintended consequences for communicating educational research findings to practitioners who may lack intimate knowledge the meaning of scale scores as well as of standardized effect size measures. One unstandardized effect size measure familiar to educators—years and months of schooling—has not been reported and used in the

literature on research synthesis (Cooper & Hedges, 1994). Instead, Cohen's (1988) rule of thumb is often applied out of context, following the rule that a medium effect ($d = .5$) is conceived as one large enough to be visible to the naked eye and thus important in a practical sense. However, it is still challenging for the lay person and even practitioners to translate a metric representing a standardized group mean difference on a more familiar yardstick such as years/months of schooling.

This work creates a context for determining the extent to which an effect represents a substantial gain in test scores. Prior research has demonstrated the theoretical and practical importance of time for learning (see AERA, 2008; Berliner, 1990; Bloom 1976; Carroll, 1963; Fisher, 1980; National Education Commission on Time and Learning, 1994; Smith et al., 2005). However, as children progress through higher grades in school, their academic growth rates change and thus the time needed to learn any particular topic also changes. National norming research, based on data from standardized K-12 reading and math achievement test publishers, reveals common patterns of academic growth across tests. Overall, there is a general deceleration of achievement growth over the entire course of schooling, even though the patterns of growth differ from subject to subject or from grade to grade (see Beggs & Hieronymus, 1968; CTB/McGraw-Hill, 1997, 2003; Harcourt, 2002, 2004; Lee, 2010; Lichten, 2004; McGrew & Woodcock, 2001).

The present research uses time-varying academic growth, applied to specific studies to address the question: "How much time is needed for students in the control group to catch up with students in the treatment group?" In other words, this raises the counterfactual question: "If students in the treatment group had been assigned to the

control group instead, how much extra time would have it taken for them to reach the level of learning actually achieved at the end of their treatment?” Since we cannot directly observe a treatment group’s outcome under the control condition during an experiment, it would be necessary to create a control group similar to the treatment group through random assignment and/or matching. A randomized trial would allow the researcher to draw inferences about causal effects based on the comparison of two separate outcomes in time units, one with treatment (observable) and the other without treatment (unobservable from the treatment group but estimated from the control group). Matched samples, carefully selected, are intended to approximate the same kind of conclusion.

The rationale for time-indexed assessment of effect sizes also comes from 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. Research has often shown that an effect is not sustainable or decaying after a treatment (e.g., large-scale preschool programs) is over (Barnett, 1995; Lee et al., 1990). However, if one takes into account the possibility that academic growth tends to become slower at the higher ages or grades regardless of treatment status (i.e., general deceleration of growth rate over time), post-treatment reductions in the gap between treatment group and control group may be interpreted differently. 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 K-12 reading and math as evidenced through empirical analyses of U.S. test publisher norms and 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. Then we examine conditions of existing test publisher norms in applying this new index in practice, such as problems with cross-sectional and aggregated data, and explain the rationale for new national norms of academic growth. 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.

Conceptual and Analytical Framework for Time-indexed Effect Size Index
 d' for pretest-posttest or repeated measures designs

In an experimental or quasi-experimental research design using both a pretest and a posttest, academic growth between the two time points is estimated using the control group gain as the basis of a time-indexed effect size calculation. Figure 1 illustrates the concept and measurement of time-indexed effect size based on hypothetical linear patterns of growth for an experimental group (E) and a control group (C). 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. Unlike a conventional effect size measure that focuses on the group difference on the vertical axis (outcome variable), we shift the focus to the horizontal axis (time variable). The time-indexed effect size (d') is the extra time (in school years/months) needed for the control group to reach Y^E , the level of outcome that the experimental group has reached at the end of treatment (see Figure 1):

$$d' = T_2 - T_1 \quad (1)$$

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)

Figure 1 about here

If the growth trajectory during the treatment period is assumed to be linear (at least for each school year for multi-year treatment study), one can estimate a constant

growth rate for each grade and get a single estimate of a treatment effect on the achievement gain per grade. The linear growth model for the control group can be expressed as $Y = b_0 + b_1 (\text{Time})$, where b_0 = the control group's average initial status when Time is equal to zero and b_1 = the control group's average growth rate per time unit (school year/month). In this case, equation (1) for time-indexed effect size d' can be solved by substituting T_1 and T_2 as follows:

$$d' = T_2 - T_1 = \frac{Y^E - b_0}{b_1} - \frac{Y^C - b_0}{b_1} = \frac{Y^E - Y^C}{b_1} \quad (2)$$

This equation can be translated into familiar terms below:

$$d' = \frac{\text{Treatment effect}}{\text{Control group growth rate}} \quad (3)$$

There is a limitation in relying exclusively on a study's own sample data to estimate how long it would take for the control group to attain a particular outcome. The treated group or the control group or both may have growth patterns unlike those of the larger population they represent. If the sample is a convenience sample or drawn from particular schools, it may be affected by local conditions including unique characteristics of the community in which students live or unique features of the schools' curricula and policies and the quality of instruction provided. This is also the case in randomized experiments in which both the experimental and control groups have been subjected to other school-wide or district-wide interventions. A control-group intervention may also be planned. For example, in Tennessee's class-size experiment, Project STAR, the full-size classes (i.e. the control group) were reduced by only several students to allay parents' fears that their children were being penalized by virtue of the experiment being

housed in the same schools (Word et al., 1990). It is important to assess not only the gains of treatment groups relative to the taken-for-granted control group, but also the gains of control or comparison groups relative to meaningful reference scales such as national/state norms. This broader contextual information on academic growth also would help guide efforts for scaling up the intervention beyond particular local study settings and time period.

d' for posttest-only design

When using a posttest-only design, the researcher does not have information on growth during the treatment time period, and thus cannot directly predict how long it would take for the control group to reach the treatment group's outcome Y^E under normal schooling conditions. One may attempt to search for similar prior research with pre-post test or repeated measures design, if available, to estimate typical control group gains under similar study circumstances (e.g., school location, racial and economic composition of study body, etc.). Alternatively, the researcher may attempt to estimate control group gains based on preexisting national norms of academic growth if the achievement test used for norms taps into the same construct and the national standardization sample matches the study's own sample well.

Standardized achievement test norms often use developmental scales that report student performance as grade- or age-equivalents (Kolen, 2006). Grade-equivalent (GE) or age-equivalent scores provide information about how a child's performance compares to that of other children at various grade or age levels. Age-equivalent or grade-equivalent scores can be obtained directly from a test publisher's manual or by fitting a curve of mean or median scale scores to the year and month of schooling in which the

test was taken (Schulz & Nicewander, 1997). Treatment effects are sometimes reported in terms of grade- or age-equivalent units, particularly in research using posttest-only designs (Finn et al., 2001; Gormley et al., 2005). The use of GE and Item Response Theory (IRT) metrics lead to different representations of individual differences in growth trajectories and thus different decisions about the efficacy of educational programs (Seltzer, Frank & Bryk, 1994).¹

Critics of GE metrics pointed out that with most test and scaling designs, a student who scores two years above her grade level on a test designed for her grade would not necessarily score at the average on a form designed for two grades higher because of curriculum-related differences in test content (Peterson, Kolen, & Hoover, 1989). However, reasonable accommodations can be made to developmental grade-level scores (Osterlind, 2006). The procedures that educational testing companies use to cope with the problems associated with conventional GE metrics include IRT-based vertical equating procedures based on sufficient test overlap between adjacent test levels and students from multiple grades who take the tests as a combined norming sample. As long as GE metrics are constructed properly and used for the sake of group comparison within an applicable range of grades, they have the potential to advance more developmentally appropriate evaluation of educational program effects.

While debates about the use of GEs focused largely on the interpretation of individual students' scores, GEs are a useful way to compare the means of several groups at a particular grade level, and can be interpreted in terms familiar to educators—months of schooling (Finn et al., 2001). While there still remain other limitations (e.g., outdated and aggregated national norms based on cross-sectional data), the merits of these

underlying ideas remain valid. 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. Some researchers have 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.

