Mathematics Achievement Growth at the Student and School Levels for Regular and Special Education Elementary Students

Joseph J. Stevens

University of Oregon

Ann Schulte

Arizona State University

Paper presented at the annual meeting of the National Council on Measurement in Education, San Francisco, CA, April, 2013. Address all correspondence to Joe Stevens, 541-346-2445, stevensj@uoregon.edu.

The research reported here was supported by the Institute of Education Sciences, U.S. Department of Education, through Grant R32C110004 awarded to the University of Oregon. The opinions expressed are those of the authors and do not represent views of the Institute or the U.S. Department of Education.

Mathematics Achievement Growth at the Student and School Levels for Regular and Special Education Elementary Students

Despite the promise of growth models, there is a paucity of research that applies these methods to students with disabilities (SWD). Published growth analyses that include special education students have shown that, although there are large differences in intercept, there are often no statistically significant differences in slope for students with disabilities (e.g., Stevens, 2005; Wei et al., 2011). The current paper reports preliminary results of analyses of the North Carolina end-of-grade mathematics test to estimate student growth during elementary school grades 3-5. The cohort analyzed is composed of all students present in 2001 in grade three and tracks students through grades four and five. We applied two-level, Hierarchical Linear Models (HLM) to examine student mathematics growth over time and the relationships of mathematics achievement with demographic variables as well as student exceptionality categories. We also used three-level HLM models to examine growth across schools and the relationships among several school level characteristics (e.g., school size, percent free/reduced lunch, percent special education, etc.) and average school mathematics achievement.

The academic achievement of students with disabilities (SWD) has been a persistent policy and practice concern for the past several decades. The majority of studies reporting the analysis of student academic growth do not explicitly analyze growth for students receiving special education. Some investigators have reported academic growth for special education students in a few selected exceptionality categories (Morgan, 2009; 2011) but studies are rare that report academic growth for many categories of exceptional students (Carlson et al., 2011; Wei et al., 2011; 2012).

The No Child Left Behind Act (NCLB, 2001) embodied federal concerns about educational quality and expressed them through enactment of assessment and accountability measures designed to ensure that all students were learning to high standards and achievement gaps between majority students and other groups of students were closing (McGuinn, 2005). NCLB requires reporting of school-level outcomes as well as the disaggregation of achievement test scores for subgroups who have historically performed poorly relative to other students. States vary considerably in their assessment instruments, definition of grade level proficiency, and models for evaluating and judging progress toward accountability goals (Heck, 2006; Linn, 2008). However, increasing awareness that the status-based accountability models promoted under the No Child Left Behind act of 2001 may not adequately capture student performance nor produce unbiased estimates of the effects of teachers and schools (Hanushek & Raymond, 2005; Heck, 2006; Linn, 2008; Linn & Haug, 2002; Schulte & Villwock, 2004; Teddlie & Reynolds, 2000; Willms, 1992; Zvoch & Stevens, 2005) has lead to flexibility under the Race To The Top program and greater attention to the use of growth models (Manna & Ryan, 2011; USDOE, 2010).

An emphasis in both NCLB and RTTT has been the idea of universal proficiency and the evaluation of the performance of student subgroups with the goal of closing achievement gaps between groups. However, we know relatively little about the impact of states' particular assessment and accountability choices on the reliability and validity of disaggregated test scores for students with disabilities (SWD), one of the targeted subgroups in NCLB. Schools' performance with this group of students has been a concern for decades (Carlberg & Kavale, 1980; Schulte, Osborne, & Erchul, 1998) and many states report that over 70% of students with disabilities perform below proficiency on annual statewide reading and mathematics tests. In a recent three-state study of schools that failed to make adequate yearly progress (AYP) targets, Eckes and Swando (2009) found that the most frequent reason for schools' AYP failure was the performance of the SWD subgroup. RTTT places greater emphasis on student growth than the status only models of NCLB. But we know little about trends in growth especially for the SWD subgroup as a whole and even less for students in specific exceptionality subgroups.

What is the course of growth over grades for students overall and in specific subgroups? Recent reports of state test score trends have indicated that students at risk educationally, which includes students with disabilities, have largely participated in their state assessments; however, the majority of their performances have not met state proficiency standards (Center on Educational Policy, 2009; Thurlow et al., 2008). And yet educators sense that students with disabilities are progressing, but just not as fast as their peers because it often takes more time for them to learn. Researchers are only beginning to take advantage of greater availability of large scale assessment data, state data systems developed under recent federal initiatives, and more sophisticated growth analytic techniques now available to provide answers to basic questions about achievement growth (Rescorla & Rosenthal, 2004).

To address these questions, there are advantages in using datasets that not only include students with disabilities but students without disabilities in order to understand how the former differ from their non-disabled peers. This approach is in keeping with the principles of "inclusion" and "least restrictive alternative" that underlie many policy decisions regarding students with disabilities, including their participation in current standards-based reforms (McDonnell, McLaughlin & Morison, 1997). By examining the relative performance and growth of students across subgroups, NCLB and RTTT policy expectations regarding achievement can be studied.

The focus of the present study is on mathematics achievement in elementary school children. Mathematics proficiency is considered to be an essential requirement to later achievement and matriculation through the scholastic system. Studies that have examined student mathematics achievement growth across grades generally agree in their findings. Most studies have found curvilinear mathematics growth over grades with relatively high growth in the early grades that progressively decelerates in later grades (Bloom et al., 2008; Carlson et al., 2011; CTB/ McGraw-Hill, 2003; Ding, Davison, & Petersen, 2005; Harcourt Educational Measurement, 2002; Lee, 2010; Morgan, Farkas & Wu, 2011; Shin, Davison, Long, Chan, & Heistad, 2013).

For instance, Lee (2010) used multiple sources of national assessment data, including long-term trend data from the National Assessment of Educational Progress, the Early Childhood Longitudinal Study-Kindergarten, and norms from two standardized achievement tests to examine mathematics growth in the United States over three decades. Lee characterized typical mathematics achievement growth for United States school children as consisting of an overall achievement gain of 6 to 7 standard deviations from kindergarten entry to high school exit. Typical gains were one standard deviation per grade in the primary grades, a half standard deviation per grade in middle school, and a quarter of a standard deviation per grade in high school. By examining average annual gains in mathematics achievement from kindergarten to Grade 12 across six nationally normed tests, Bloom et al. (2008) reached similar conclusions about the general pattern of decelerating mathematics achievement growth over time.

Although some studies have found slight variations in the general pattern described by Bloom et al. (2008) and Lee (2010), such as increases in growth from one grade to the next within an overall deceleration in growth with increasing grade (e.g., Muthen & Koo, 1998;

Bodovski & Farkas, 2007), the general pattern is largely consistent. Decelerating mathematics growth has been attributed to a number of factors including the decreasing rate of growth in children's cognitive capacity with age, the increasing complexity of mathematics content (Lee, 2010), changes in the structure of the content assessed, and changes in assessments and instrumentation.

Central to NCLB and RTTT policy is the goal of closing or eliminating the achievement gap between student subgroups. Evaluation of progress towards this goal is enhanced through the analysis of growth over time for student subgroups. There is a great deal of variability in study findings evaluating the achievement gap. On a variety of mathematics and reading assessments including district, state, and federal tests, a number of studies have found increases in achievement gaps over time (Abedi et al., 2005; Butler & Castellon-Wellington, 2005; Chang, Singh, & Filer, 2009; Grimm, 2008; Muthén & Khoo, 1998; Williamson, Appelbaum, & Epanchin, 1991). Less frequently, studies have reported stable achievement gaps over time (Anderson, Wilson, & Fielding, 1988; Baker et al., 1984; Jordan, Kaplan, & Hanich, 2002; Juel, 1988; McGee, Williams, Share, Anderson, & Silva, 1986; Scarborough, 1998; Shaywitz et al., 1995). There are also a number of studies that have reported decreases in the achievement gap, again across a variety of mathematics and reading assessments including district, state, and federal tests (Bast & Reitsma, 1997, 1998; Ding, Davison, & Petersen, 2005; Han, 2008; Galindo, 2010; Leppanen et al., 2004; Parrila et al., 2005; Scarborough & Parker, 2003; Shaywitz et al., 1995; Tate, 1997). Differences in study findings may be due to a variety of factors including differences in age or grade, academic content, analytic methodology, assessment instruments (e.g., content covered, difficulty, floor and ceiling effects, equating limitations), student groups studied, and inclusion or exclusion of students from study analytic samples. While there are differences across studies, in a review of studies of mathematics achievement, Shin et al. (2013) found that achievement gaps generally tended to increase or remain stable across grades.

Examination of achievement gap differences has largely focused on subgroups defined by NCLB: economic disadvantage, language proficiency, race/ethnicity, and special education status (represented as a dichotomy). There is evidence that economically disadvantaged students start school behind their more advantaged peers (Clements 2004; Davison, Seok Seo, Davenport, Butterbaugh, & Davison, 2004; Denton and West 2002; Denton, West, and Walston 2003; Hart & Risley, 1995; Lee & Burkham, 2002) and these differences persist throughout the school years (Butler & Castellon-Wellington, 2005; McCoach, O'Connell, Reis, & Levitt, 2006; Molfese, Modglin, & Molfese, 2003; Wright & Li, 2008). In several studies, gaps in academic performance between student subgroups have been observed in initial performance and have been observed to persist across grades (Chatterji 2005; LoGerfo, Nichols, & Reardon 2006; Morgan, Farkas, & Wu 2007; Princiotta, Flanagan, & Germino-Hausken 2006).