Cross-sectional vs. Longitudinal Data-based Norms of Academic Growth

In this study, we constructed national norms of academic growth for K-12 reading and math achievement through meta-analytic synthesis of existing cross-sectional test publisher norms and existing longitudinal datasets (see Appendix for descriptions of the tests and standardization samples). 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 information from the combination of three separate test publisher norms of reading and math achievement for K-12 students: Stanford Achievement Test (SAT), TerraNova (TN), and Metropolitan Achievement Test (MAT). These all employ IRT vertical scaling methods for equating across grades and provide comparable measures of reading and math achievement across grades within tests as well as between tests.

We also used two national longitudinal datasets, the Early Childhood Longitudinal Study-Kindergarten (ECLS-K) and the National Education Longitudinal Study of 1988 (NELS:88) and to construct our own national norms of academic growth. These two 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-K, launched in 1998, followed academic growth trajectories from Kindergarten to grade 8. The NELS, launched in 1988, tracked individual students' academic growth from grade 8 to grade 12.

Longitudinal analyses of the 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. In order to track reading and math achievement for the "typical" student (i.e. those who spent one year in kindergarten, and who entered grade 1 the following year and grade 3 two years later, etc.), 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. The sample size for the analysis of the ECLS-K data was 5,959. As with the selection criteria for ECLS-K data, 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. Students who were retained in any grade, 8 through 12, who dropped out, or who graduated ahead of their class were excluded. The sample size for the analysis of the NELS:88 data was 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.

Synthesizing National Norms of Academic Growth in Reading and Math

The cross-sectional and longitudinal data revealed that the patterns of academic growth are not always consistent from subject to subject and from grade to grade. Nevertheless, there are some common patterns of growth across tests such as decelerating growth over the course of schooling. Figures 2 and 3 show national average K-12 reading and math achievement trajectories, with cumulative gain scores from fall K through spring grade 12 in standard deviation units, across all five tests. For the sake of illustration, we juxtapose ECLS-K curves (for K-8) and NELS curves (grades 8-12) together to show possible full K-12 range of longitudinal growth trajectories; they are combined by adding NELS 8-12 gains on top of ECLS-K K-8 gains, based on the fact that they both used comparable grade 8 spring reading and math assessments.² It needs to be noted that the values of vertical axis on Figures 2 and 3 include both school year and summer gains across K-12, whereas time-indexed effect size calculations shown later use only the school year portion of the gains in individual grades separately and thus does not involve any statistical linking between ECLS-K and NELS.

There is a high degree of consistency between test publisher norms; all three tests' growth curves are highly similar in both subjects. In contrast, the comparison of test publisher norms with longitudinal growth norms shows discrepancies. It suggests that cross-sectional data may underestimate real gains over time during the elementary grades that longitudinal data are better able to capture. Particularly, the ECLS-K growth curve outpaces test publisher growth curves. During high school (grades 8-12), however, the gains are not substantially different between NELS and other standardized tests. By and large, the gap between the two types of growth norms begins to widen during the early elementary school period with different growth rates and sustains through high school level.

Figures 2 and 3 about here

We combined the multiple sources of data to create our own adjusted national longitudinal norms of academic growth. Based on the new national norms, we built a table of conversion for translating standardized group mean differences (Cohen's d) into years/months of schooling (d') by subject and grade (see Table 1). The process of constructing national longitudinal growth norms followed three stages: (1) reanalysis and synthesis of existing national test publisher norms (SAT, MAT and TN), (2) creation of national longitudinal growth norms (ECLS-K and NELS:88), and (3) synthesis of stage (1) and (2) results to construct adjusted longitudinal growth norms.

Table 1 about here

The first stage involved several steps. First, estimates of the standardized reading and math achievement scores were obtained for each norming sample in the fall and spring of grades K-12. Second, standardized achievement gain scores during the school

year (i.e., between fall and spring assessments in the same grade) and summer (i.e., between spring and fall assessments between adjacent grades) were obtained by computing the mean scale score differences between two time points and dividing them by pooled standard deviations (i.e., pooling two standard deviations from adjacent assessments). The fall-to-spring standardized gain scores were adjusted for the difference in testing time intervals (approximately 6-7 months) to obtain full 10-month equivalent school year gain. The formula for 10-month school year standardized gain, g is as follows:

$$g = \left[\frac{(\bar{Y}_{t+1} - \bar{Y}_t)}{\sqrt{(s^2_{t+1} + s^2_t)/2}} \right] \cdot \left[\frac{10}{\Delta(t)} \right] \quad (4)$$

where \bar{Y}_t = mean of test score at time point t ; s^2_t = variance of test score at time point t ; $\Delta(t)$ = elapsed time in months between two successive rounds of assessments at time t and $t+1$.

Third, average yearly achievement gain scores were calculated by averaging the g values across all three tests for each grade and subject. The averaging of all three tests' growth norms was weighted by their approximate norming sample sizes (weight = .20 for MAT, .34 for TN, and .46 for SAT).³ The end product of this first stage synthesis, g_c , appears as "cross-sectional growth norms" in the column (1) of Table 1.

The second stage created national norms of academic growth based on the analysis of ECLS-K and NELS:88 data. We created standardized measures of reading and math achievement gain scores (in pooled standard deviation units) between successive grades. Because ECLS-K and NELS assessments do not cover all grades, gains were computed only between successive waves of assessments available in the

datasets (i.e., 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 (4) 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.

However, the assumption of linear growth (equal increment to achievement by grade) during any missing grades was not supported by decelerating growth patterns reported in the test publisher norms and prior research. Moreover, the well-known phenomenon of unequal growth between school year and summer break periods (Cooper et al., 1996; Alexander, Entwisle, & Olson, 2001; Heyns, 1978) also made it difficult to estimate gains during school years only based on total gains measured between the spring of two different grades that did not separate school year and summer periods.

To address these limitations, the third stage involved adjusting the longitudinal growth norms based on the first stage results, that is, the distribution of school year (fall to spring) and summer (spring to fall) gain scores for each individual grade from cross-sectional test publisher norms. Except for ECLS-K kindergarten and grade 1 data with both fall and spring assessment measures, we incorporated cross-sectional grade-by-grade norms into longitudinal norms for grades 2-12. Based on a high correlation between cross-sectional and longitudinal gains ($r = .88$ for reading and $r = .78$ for math), it was reasonable to assume that, despite overall discrepancy in the size of total gains,

longitudinal growth norms followed the same distributions of gains as cross-sectional norms in terms of the proportion of growth in any given month.

Specifically, we retained the value of total multi-grade gains obtained from the second stage longitudinal data analysis, but prorated the total gains according to the relative proportion of grade-by-grade increments from the first stage analysis. For example, the second stage analysis of ECLS-K data showed the overall standardized reading achievement gain of 2.17 between spring grade 1 and spring grade 3. The corresponding total gain during the same 2-year period from the first stage analysis of test publisher norms was 1.51, which can be broken down into four subperiods: .05 during 2-month summer before grade 2 (3.2%), .86 during 10-month school year in grade 2 (56.8%), .04 during 2-month summer before grade 3 (2.7%), .57 during 10-month school year in grade 3 (37.3%). We borrowed this information about the percentages of school year gains and divided the ECLS-K grades 1-3 total gain of 2.17 into 1.23 ($g = 56.8\%$ of $2.17 = 1.23$) for grade 2 and .81 ($g = 37.3\%$ of $2.17 = .81$) for grade 3; the rest are 2-month summer gains in each grade not used for the d' calculation. In the same way, we computed percentages and estimated gains in reading and in math for each of the other grades.

The end product of the third stage analysis is g_l , labeled as “longitudinal growth norms” in column (2) of Table 1. These final g values (estimated standardized gains per school year) 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:

$$d' = \frac{d}{g_l} \quad (5)$$

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_t values in Table 1. We followed the same steps to construct the conversion table for demographic subgroups based on their national longitudinal growth norms, adjusted for the weighted average of multiple national test publisher norms.

Table 2 about here

Although the values of d' and d can change in the opposite directions from a lower grade to the upper grade as a result of diminishing growth rate, they should have the same signs at the same grade. Since the value of standardized gain per year is positive for every grade, positive treatment effect would always produce positive value of d' . If the treatment effect is zero, then d' also becomes zero. If the treatment effect is negative, then d' is also negative. A zero value of d' means that there is no gain or loss in time, while any negative value of d' suggests that there is time lost as a result of the treatment.

According to the conversion table for reading, the effect size for a reading program with $d=0.2$ (i.e., 20% of one standard deviation) in Kindergarten would be equivalent to one month of schooling ($d' = 0.1$). 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 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.

An exceptional spurt occurs at grade 12 due to the fact that, according to the publishers' norms, achievement gains between grade 11 and grade 12 dropped substantially and the growth curve becomes almost flat. On the other hand, the 12th grade anomaly seems to be more serious in reading than in math as a result of relatively faster rate of growth in math than in reading. In any case, extra caution is needed when applying our national norms or conversion table to studies that involve grade 12.

We followed the same procedures to explore subgroup differences in academic growth trajectories. Achievement gaps between subgroups (e.g., race and parental education) emerged at the beginning of kindergarten and tended to widen to some extent over the course of schooling. Table 3 illustrates separate estimates of average growth rates for different racial/ethnic groups based on their longitudinal growth trajectories. However, it remains to be examined whether differences between student subgroups warrant different growth norms and how such separate norms might be applied to educational program evaluation.

Table 3 about here

Applications of Time-indexed Effect Sizes

Example of experimental research: Project STAR class size effects

In this section, an application of time-indexed effect size measures to experimental research is demonstrated by using data from Project STAR, the Tennessee Class Size Reduction Study 1985-1988 K-3 data files. Project STAR involved randomized controlled trials of class size reduction with about 6,500 students in K

through 3rd grade from 79 schools in 42 Tennessee school districts; students were randomly assigned to either a smaller class (13 to 17) or a larger class (22 to 26). Standardized tests used for reading and math achievement measures were the Stanford Achievement Tests (SAT; Psychological Corporation, 1985). A previous study using the STAR data found that students attending small classes performed better academically on all achievement tests in each grade compared to students in regular-size classes (Finn et al., 2001). Effect sizes were expressed in grade equivalents using the SAT norms based on a series of cross-sectional norming samples. These indicated that the benefits increased with each additional year a student spent in a small class.

Our analyses reexamined the effects of small class on SAT total reading and total math scores, and compared the results based on two different effect size metrics, d and d' . Our results differ from those of Finn et al. (2001) in several ways. First, the sample used here included only 2,432 students who participated in the study for four consecutive years beginning in kindergarten and remained in the same class type (small or regular). Second, in contrast to grade equivalents, the d' scale is based on longitudinal growth norms that show more rapid growth than do test publishers' norms, and this results in smaller time-indexed effects.

Figures 4 and 5 demonstrate how the effects of small classes in Project STAR changed from K to grade 3 as a result of the choice of different effect size metrics. The upper panel of Figure 4 (reading) and Figure 5 (math) is based on Cohen's d , group mean differences in standard deviation units, showing that small classes were beneficial for both subjects in all grades. In reading for all students, the small-class advantage declined

in grade one and grade two and increased in grade three. In math for all students, the small-class advantage decreased in each subsequent year.

The lower panel of Figure 4 and Figure 5 is based on time-indexed effect size d' , group mean differences in units of school years. These results show that the small class effect in reading remained fairly stable in grades K through two and increased in grade 3. We estimate that it would take students in larger classes about two and half months to catch up to the reading performance of students in small classes in grade three. In math, the results did not deteriorate after kindergarten but remained stable through third grade. In each grade, 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.

Figures 4 and 5 about here

For the lower panel of Figure 4, this study applied our longitudinal growth norms instead of the test publisher's national norms to calculate time-indexed effect sizes. Applying national norms as opposed to local norms to calculate time-indexed effect size could be potentially misleading and biased if there are significant differences between the national and local samples in their social demographic and educational profiles which can influence academic growth curves.

Project STAR researchers did not examine control group's gain relative to the national test publisher norms, but this study compared some characteristics of the STAR sample to national figures. The STAR sample had different demographic profiles from both test publisher standardization sample and our longitudinal sample.⁴ It appears that the average student in the STAR sample had socioeconomic and academic disadvantages relative to the average student in the national population. This raises a question about the

assumption that the control group in the STAR sample would have had the same reading and math growth trajectories as students across the U.S.

According to several analyses of the STAR data (Finn & Achilles, 1989; Goldstein & Blatchford, 1998; Hedges, Nye, & Konstantopoulos, 2000), minorities or students from low-income homes benefitted more from attending small classes than did students who were White or from higher-SES homes. Thus we computed time-indexed effect sizes for Black and White students using our separate growth norms for these groups (see Table 3). Those subgroup results are shown in Figures 4 and 5. For White students, the patterns of d were uneven across the grades in both subjects. The time-indexed effect sizes d' was somewhat more even across the grades, especially for mathematics. All effect-size measures except one were larger for Black students than for white students. The difference between d and d' was more evident for Black students than for White students. Cohen's effect-size measure in reading was stable from kindergarten through second grade and then increased in grade 3. The time-indexed measure, however, indicated that more time was needed to catch up in grade two than in grade one, and more in grade three than in grade one—indeed, almost half of a school year (5 months). In math, d increased from kindergarten to grade one and then decreased in each subsequent year. The time needed to catch up (d'), however, increased monotonically with each additional year. Black students benefit from early grade (K-3) small classes, up to a maximum of 4 months. In brief, we note that (1) the d s for Black students are higher than for White students and the d' even more so, and that (2) even though the d s for Black students are relatively flat over the four grades, the d' s increase

considerably; in the same vein, even though the d s for White students decline over the four grades, the d 's remain more constant and stable.

Example of nonexperimental research: NAEP racial achievement gaps

Achievement gaps constitute important barometers in educational and social progress. The National Assessment of Educational Progress (NAEP), the so-called nation's report card of student achievement, provides information on the achievement gaps among different racial and socioeconomic groups in core academic subjects. Despite the advantage of providing national snapshots, NAEP is cross-sectional, and thus does not allow us to track changes in the size of gap for the same cohort of students. Using longitudinal data sets, prior research on Black-White achievement gap showed that the racial gap in reading and math emerges prior to school entry and widens over the course of schooling (Fish, Lee, & Chilungu, 2007; Fryer & Levitt, 2004; Philips, Crouse, & Ralph, 1998). However, the size of change in racial achievement gap has been often reported and interpreted in terms of standardized group mean differences so that information on its practical significance was not clear and straightforward. Our understanding of the widening Black-White achievement gap phenomena can be enriched with uses of time-indexed effect size metric based on national longitudinal education datasets which provide information on academic growth in core subjects.

While the gap between White and Black student groups persist across grades, it is clear that the racial achievement emerges prior to school entry and widens to some extent over the course of early elementary education. The absolute size of the Black-White achievement gap (as measured by IRT scale score differences) appears to widen slowly, but this assessment would underestimate the practical significance of the widening gap.

How long would it take for Black students to catch up to the current performance level of their White peers? If we evaluate the achievement gap from the viewpoint of “time needed to catch up” (based on “Black” annual growth rate g_{l-b}), it becomes clear that the gap widens at a more rapid pace making it harder and harder to narrow over time.

Table 4 shows contrast of changes in Black-White reading and math achievement gaps in the original NAEP scale score units, standard deviation units (d) and school year/month units (d'). The Black-White gap in d remains largely constant from grade 4 through grade 12 in both subjects. In contrast, the gap in d' increases about 9 times in reading and 5 times in math during the same schooling period. The exponential increase of time-indexed achievement gap is attributable to the rapid deceleration of academic growth rates at the upper grades. In reading, the Black-White achievement gap may be equivalent to one and half years of schooling ($d'=1.48$) at grade 4 when Black students' academic achievement grows fast ($g_{l-b}=.48$), and the gap would enlarge to 4 years and 6 months ($d'=4.65$) at grade 8 when their growth rate slows down to .17. Then it becomes thirteen and half years ($d'=13.6$) by the end of grade 12 when Black students' academic growth almost stalls ($g_{l-b}=.05$). In math, the Black-White gap in d' increases from 1 year 3 months to 3 years and 4 months and then to 6 years and 3 months between grades 4, 8 and 12.