Only a handful of studies, however, have examined student growth and gaps in achievement for students in specific exceptionality categories. With some exceptions (e.g., Carlson et al., 2011; Wei, 2011; Wei,2013), most studies of achievement growth for students with disabilities have been limited to one or two disability categories, a limited age span, and relatively small sample sizes. For example, the widely-cited study by Francis, Shaywitz, Stuebing, Shaywitz, and Fletcher (1996) examined growth in reading for 32 students with reading disabilities, 37 students with low reading achievement, and 334 non-disabled students, conducting reading assessments annually for students from age 7 to 17.

Despite lower scores overall for students with disabilities, there is considerable heterogeneity in achievement across, as well as within, exceptionality categories. At the elementary school level, students with speech/language and visual impairments have score distributions similar to their same-age, non-disabled peers in the area of math calculation. In contrast, over 80% of students with intellectual disabilities score in the bottom quartile (Blackorby et al., 2005). In terms of within-category heterogeneity, although students with learning disabilities score well below the mean in reading comprehension at high school exit, a significant percentage (10%) score in the top quartile (Wagner, Newman, Cameto, & Levine, 2006). This heterogeneity has led some to suggest that the single category, “students with disabilities,” is not a meaningful one as the group includes children with cognitive, physical, and behavioral impairments that can range from mild to severe, with varying impact on schooling and achievement.

Studies of student growth have also been limited by features of study design or analysis. Some studies have used too few time points to meaningfully evaluate growth functions. For example, Scarborough & Parker (2003) used pre-post tests from second through eighth grade to evaluate growth functions but the use of two waves of data precludes the possibility of actually evaluating the functional form of growth. Another concern in the design of growth studies is the amount of student attrition and mobility from one school to another as well as the methods used to model and estimate growth in the face of missing data. Study results may also be biased if relevant variables are omitted from analysis. For example, some growth studies do not include student demographic or background characteristics. Because these variables are often intercorrelated with student subgroup status variables, estimates may be biased. For example, Kieffer (2008) examined the extent to which estimates of a subgroup’s achievement growth trajectories changed depending on whether additional characteristics of the student, family, or school were included in analysis.

Another variable of importance in estimation of growth trajectories involves the time at which model predictors are measured. For example, most longitudinal studies use information on student special education status defined at the first wave or measurement occasion (Carlson, 2011; Wei, 2012). However, students enter and exit from special education status over time (Schulte et al., 2012).

The purpose of the present study was to examine mathematics growth for students with and without disabilities in grades 3 through 5 using multilevel longitudinal models. Another goal of the study was to provide empirical benchmarks that provide additional contextualization and interpretation of growth by describing all students' mathematics achievement from year to year and in comparison to non-SWD and regular education students. A third goal of the study was to examine school level differences in mathematics achievement and growth and determine whether achievement was associated with school characteristics.

Methods

Sample

The original data set can be conceptualized as a population in that it represents all students in North Carolina in the third grade for the first time (i.e., not repeating third grade) that were present in the state’s large scale achievement database in 2000-01 (*N* = 103,123; see first columns of Table 1 labeled Total Sample). To create an analytic sample that was appropriate for our research questions, we systematically excluded a number of individuals from the population. Students who did not follow the typical grade level sequence from grades 3 to 7, largely due to grade retention (*N* = 8,315; 8.1%) and students who never participated in the large scale mathematics test in any of the grades 3-7 (*N* = 1,207; 0.1%) were excluded from the analytic sample used to report results below (see second column of Table 1). In addition, we also excluded students with missing values on predictor variables. Fourteen students were missing an ethnicity code, three students had no sex code and 888 (1%) students had no value for parental education and were excluded. We also excluded thirty students coded as “other” ethnicity in 2001 because the state dropped the category in subsequent years. Students in exceptionality categories with a sample size smaller than 100 (see second column of Table 1) and 255 (< 1%) students with missing exceptionality codes were excluded. Overall, after exclusions, the analytic sample represented 89.3 percent of the population of students in the state system data file. It can be seen in Table 1 that the analytic sample was quite similar to the population in most respects but, as evaluated by single degree of freedom chi-square tests, the analytic sample was characterized by significantly fewer students with autism, emotional disturbance, intellectual disabilities, and hearing impairments. There were no statistically significant differences between the two samples on any of the other variables listed in Table 1.

We forward matched the cohort of students who were present in the data base in 2000-2001 in Grade 3 to all succeeding years through Grade 7 (2004-05). By forward matching we mean that new students entering the system in future years were not added to the cohort (e.g., new grade 4 students in 2001-02, new Grade 5 students in 2002-03, etc.). We tracked the cohort only through 7th grade because the state administered a new edition of the mathematics test in 2005-06 and we wanted to avoid confounding mathematics growth with any changes in performance due to changes in test edition. We included any student who had at least one mathematics score during the study period. Of the 92,045 students in the analytic sample, 80.9 percent had mathematics scores in all five years, another 5.1 percent had scores in four years, 4.3

percent had scores in three years, 3.8 percent had scores in two years, and 6.0 percent had only one mathematics score during the five-year study period. Achievement data for a student could be missing in a year because a student moved out of state or to a private school that did not participate in the testing program, or because the student took an alternate assessment in that year.

For the school level analyses we used listwise deletion to remove students who did not attend the same school in all three grades (3-5). Across grades four and five, 33,196 (36.5%) of students changed schools. Deletion of these cases resulted in an analytic sample of 57,692 cases for the three-level study.

Measures

For all analyses reported, the outcome measure used was the standardized, second edition EOG test score in mathematics for each student. The North Carolina EOG Mathematics Tests are multiple-choice tests designed to measure the goals and objectives of the mathematics curriculum adopted by the North Carolina State Board of Education for each grade. The mathematics competency goals and objectives are organized into four strands: (a) number sense, numeration, and numerical operations; (b) spatial sense, measurement, and geometry; (c) patterns, relationships, and functions; and (d) data, probability, and statistics (Public Schools of North Carolina, 1996a; 1996b). The mathematics test is comprised of two sections: Calculator Inactive and Calculator Active. The Calculator Inactive section is comprised of 28 items that assess students' ability to perform mathematical computations without a calculator. The section has a time limit of 60 minutes for third through fifth grade students. For the Calculator Active section of the test students are allowed to use calculators, rulers, and protractors. The section is comprised of 54 items with a time limit of 135 minutes for third through fifth grade students. Overall, the EOG mathematics test is comprised of 82 items to be completed in 195 minutes.

Each student's score on the mathematics test is determined by calculating the number of items correctly answered. Scores from the North Carolina mathematics tests are reported as developmental scale scores and proficiency levels.. Raw scores (total number of items correct) are converted to scale scores using a 3PL IRT model as described by Thissen and Orlando (2001, pp. 119-130) following procedures described by Williams, Pommerich, and Thissen (1998). The developmental scale defines means and standard deviations for each grade level ranging from the Grade 3 through Grade 8. To allow comparisons across years, the state has performed vertical linking using a common items design in which adjacent grades are forward linked. That is, approximately 10 items for third grade students also appeared on the fourth grade test and 10 items for fourth grade students were included on the fifth grade test and so forth. IRT equating methods were then applied to create a linkage of each EOG test to the sequent EOF test in order to create a common developmental scale that spans all grades.

In previous analyses (PSNC, 2006a), internal consistency reliability estimates (coefficient alpha) for the EOG-Math Tests ranged from .94 to .96 for grades three through eight. These estimates were the same when examined separately by gender, ethnicity, disability status, or Limited English Proficiency status. The standard error of measurement for all students taking the test ranged from two to six points; the standard error of measurement for 95% of the students taking the test who score within two standard deviations of the mean was two to three points.

Evidence of both content and criterion-related validity has been gathered for the EOG-Math Tests as well (PSNC, 2006a). Evidence of content validity was examined through analyses of how well items were aligned with the four basic strands that are emphasized in the curriculum.

North Carolina teachers also reviewed the math test items and rated them as having the following qualities to a “high” or “superior” degree: (a) tapped grade-level curriculum objectives; (b) reflected the grade-level curriculum taught in their individual school; (c) were clearly and concisely written; (d) were not biased against students of different races, genders, socioeconomic statuses, or geographic locations; and (e) had only one answer. Evidence of criterion-related validity was demonstrated through moderate to strong Pearson correlation coefficients with associated variables (e.g., assigned achievement level by expected grade, teacher judgment of achievement by assigned achievement level, teacher judgment of achievement by expected grade, teacher judgment of achievement by math scale score, and expected grade by math scale score; PSNC, 2006).

Procedures

North Carolina began its current large-scale assessment and accountability program in reading and mathematics in 1993. Student- and school-level test data from the program’s inception to the present, along with demographic and other descriptive information, are housed for research purposes at the North Carolina Education Research Data Center (NCERDC) at Duke University. The testing program and procedures used to build the longitudinal dataset used in the present study are described below.

Test administration. The North Carolina EOG mathematics tests are part of the North Carolina Department of Public Instruction’s school accountability program and are administered to students in May of each year, typically in the general education classroom by general education teachers and proctors. Generally, state testing guidelines require students who are pursuing the state’s standard course of study to be included in testing, regardless of special education status. However, there are procedures to exempt students from testing or allow an alternate assessment for a variety of reasons, including limited English proficiency or determination by an IEP team that a student with a disability should not participate in testing (PSNC, 2007). For the 2000-01 school year, the participation rate for the EOG-Mathematics for students in the third grade was 98% overall, and 81.6% for students with disabilities. Within the special education population, the participation rate by disability varied from 0% for students with severe/profound mental retardation to 91% for students with learning disabilities.