Table 4 about here

The 12th grade Black-White gaps in school year units are incredibly large and they represent extra amount of time needed for average Black 12th graders to reach the current achievement level of average White 12th graders. This estimate is based on a hypothetical situation where Black 12th graders would continue to learn content in the same grade and

maintain the same growth rate. The achievement gap may reflect corresponding content gap in terms of advanced course-taking. For example, according to the 1999 NAEP survey of 17-year-olds, about two-thirds of White students had taken algebra II or precalculus/calculus, whereas only 56 percent of Black students had done so (Campbell, Hombo, & Mazzeo, 2000). Nevertheless, the time-indexed measure of Black-White achievement gaps may have been exaggerated if Black students' true academic growth rates were underestimated due to possible deterioration of test-taking motivation at the end of high school. The validity of interpretations remains questionable, since they may not be meaningful in light of the entire growth trajectory; the 12th grade growth rate of being near zero is an unexpected deviation from slow but steady pattern of achievement gains observed during the lower grades in high school.

Validation and Limitations of Time-indexed Effect Size

The index, d' , is proposed as a supplement rather than an alternative to d , the usual effect size index for comparing two groups. Essentially, the index estimates the additional time that the control group would need to reach the attainment of the treatment group. The validation of the proposed time-indexed effect size requires evaluation of its assumptions, including implicit or hidden assumptions (Kane, 2006).

We proposed to interpret an effect size that communicates the time in years and months for the control group to catch up to the treatment group. The time-indexed effect size assumes (a) linear growth by the control group over time at the rate estimated for that particular grade and (b) no movement by the treatment group. Both assumptions (a) and (b) may not reflect realities and there may be more plausible alternatives. Some practitioners might plausibly interpret the "catching up" scenario to involve alternatives to

(a) for example, that control groups would suddenly be given the treatment or otherwise alternative trajectory) or (b) for example, that the treatment group continues to grow while the control is catching up. The assumption (a) does not mean that the control group continues to have the same growth rate in subsequent grades. Without any special intervention beyond regular schooling, both cross-sectional and longitudinal growth curves show that sustained linear growth is clearly not the case. Since the growth trajectory changes over time, it is difficult to predict the future changes beyond the current grade in which a particular study collected data. The assumption (b) does not suggest that the treatment group's outcome observed at the end of treatment remains constant. Without invoking any ungrounded predictions of possible future changes beyond a study period, our index is intended to give current estimate of the learning gap in time based on actual observed growth rate at the same grade in which the test score gap occurred. We acknowledge that this is not the only way that a time-indexed effect size could be constructed, and caution should be exercised in the interpretation of d' .

The use of a time-indexed effect sizes requires that measures of academic achievement be on common scales that allow the researcher to compare an individual's or a group's performance at different time points. It is difficult to find interval scale measures that are equally applicable, reliable, and valid in children of various ages and that are known to measure the same construct at different ages (Baltes, Reese, & Nesselroade, 1977; Bergman, Eklund, & Magnusson, 1991; McCaffrey et al., 2003; Peterson, Kolen, & Hoover, 1989; Schaie, 1965). A fundamental premise of vertical scaling is measurement equivalence based on sufficient continuity of curriculum and assessment across grades K-12 in reading and math that warrants a common scale in each

subject. Developmental scales of child achievement had been created using Thurstone scaling, but a major advance was made later through test equating based on item response theory modeling (see Kolen & Brennan, 2004; Lord, 1980; Wright & Stone, 1979). IRT was used for all five tests used for our national norms so that the items measuring performance in a particular content domain can be placed on a common scale of difficulty, and thus, all scores can be placed on a common achievement scale. The validation of a vertical scale requires that curricula be compared across grades to justify the exchangeability of tests from one grade to another. The technical reports of all five tests provide adequate information on test design and supporting evidence for cross-grade vertical scales.⁵

Even with cross-grade vertical scales, a problem would occur if one is attempting to assess the degree of learning gap or effect size in years/months of schooling by making reference to a future grade level. We emphasize the pitfalls of this type of misinterpretation. For example, the time-indexed effect size of “two years” ($d'=2.0$) does not mean that the treatment group performs at the same level as the average student of two grades above the control group’s current grade level on a test designed for two grades higher. Rather, the “two-year” effect size should be interpreted as that, given the current rate of growth, it would take about two years worth of schooling time for the control group to reach the same end-of-treatment performance level of the treatment group on a test suitable for their current grade.

We also caution that applying national norms to a local study is prone to possible misuse and misinterpretation. Norms are not standards. In the past, arguments have been forwarded that all students should perform at or above the national norm (e.g., all sixth

graders reading at or above the sixth grade equivalent). This kind of inconsistency could be avoided if we switch the focus of evaluation from status to growth. Nevertheless, application of national norm to a local experimental study is based on the assumption that the control group would grow at the same rate as the national norm. 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. Some adaptation of national growth norms may be needed based on sample differences (e.g., demographics) between the study and national norming sample.

Lastly, it would have been more desirable to use longitudinal data from the same cohort for the entire period of K-12, but there is no such data yet. ECLS-K is available only for grades K-8, whereas NELS:88 covers only grades 8-12. We created K-12 growth curves and national norms in reading and math achievement by combining K-8 norms derived from ECLS-K and grades 8-12 norms derived from NELS:88. Because these two datasets were collected from different cohorts at different time periods, the time gap between the two datasets may pose potential threat to validity of using them together if there were significant change during the interim period. The growth trajectories may have changed among different cohorts over the long run. The Educational Longitudinal Study 2002 (ELS: 2002) has more recent information yet the data are limited to an assessment of grades 10 and 12 in math ($N = 12,652$ students). The comparison of NELS (1990-1992) and ELS (2002-04) growth norms in terms of grades 10-12 math standardized

gains did not detect significant differences between the two cohort groups. Nevertheless, potential changes in national academic growth norms, including varying degree of progress at different grade levels, remains an issue. A cross-cohort comparison of NAEP reading and math achievement gains over the past few decades revealed a tripartite pattern where American students have been gaining ground at the pre/early primary school level, holding ground at the middle school level, and losing ground at the high school level (Lee, 2010). If NCES continues to collect comparable longitudinal data for subsequent cohorts, it is feasible to update our national norms of academic growth.

Summary and Conclusion

Our current capacity to understand or provide a context for interpreting the size of an effect is limited. While estimating treatment effects is a technical issue, interpreting the size of an effect is a judgment. Despite the de facto requirement of effect size reporting for practical significance in publication, context-free and mechanical effect size reporting practices cannot help advance our informed judgment about educational program effects without the understanding of developmental context. Conventional effect size metrics such as Cohen's d are standardized group mean differences based on the distribution of a student outcome variable at one particular time point. These measures do not take into account the aforementioned time dimension—varying length of time needed to learn at different age or grade levels. This article 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.

The phenomenon of diminishing rate of growth in reading and math achievement has been observed in both cross-sectional and longitudinal national data (see Figures 2

and 3). If the achievement gap is declining in content but growing in the time it takes to learn that content, which is the accurate characterization of change in the gap? Is the gap widening or narrowing? We claim that the answer can be both and make it a central argument for the use of time-indexed effect sizes. This approach requires assessing the gap on this national growth curve not only from its vertical axis viewpoint (i.e., content knowledge/skills measured in standard score units) but also from its horizontal axis viewpoint (i.e., time measured in school year units).

There are challenges and issues for validating and applying this idea to actual research. Time-indexed effect size requires information on typical academic growth in the control group (experimental research) or the reference group (non-experimental comparative research). For research designs using a posttest only, researchers cannot directly assess achievement gains but may utilize information from existing test publisher norms that provide conventional age- or grade-equivalent metrics. However, the test publishers' norms are based on cross-sectional data of different cohort groups at a single year. Moreover, the assumption that the study sample would grow at the same rate as the national norms could be erroneous. Therefore, this application is highly prone to errors and thus should be accompanied with a strong caveat in order to guard against possible misuse and misinterpretation of the national norms. In this case, it is desirable to use our longitudinal growth norms that can provide more valid reference of the national average growth trajectory for all and subgroups. The findings on the variability of academic growth curves among different subgroups of students and schools challenge one-size-fits-all approach based on conventional GEs.