In addition to an exemption from testing or participation in an alternate assessment, , students could receive a variety of accommodations during the EOG tests, if use of the accommodations had been specified in their IEP, Section 504 Plan, or Limited English Proficiency Plan (PSNC 2006a; 2006b). Available accommodations included extended testing time, taking the test in a separate setting, having a translator, using an abacus, using a test with large print, and marking in the test booklet. The most common accommodations provided were: extended testing time, taking the test in a separate room, having the test read aloud, marking in the test booklet, and taking the test in multiple sessions.

Determination of special education status. The identification of students as disabled and in need of special education services occurred at the school level, based on procedures described in the state's *Policies Governing Services for Children with Disabilities* (PSNC, 2007). In the state testing database, students were classified into 1 of 17 categories in terms of their exceptionality. Two of these categories were for students who were not receiving special education services (a) regular education students, not identified as exceptional, and (b) academically/intellectually gifted. The remaining 15 categories paralleled the disabilities categories in IDEA, although the terms sometimes varied somewhat from the IDEA designations (e.g., "Behaviorally-Emotionally Handicapped" for IDEA's "Emotional Disturbance") and some

IDEA categories were split (e.g., there were three state designations for mental retardation corresponding to degree of intellectual impairment).

Longitudinal database construction. The study dataset was constructed from multiple annual student electronic files available from the NCERDC. At each school, there is one EOG record for each student who was a member of that school at the time of test, even if the student was absent or exempt from testing. Also, included in the EOG record is student demographic information, which testing accommodations were used during that year's assessment, and disability classification, all coded by school personnel at the time of testing. NCERDC compiles these EOG records by grade and year, standardizes district and school codes, conducts internal consistency and validity checks on the files, and assigns a unique identifier for each student that can be used to match student records across years.

To create the longitudinal records, we conducted additional data quality checks on the EOG files for the years 2001 to 2005, and then used a set of second annual files providing test scores, student demographic and attendance information to resolve discrepancies found in the testing records (e.g., two different test score records with the same student identification number in the EOG file). We then merged annual files by student identification number to create the longitudinal dataset.

Analytic Methods

We used several analyses to describe and estimate the growth in mathematics achievement of students with and without disabilities. First we applied descriptive methods to report mean growth for all students by exceptionality category. We then modeled student growth using multilevel, longitudinal analyses (Raudenbush & Bryk, 2002). Because the measurement occasions studied (EOG test in grades 3-5) were equally spaced, we centered time at the first occasion (grade 3) and used integer values for subsequent time points to indicate their relative distance from the first time point. Thus a student with measurements at all time points would have a time value coded as 0, 1, 2 corresponding to scores in grades 3, 4, and 5. Because the multilevel modeling approach to longitudinal analysis does not require data to be balanced within subjects (Raudenbush & Bryk, 2002), we included data from any available time point for a student even when data from other time points were missing.

In the two-level multilevel analyses, we first applied unconditional growth models. The next model step applied conditional multilevel growth models to examine model results for each student exceptionality category. The last step in our analyses was to add student characteristics and demographic variables as predictors of mathematics achievement. All multilevel analyses were conducted using HLM 7.0 (Raudenbush et al., 2011), full maximum likelihood estimation, and specification of model parameters as random effects. The conditional models included a level-1 model that specified student mathematics scores predicted by time of measurement and a level-2 model composed of the prediction of level-1 model parameters as a function of student exceptionality categories and demographic characteristics. The initial level-1 model can be described as follows:

$$(Y_{ti}) = \pi_{0i} + \pi_{1i}(\text{time}) + e_{ti} \qquad (1)$$

where Y is the mathematics scale score for student i at time t and π_{0i} is the initial status or intercept for student i at time 0, π_{1i} is the initial linear rate of change, and e_{ti} is the residual for each student.

At level 2, the level-1 parameters were modeled using student level predictors. We left all dichotomous predictors uncentered. However, we grand-mean centered the parental education variable that had multiple categories. With this approach, coefficients can be interpreted for dichotomous predictors as the effect for the group coded one and for the continuous predictor (parental education) as the effect for a student with the mean value of the predictor. The level-2 equations (for the study's full set of predictors) for the mathematics initial status and growth rate parameters are as follows:

$$\textit{Initial Status},\ \pi_{0i} = \beta_{00} + \sum \beta_{0k}\ (\textit{Predictor}_i) + r_{0i} \qquad (2)$$

$$\textit{Rate of Change},\ \pi_{1i} = \beta_{10} + \sum \beta_{1k}\ (\textit{Predictor}_i) + r_{1i} \qquad (3)$$

where π_{0i} represents each student's grade 3 mathematics score, β_{00} is the grand mean average mathematics score at grade 3 for all students, each β_{0k} represents the grand mean average partial regression coefficient relating the predictor of interest to students initial status, and r_{0i} is the residual between the fitted predictor value for each student and the student's observed mathematics score. Each individual's initial rate of change, π_{1i}, was modeled as a function of the grand mean average mathematics initial rate of change for all students, β_{10}. Each β_{1k} represents the grand mean average partial regression coefficient relating the predictor of interest to student's initial rate of change, and r_{1i} is the residual between the fitted predictor value for each student's initial rate of change and the observed initial rate of change.