Meanwhile, research with pretest-posttest or repeated measures design affords information on achievement gains in the sample so that the researcher can directly estimate the control group or reference group's typical growth rates. However, even when local norms can be created with repeated measures, the researchers can still benefit from comparing their local norms with comparable national norms in corresponding subjects and grades. It would be useful to examine whether the control group or reference group's growth is typical or abnormal in comparison with the national norm group's growth. When the local control or reference group's growth turns out to fall significantly above or below the national norms, the reporting and interpretation of treatment effect size or group difference can be revisited and enriched with reference to the national population.

Applications of the time-indexed effect size d' to the two examples of prior research provide new insights and raise new questions about the findings. For Project STAR, it appears that the effect of small class size on academic achievement does not diminish at the upper grades as much as prior research indicated. For Black-White achievement gaps, it turns out that the size of gaps widens much faster over the course of schooling than prior research suggested. These findings help researchers and evaluators become more aware of potential biases and limitations in relying on any single metric for strength of effect measures. Through continuing validation, the proposed effect size metric presented in school time frame might be a step toward bridging the gap between educational research and practice and allowing researchers to communicate their findings with educators in more meaningful ways.

References

- Alexander, K. L., Entwisle, D. R., & Olson, L. S. (2001). Schools, achievement, and inequality: A seasonal perspective. *Educational Evaluation and Policy Analysis*, 23, 171-191.
- American Educational Research Association (2008). Time to learn. *Research Points*, 5(2), 1-4.
- American Psychological Association. (2001). *Publication manual of the American Psychological Association* (5th ed.). Washington, DC: Author.
- Baltes, P. B., Reese, H. W., & Nesselroade, J. R. (1977). *Life-span developmental psychology: Introduction to research methods*. Monterey, CA: Brooks/Cole.
- Barnett, W. S. (1995). Long-term effects on cognitive development and school success. *The Future of Children*, 5(3), 25-50.
- 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.
- Bergman, L, Eklund, G, & Magnusson, D. (1991). Problems and methods in longitudinal research: stability and change. In Magnusson D, Bergman L, Kudinger G, et al, (Eds), *Studying individual development: problems and methods*. New York: Cambridge University Press.
- Berliner, D.C. (1990). *The Nature of Time in Schools: Theoretical Concepts, Practitioner Perceptions*. New York: Teachers College Press.
- Bloom, B. S. (1964). *Stability and Change in Human Characteristics*. New York, John Wiley & Sons.

- Bloom, B. S. (1976). *Human characteristics and school learning*. New York: McGraw-Hill.
- 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.
- Campbell, J. R., Hombo, C.M., & Mazzeo, J. (2000). *NAEP 1999 trends in academic progress: Three decades of student performance*. Washington, DC: OERI, U.S. Department of Education.
- Carroll, J.B. (1963). "A Model of School Learning," *Teachers College Record, 64*(8), 723–733.
- Cohen, J. (1994). The earth is round ($p < .05$). *American Psychologist, 49*, 997-1003.
- Cohen, J. (1988). *Statistical power analysis for the behavioral sciences (2nd ed.)*. Hillsdale, NJ: Erlbaum.
- Cooper, H., and Hedges, L. V. (1994) (Eds.). *The handbook of research synthesis*. New York: Russell Sage Foundation.
- Cooper, H., Nye, B., Charlton, K., Lindsay, J., & Greathouse, S. (1996). The effects of summer vacation on achievement test scores: A narrative and meta-analytic review. *Review of Educational Research, 66*, 227-268.
- CTB/McGraw-Hill (1997). *Technical Bulletin 1*. Monterey, CA: Author.
- CTB/McGraw-Hill. (2003). *TerraNova 2nd Edition CAT Technical Report*. Monterey, CA: Author.
- Effect size (2008, March 22). In *Wikipedia, the free encyclopedia*. Retrieved April, 10 2008 from http://en.wikipedia.org/wiki/Effect_size.

- 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.
- Fish, R. M., Lee, J., & Chilungu, E. N. (2007). Tracking Racial Achievement Gaps from K through 12: White, Black, Hispanic and Asian Trajectories of Reading and Math Achievement. (pp. 1- 20). In Lee, J. (Ed.). *How national data help tackle the achievement gap*. Buffalo, New York: SUNY Buffalo GSE Publications.
- Fisher, C.W., et al. (1980). "Teaching Behaviors, Academic Learning Time, and Student Achievement: An Overview." In C. Dehman, A. Lieberman (Eds.), *Time to Learn* (pp. 7–22). Washington, DC: National Institute of Education.
- Fryer, R.G., & Levitt, S.D. (2004). Understanding the black-white test score gap in the first two years of schooling. *The Review of Economics and Statistics*, 86(2), 447-464.
- Goldstein, H., & Blatchford, P. (1998). Class size and educational achievement: A review of methodology with particular reference to study design. *British Educational Research Journal*, 24, 255-268.
- 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 (2002). *Metropolitan 8 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.

- Hedges, L. V., Nye, B., & Konstantopoulos, S. (2000). The effects of small classes on achievement: The results of the Tennessee class-size experiment. *American Educational Research Journal*, *37*, 123-151.
- Heyns, B. (1978). *Summer learning and the effects of schooling*. New York: Academic Press.
- 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.
- Kolen, M. J. & Brennan, R. L. (2004). *Test Equating, Scaling, and Linking: Methods and Practices* (2nd ed.). New York: Springer-Verlag.
- Kolen, M. J. (2006). Scaling and norming. In R. L. Brennan (Ed.), *Educational Measurement* (4th ed., pp. 155-186). Westport, CT: Praeger Publishers.
- 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, V. E., Brooks-Gunn, J., Schnur, E., & Liaw, F. R. (1990). Are Head Start effects sustained? A longitudinal follow-up comparison of disadvantaged children attending Head Start, no preschool, and other preschool programs. *Child Development*, *61*, 495-507.
- Lichten, W. (2004). On the Law of Intelligence. *Developmental Review*, *24*(3), 252-288.

- Lord, F. M. (1980). *Applications of item response theory to practical testing problems*. Hillsdale, N. J.: Erlbaum.
- McCaffrey, D.F., Lockwood, J.R., Koretz, D.M., & Hamilton, L.S. (2003). *Evaluating Value-Added Models for Teacher Accountability*. Santa Monica, CA: Rand.
- McGrew, K. S. & Woodcock, R. W. (2001). *Woodcock-Johnson III Technical Manual*. Itasca, IL: Riverside Publishing.
- Najarian, M. Pollack, J.M., & Sorongon, A.G. (2009). *Early Childhood Longitudinal Study, Kindergarten Class of 1998–99 (ECLS-K), Psychometric Report for the Eighth Grade* (NCES 2009–002). Washington, DC: National Center for Education Statistics.
- National Education Commission on Time and Learning. (1994). *Prisoners of Time: Report of the National Education Commission on Time and Learning*. Washington, DC: Author.
- 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.
- Phillips, M., Crouse, J., & Ralph, J. (1998). Does the black-white test score gap widen after children enter school? In C. Jencks & M. Phillips (Eds.), *The black/white test score gap*. Washington, DC: Brookings.
- Pollack, J., Atkins-Burnett, S., Rock, D. and Weiss, M. (2005). *Early Childhood Longitudinal Study, Kindergarten Class of 1998–99 (ECLS-K), Psychometric Report for the Third Grade* (NCES 2005–062). Washington, DC: National Center for Education Statistics.

- Psychological Corporation. (1985). *Stanford Achievement Test Series Technical Data Report*.
- Ramey, C.T., & Ramey, S.L (1998). Early intervention and early experience. *American Psychologist*, 53(2), 109-120
- Raudenbush, S.W., & Bryk, A.S. (2002). *Hierarchical Linear Models: Applications and Data Analysis*. 2nd Ed. Newbury Park, CA: Sage.
- Rock, D. A. & Quinn, P. (1995). *Psychometric report for the NELS:88 base year through second follow-up (NCES-95-382)*. Washington, DC: National Center for Education Statistics.
- Schaie, K.W. (1965). A general model for the study of developmental change. *Psych. Bulletin*, 64, 92-107.
- 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.
- Singer, J.D. & J.B. Willett. (2003). *Applied Longitudinal Data Analysis: Modeling Change and Event Occurrence*. New York: Oxford University Press.

- Smith, B., et al. (2005). "Extended Learning Time and Student Accountability: Assessing Outcomes and Options for Elementary and Middle Grades," *Educational Administration Quarterly*, 41(2), 195–236.
- 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*, 25(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>]
- Word, E., Johnson, J., Bain, H. P., Fulton, D. B., Boyd-Zaharias, J., Lintz, M. N., Achilles, C. M., Folger, J., & Breda, C. (1990). *Student/teacher achievement ratio (STAR): Tennessee's K-3 class size study*. Nashville, TN: Tennessee State Department of Education.
- Wright, B. D., & Stone, M. H. (1979). *Best Test Design: Rasch Measurement*. University of Chicago: MESA Press.
- Yen, W.M., (1986). The choice of scale for educational measurement: An IRT perspective. *Journal of Educational Measurement*, 23, 299–325.

Appendix.

Descriptions of test data used for K-12 reading and math achievement growth norms

	MAT	TN	SAT	ECLS-K	NELS:88
Norming Sample	Norming for MAT 8 th edition was based on a stratified nationally representative sample in 1999-2000. N \approx 80,000 for both spring and fall across grades K-12	Norming for TN 2 nd edition was based on a stratified nationally representative sample in 1999-2000. N = 149,798 for spring and N=114,312 for fall across grades K-12	Norming for SAT 10 th edition was based on a stratified nationally representative sample in 2002. N \approx 250,000 for spring and N \approx 110,000 for fall across grades K-12	Norming was based on a stratified nationally-representative sample of Kindergartners in Fall 1998 with follow-through (spring K, grades 1, 3, 5, 8). N = 5,959 students.	Norming was based on a stratified nationally-representative sample of 8th graders in Spring 1988 with follow-through (grades 8, 10 and 12) N = 10,879 students.
Test Measures and Reliabilities	Total reading includes reading vocabulary and reading comprehension; Total math includes concepts & problem solving and computation; reliability ranges .93-.97 in reading and .91-.94 in math	Reading composite is the average of reading and vocabulary; Math composite is the average of math and math computation; Reliability ranges .88-.95 in reading and in math	Total reading includes reading vocabulary and reading comprehension; Total math includes math problem solving and procedures; reliability ranges .93-.97 in reading and .82-.95 in math	Reading composite covers basic reading skills, vocabulary, and reading comprehension skills; Math composite includes number operations, measurement; geometry, algebra etc.; reliability ranges .93-.97 in reading and .92-.95 in math	Reading composite covers vocabulary and reading comprehension skills; Math composite includes algebra, geometry, and advanced topics; reliability ranges .80-.85 in reading and ranges .89-.94 in math

Note. The test information and data sources are as follows:

MAT: Harcourt (2002). *Metropolitan8 Form V Technical Manual*. San Antonio, TX:

Author. Appendix J Table J-1 and J-2 (Scaled score summary data by grade).

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

CA: Author. Table 59 (scale score descriptive statistics form c fall and spring reading composite and math composite scores)

SAT: Harcourt (2004). *Stanford Achievement Test 10th Edition Technical Data Report*.

San Antonio, TX: Author. Appendix K Table K-2 to K-29 (Mean scale scores and related summary data for grades K-12)

ECLS-K: National Center for Education Statistics (NCES). Pollack et al. (2005);

Najararian et al. (2009).

NELS:88: National Center for Education Statistics (NCES). Rock & Quinn (1995)

Table 1

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

grades	Reading		Math	
	(1) Cross-sectional Growth Norms g_c	(2) Longitudinal Growth Norms g_l	(1) Cross-sectional Growth Norms g_c	(2) Longitudinal Growth Norms g_l
	K	1.87	1.66	1.24
1	1.34	1.76	0.94	1.66
2	0.86	1.23	1.28	1.27
3	0.57	0.81	0.96	0.95
4	0.36	0.54	0.70	0.77
5	0.34	0.50	0.68	0.73
6	0.35	0.35	0.60	0.44
7	0.27	0.27	0.45	0.33
8	0.20	0.20	0.30	0.22
9	0.29	0.26	0.41	0.47
10	0.45	0.40	0.41	0.47
11	0.27	0.37	0.29	0.67
12	0.04	0.06	0.06	0.15

Table 2

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

grades	Reading			Math		
	d			d		
	small 0.2	medium 0.5	large 0.8	small 0.2	medium 0.5	large 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 K-12 Reading and

Math by Race/Ethnicity: Standardized Achievement Gains per School Year (*g*)

Grades	Reading					Math				
	White <i>gl-w</i>	Black <i>gl-b</i>	Hispanic <i>gl-h</i>	Asian/ Pacific Islander <i>gl-ap</i>	American Indian/ Alaska Native <i>gl-aa</i>	White <i>gl-w</i>	Black <i>gl-b</i>	Hispanic <i>gl-h</i>	Asian/ Pacific Islander <i>gl-ap</i>	American Indian/ Alaska Native <i>gl-aa</i>
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

National average White-Black achievement gaps based on NAEP reading and math assessments in the units of standard deviation (d) and years/months of schooling (d')

Subject	Grade	Scale score gap	Standard deviation	Standardized gap (d)	Standardized gain per year for Black (g_{l-b})	Time-indexed gap (d')
Reading						
	4	25	35	0.71	.48	1.48
	8	27	34	0.79	.17	4.65
	12	26	38	0.68	.05	13.6
Math						
	4	26	29	0.90	.70	1.29
	8	32	36	0.89	.26	3.42
	12	30	34	0.88	.14	6.29

Note.

The above NAEP assessments were administered in 2009 for grades 4 and 8 and in 2005 for grade 12. d is obtained by dividing scale score gaps by standard deviations, and d' is obtained by dividing d by g_{l-b} .

Figure 1

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

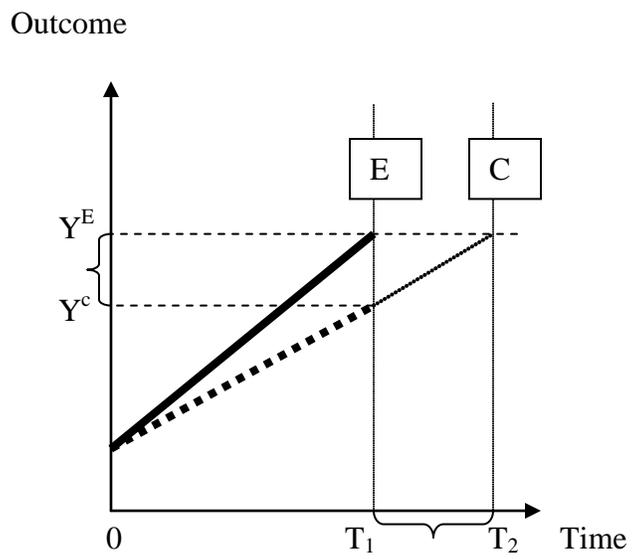


Figure 2

K-12 reading national average achievement trajectories based on longitudinal datasets

(ECLS-K and NELS) and cross-sectional test publisher norms (MAT, SAT, TN)