Another goal of our analyses was to provide empirical benchmarks (Bloom et al., 2008) of achievement scores as a contextual aid in interpreting students' mathematics growth. We used two methods to provide interpretive benchmarks: (a) year to year growth effect sizes, and (b) achievement gap effect sizes. We estimated effect sizes (ES) for regular education students and students from each exceptionality category by examining the mean difference from one year to the next in ratio to the pooled standard deviation for the two years. To estimate achievement gaps, we compared the mean mathematics performance in each year for students in a particular exceptionality category to the mean mathematics performance of regular education students or to all non-SWD students (i.e., regular education and academically/intellectually gifted students combined). This mean difference was divided by the standard deviation of the scores for all students in that grade.

In order to estimate school level differences in mathematics performance we also applied three-level longitudinal models. The first two levels of the three-level models (time and students) were exactly as presented above in equations 1-3. A notable difference in the three level models was that all level two predictors other than the intercept and slope were treated as fixed effects. This specification was used to ease model estimation and because our interest in the three-level models was only in the differences in school average intercepts and school average slopes. Six school characteristics (proportion regular education students, proportion white students, average parental education in the school, proportion limited English proficient students, and school size) were used as predictors of level two school intercepts and slopes.

Results

The analytic sample was used to describe growth trajectories for each student group, estimate growth trajectories using multilevel methods, and create empirical benchmarks to further characterize mathematics growth by student group. We report results separately for regular education students and students identified as Academically/Intellectually Gifted students. However, because these two groups are not distinguished from each other in many accountability applications, we also provide information below in which these two groups are combined and described as Non-SWD students.

Average Observed Growth by Exceptionality Category

Table 2 shows means and standard deviations of mathematics scale scores by grade for all students, all non-SWD students, regular education students, Academically Gifted students, and students in each exceptionality category that had a sample size greater than 100 in 2000-01. It can be seen in Table 2 or Figure 1 that most average trajectories are curvilinear in form, with decelerating growth across grades. As would be expected, the Academically Gifted group of students had the highest trajectory at initial status in grade 3 and showed a somewhat higher slope than other student groups. In terms of level of performance, the order of growth trajectories from highest to lowest performing were the regular education, Speech-Language Impairment, Hearing Impairment, Autism, Specific Learning Disabled, Other Health Impairment, Emotionally Disturbed, and Intellectual Disability. While there was some crossing of growth trajectories in the middle of the distribution, generally there was some differentiation between subgroups of students in initial status. All students groups demonstrated growth over time although there are some differences apparent in rate of change and rate of curvature for some groups. These observations are tested further using the statistical models described in the next section.

Two-Level Growth Models

The first model applied was a fully unconditional random effects model that only estimated grand means and variance components. We then applied a two-level longitudinal models that represented student mathematic scores as a linear function of time (see first columns of Table 2). On average, across all students, the estimated mean mathematics scale score in grade 3 was 251.08. The average initial rate of change was a statistically significant growth rate of 6.92 scale score points per grade ($z = 520.02$, $SE = 0.01$, $p < .001$). Inspection of the variance components for each model parameter (intercept, initial rate of change, curvature) in this model and also in the following two models showed that there was significant variation in parameters across students for all three random effect model parameters ($p < .001$).

We next applied two conditional models that added predictors to the quadratic model. In the first model we added dummy coded predictors that reflected students' exceptionality category. Multilevel model results for the exceptionality predictors are shown in the middle section of Table 2. It can be seen in Table 2 that all differences in intercept or initial status in grade 3 between regular education students (i.e., the group coded zero on all vectors), Academically Gifted students, and students in all exceptionality categories were statistically significant (*p* < .001). While Academically Gifted students scored about 10 scale score points higher in grade 3 than regular education students, all of the exceptionality students had significantly lower initial mathematics performance in Grade 3. Students with an Intellectual Disability showed the largest differences in initial status, on average 14 scale score points lower than regular education students. The group showing the smallest contrast with regular education students in Grade 3 was students with a Speech-language Impairment who scored an average of 1.45 scale score points lower.

Three exceptionality groups, students with a Behaviorally Handicapped, students with a Hearing Impairment and students with a Speech-language Impairment did not differ significantly in initial rate of change in contrast to the regular education students. The remaining comparisons between student groups all showed statistically significant (*p* < .001) lower initial rates of growth in contrast to the regular education students. The smallest difference was observed for the Academically gifted students but annual decreases in initial growth rate of from .86 to 1.69 scale score points were observed for the remaining exceptionality groups.