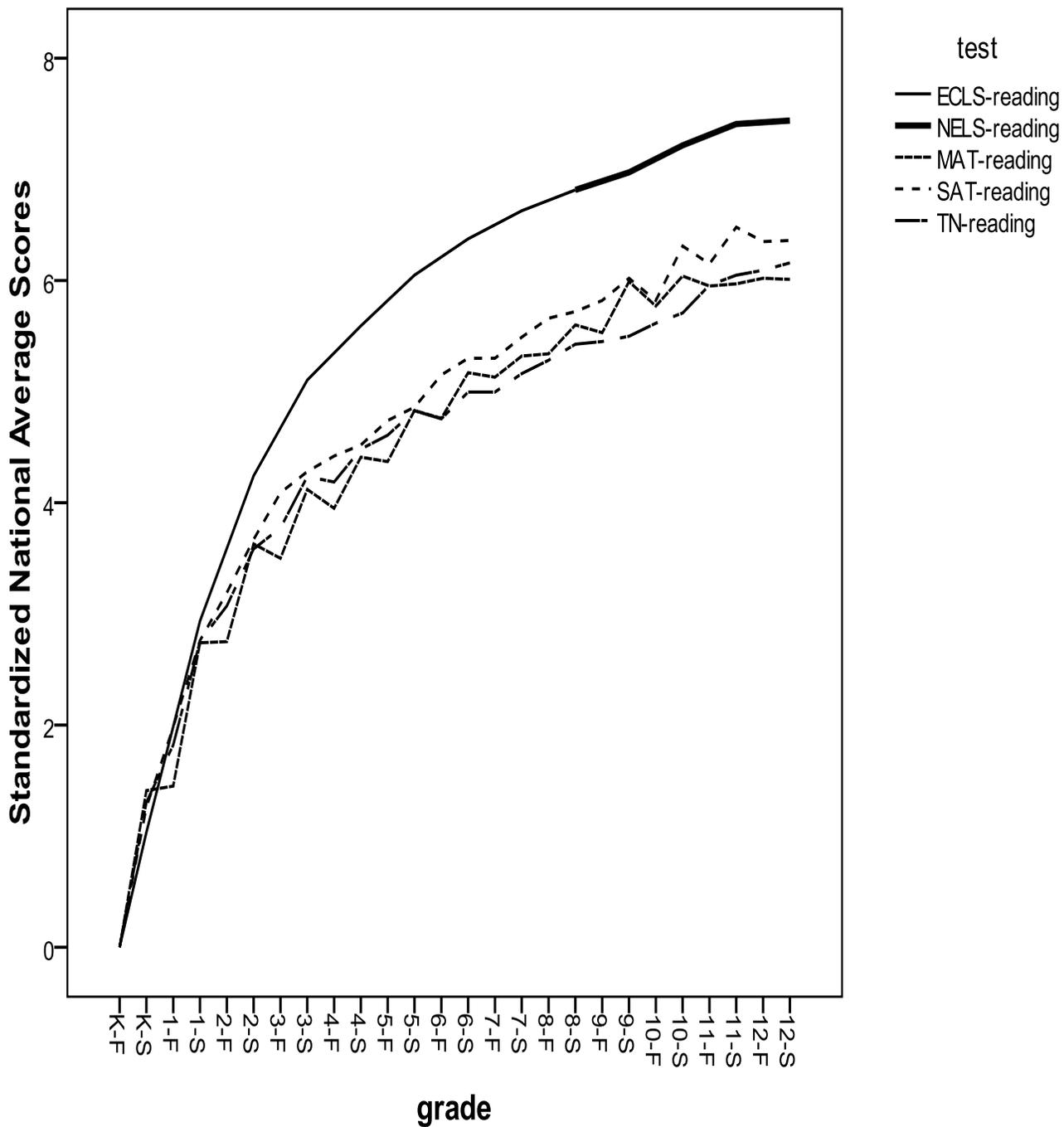


Figure 3

K-12 math national average achievement trajectories based on longitudinal datasets

(ECLS-K and NELS) and cross-sectional test publisher norms (MAT, SAT, TN)

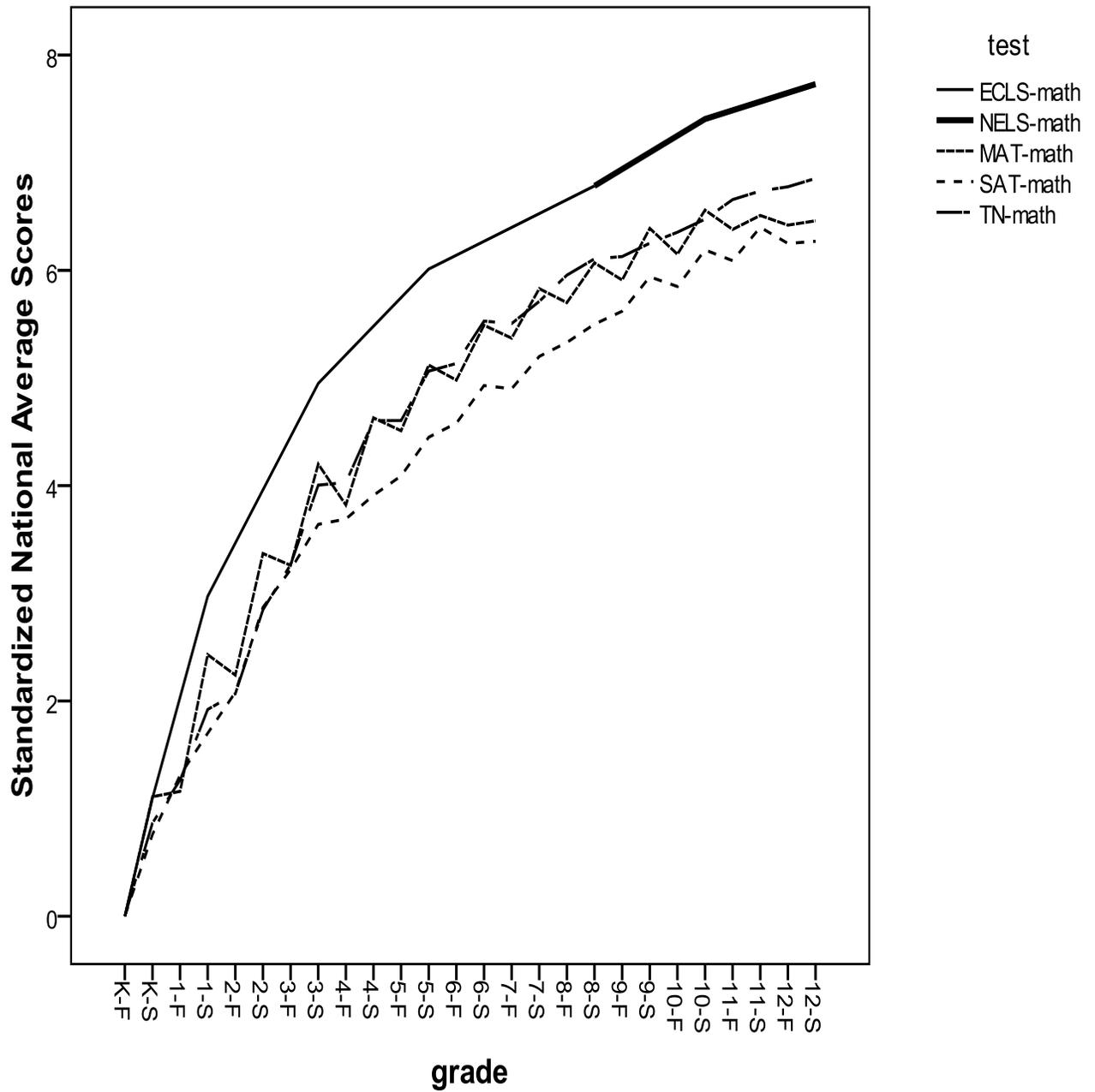


Figure 4

Project STAR small class effects in K-3 reading based on d (standard deviation units) in the upper panel and d' (school year/month units) in the lower panel