We then expanded the multilevel growth model by adding an additional set of predictors representing student demographic and background characteristics. As can be seen in the right-most columns of Table 2, all predictors were significantly related to students' initial achievement level in grade 3. The estimated grand mean initial status or intercept (252.63) now represents the average mathematics achievement in grade three for white male students, who are not limited-English proficient or receiving free lunch with an average parental education. While the magnitude of several parameter estimates vary slightly, hypothesis testing results for the grand mean of parameters and for specific exceptionality category contrasts are very similar to those for the previous model with three exceptions: the initial rate of change for students with Autism was no longer statistically significant and the initial rate of change for students with a Speech-language Impairment were now statistically significant in this model when they had been nonsignificant in the previous model.

Examination of results for the new predictors showed that females (-0.44), limited English proficient students (-2.69), free lunch recipients (-1.25), Black students (-4.13), Hispanic students (-0.94), and American Indian students (-1.85) all had significantly lower initial mathematics performance in Grade 3. Results for rate of change showed that all predictors except sex and limited English proficiency showed statistically significant differences in rate of change in comparison to the reference group. Students with higher levels of parental education (0.05), who were Asian (1.28), Black (0.22), or Hispanic (0.93) showed larger rates of change. Students who were free lunch recipients (-0.19) or American Indians (-1.46) showed lesser rates of change.

Empirical Growth Benchmarks

To provide additional context for interpretation of differences in student growth we examined two representations of student group differences: (a) change in mathematics growth from year to year expressed as an effect size, and (b) achievement gap effect size between student exceptionality groups and either non-SWD students or regular education students.

Mathematics growth effect sizes. Table 4 shows year-to-year growth expressed as an effect size for all non-SWD students, regular education students, Academically Gifted students, and students in each exceptionality category. As can be seen in the first column of Table 4, student growth from grade 3 to grade 4 is substantial for all student groups, ranging from 0.72 for the Other Health Impaired group to 1.19 for the Intellectual Disability group. Inspection of the grade transition effect sizes from grades 3-4 to grades 4-5 showed that these effect sizes diminished for all student groups across grades.

Achievement gap effect sizes. Another representation of mathematics achievement for student groups can be obtained by inspecting the relative growth of one student group in comparison to another. Table 5 shows effect size comparisons of students by exceptionality group in comparison to Non-SWD students in the upper portion of the table and in comparison to regular education students in the lower portion of the table. In comparison to Non-SWD students, the student exceptionality group mathematics scores were lower by from .30 to almost two standard deviations in grade three. Inspecting achievement gap differences over time, it can be seen that in general, although there is some variation from grade to grade, achievement gap differences remain relatively stable or increase from grade 3 to grade 5. As shown in the lower portion of Table 5, when examined separately, Academically Gifted students perform approximately one and a quarter standard deviations higher than regular education students over grades 3-5. Achievement gap effects sizes for students in specific exceptionality categories are somewhat smaller than the comparisons in the upper portion of the table for all non-SWD students but generally show the same pattern of differences by group and over time.

School Level Results

Another goal of the present study was to examine mathematics achievement aggregated to the school level. As described earlier, these analyses were performed on a reduced sample that had stable enrollment at the same elementary school in Grades 3-5. The same unconditional and conditional HLM models reported above were analyzed while adding a third level of analysis representing school membership. For brevity, we only report here the final results of the fully conditional three-level model (see Table 6). As can be seen in the table, all student level predictors of mathematics intercept were statistically significant except the contrast between Asian and white students. At the school level, school average intercept was significantly associated only with parental education with an increase of .06 scale score points with each additional level of parental education over the grand mean. At the student level, limited English proficient students, females, students with an Emotional Disturbance, students with a Hearing Impairment, students with a Speech or Language Impairment, Black students, and American Indian students did not differ significantly from the grand mean slope. Higher levels of parental education, Academically gifted students, and Hispanics all showed significantly higher slopes. Students receiving a free or reduced price lunch, students with an Intellectual Disability, students with an Other Health Impairment, and students with a Specific Learning Disability all showed significantly lesser slopes than the grand mean. At the school level, only one school characteristic was significantly associated with school slope; larger schools were associated with lesser average growth rates. Figure 2 shows estimated mathematics achievement growth as a function of regular vs. special education status and at the 25th and 75th percentiles of parental education. Figure 3 shows estimated mathematics achievement growth as a function of proportion of free lunch students and school size.

We also examined the relationship between school average intercept and school average slope. Figure 4 shows estimated school intercept plotted against estimated school slope with schools coded as low or high proportion students with disabilities. The reference lines in the figure indicate median intercept and slope. It can be seen in the figure that, although the relationship between intercept and slope at the student level is strongly positive in correlation (i.e., a Matthew effect), at the school level average intercepts and average slopes are negatively correlated (*r* = -0.29).