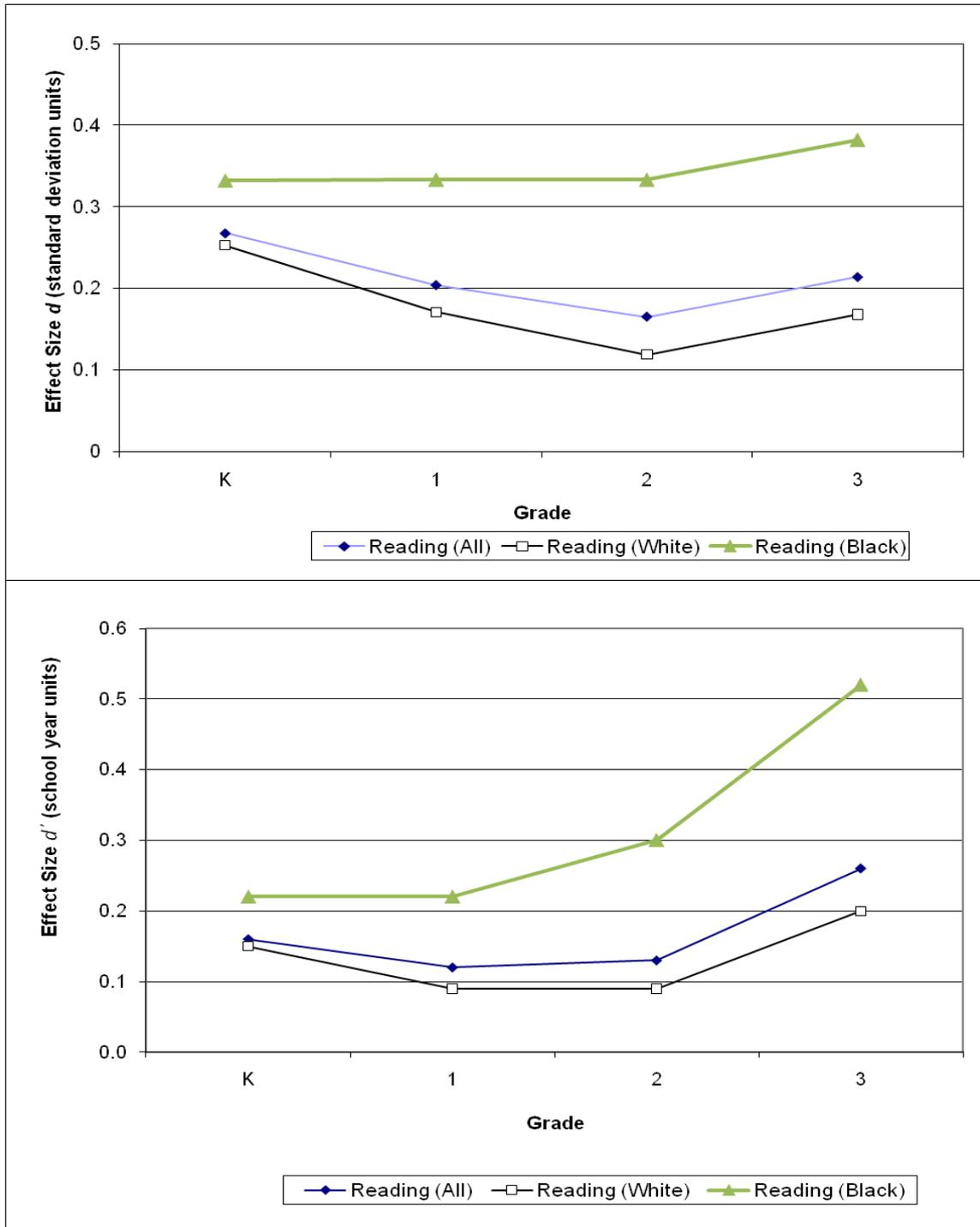
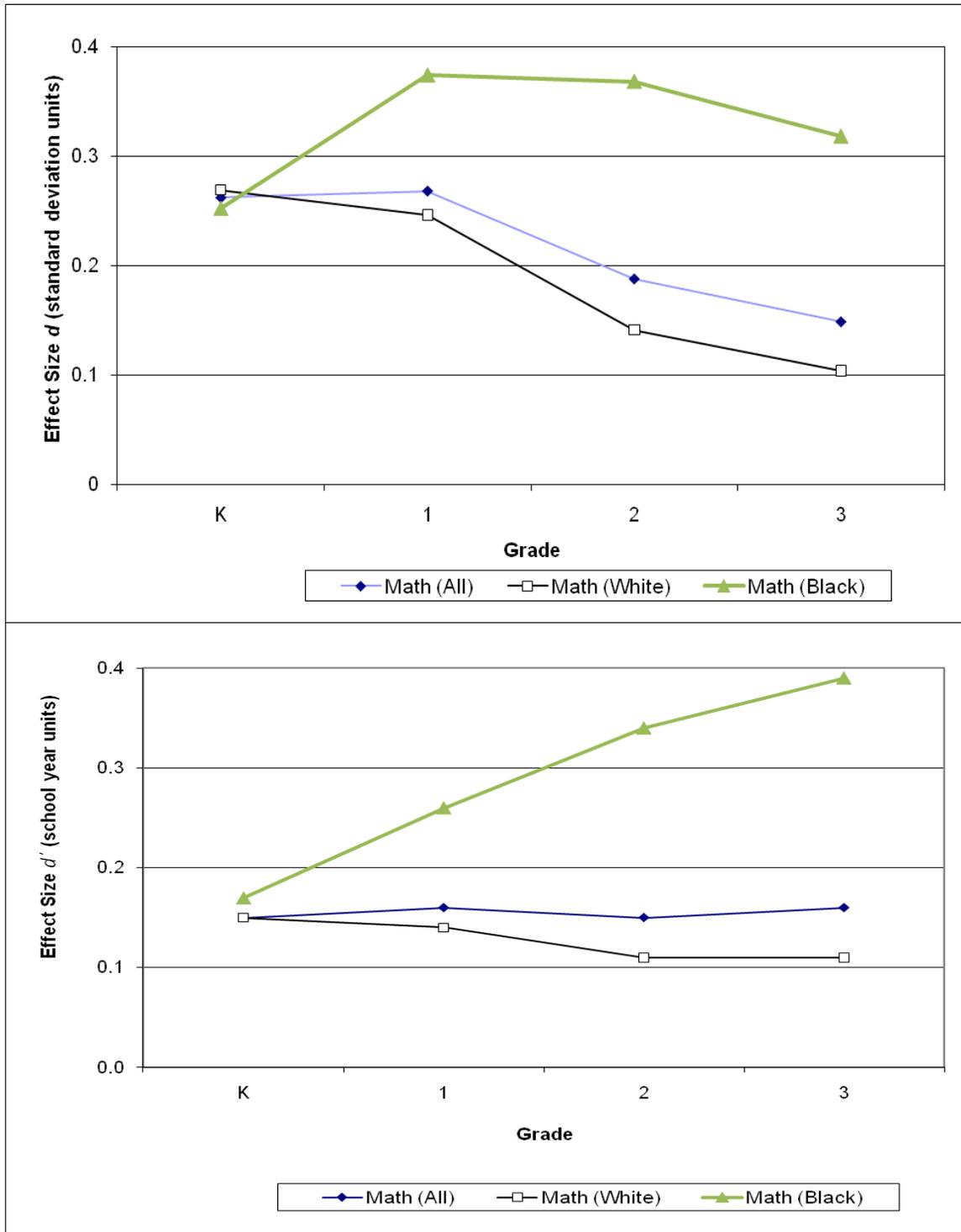


Figure 5

Project STAR small class effects in K-3 math based on d (standard deviation units) in the upper panel and d' (school year/month units) in the lower panel



Notes

¹ This difference is attributable to the fact that GE variance is bound to increase if IRT metric shows a pattern of constant within-grade variance and decelerated growth in the mean (Yen, 1986; Schulz & Nicewander, 1997).

² ECLS-K and NELS 8th grade tests have close alignment with each other, as both adopted similar assessment frameworks and test items (Najarian, Pollack, & Sorongon, 2009).

³ Since sampling designs were similar across the three tests, only sample size differences were considered for weighting (see Appendix). Our sensitivity analysis revealed that the results of synthesis without use of differential weights were very similar. The growth norms were highly convergent among the three tests with any paired correlation coefficients at or above .99.

⁴ Forty-seven percent of the STAR kindergarten sample attended rural schools. Approximately 48 percent of the STAR students were on free or reduced lunch compared to approximately 29 percent of public school students nationally (in 1987-1988). Approximately 33 percent of the STAR sample was minorities, of which 98 percent were Black. In contrast, the entire SAT7 spring standardization sample consisted of 22 percent minority students of which 55 percent were Black (Psychological Corporation, 1985). Similarly, our longitudinal sample (ECLS-K) had 34 percent minority students (47 percent Black).

⁵ For MAT, TN and SAT, they used equating of levels program in which students took two adjacent levels of the tests. MAT 8th edition comprises a battery of fourteen overlapping test levels (Harcourt, 2002). TN 2nd edition comprises a battery of twelve overlapping test levels (CTB/McGraw-Hill, 2003). SAT 10th edition comprises a battery of thirteen overlapping test levels (Harcourt, 2004). For ECLS-K and NELS:88, they both used equating based on a common set of anchor items across adjacent grade forms and most content areas represented in all grade forms (Najarian et al., 2009, Pollack et al., 2005; Rock et al., 1995).