In previous research (e.g., Stevens, 2005) we have argued that evidence for the adequacy and validity of accountability models can be garnered through a process of pattern matching (Shadish, Cook, & Campbell, 2002) that examines the correlation of school performance estimates (in the present study intercept and slope) with confounding factors that can undermine the adequacy of an estimate of school effects. We examined the magnitude of correlations with six predictors that are possible confounding factors in the estimation of school effects: proportion Regular Education students, proportion White students, proportion free lunch students, proportion limited English students, level of parental education, and school size. Table 7 shows the correlation of school intercepts and slopes with these school characteristics. As can be seen in the table, the relationship between school intercept is substantially larger in almost all cases than for school slope.

Discussion

The present study examined mathematics achievement growth for regular education students, academically gifted students, and students in seven specific exceptionality categories. The study examined a large and inclusive statewide database used for operational accountability reporting and decision-making.

Major Findings

Students in all exceptionality categories showed statistically significant growth in mathematics achievement over the studied grades. There was also statistically significant variability by student subgroup in mathematics achievement both at the initial assessment in grade three and in rates of growth over time. These general results are quite consistent with other published studies. There is evidence that there are gaps in achievement when students enter school, at the time of first assessment on state mandated accountability tests, and in rates of growth. In the present study, the highest performing student groups were Academically Gifted students followed by regular education students. The lowest performing student group was students with Intellectual Disabilities.

Findings in Relation to Previous Research

These results are generally consistent with other recent studies of mathematics achievement growth (Bloom et al., 2008; Francis et al., 1996; Morgan et al., 2009; Morgan et al., 2011; Shin et al., 2013; Wei et al. 2011; 2012) which all reported curvilinear growth functions with decelerating growth over age or grade. In the present study, given three waves of measurement, we could only apply linear growth models. However, when additional waves are available, decisions on functional form should be made based on purpose of analysis and study. Linear models are more parsimonious but may under-represent the complexity of change for certain groups, assessments, or ranges of measurement and occasions/grades spanned.

As in a number of previous studies, we found statistically significant differences in mathematics performance as a function of student background characteristics including race/ethnicity, parental education, free lunch status, and language proficiency. It is also noteworthy that the magnitude of some coefficients for specific exceptionality categories changes when other student background characteristics are taken into account. This underscores the importance of including a variety of student descriptors in analyses to disentangle effects among variables (Kieffer, 2008). One of the predictors more strongly and consistently related to outcomes, parental education, confirmed previous findings that parent education level is significantly related to student achievement (Holman, 1995; Phillips et al., 2002).

Another important feature of the present study was the provision of additional benchmark information for interpreting growth results as recommended by Bloom et al. (2008). Common and often unanswered questions in evaluating student changes in proficiency include: "How much growth was there?," "How does growth for different students or groups compare?," and "Are gaps between student groups decreasing." Our results show that the amount of mathematics growth expressed as an effect size decelerates from the earlier grades into middle school for all student groups studied. These results are consistent with many other studies (e.g., Shin et al., 2013) although there is no clear understanding of why this decrease occurs. Some have proposed that the decrease occurs as mathematical content becomes more difficult, more elaborated, and branches into more distinct and complex areas of mathematics (e.g., algebra, geometry, etc.; Carraher & Schliemann, 2007; Kieran, 2007; Shinn et al., 2013). Another possibility is that later learning is hampered by incomplete learning and mastery in earlier grades, essentially compounding content difficulty as time passes. A third possibility is that changes in student motivation, academic self-concept, and beliefs in the worth and utility of mathematics may decrease over time. Another possibility less often cited is that the nature of instruction, assessment, or the interaction of the two results in instruments that are less sensitive to student achievement occurring in later grades.

The laudable goal of closing the gap between SWD and Non-SWD students is an important policy target. However, the policy may not fully embrace empirical evidence about student growth. In order for those who are "behind" to catch up with other students it is necessary that they not only make academic progress but that they make progress at faster, more accelerated growth rates that their higher performing, and often more advantaged peers. Results presented here show that in fact, SWD generally are not closing the achievement gap relative to regular education students or to all Non-SWD students. As noted elsewhere (Morgan, 2011) this raises important questions about how much growth should be expected for these students as well as what reasonable expectations are for "proficiency" for these students and for narrowing the achievement gap.

In another study of achievement growth for students with disabilities, Zvoch and Stevens (2005) found that students in special populations demonstrated achievement growth at a rate that was indistinguishable from their counterparts in the data submitted for accountability reporting. However, data from a more inclusive student cohort revealed that ethnic minority, impoverished, and special education students grew at a slower rate than their peers. Thus, whereas in one student sample it appeared as though schools were under-serving special student populations (i.e., achievement gaps widened over time), in the other it appeared as though traditionally disadvantaged students kept pace with their more advantaged peers. This study highlights the importance of carefully attending to missing data in assessing growth of students with disabilities.

More recently, Schulte (2010) extended her initial investigations of achievement outcomes and growth for students in special education in a single district by using multiple years of extant state-level longitudinal data from NC's testing program. These analyses confirmed that

SWDs differ markedly from non-identified students in terms of achievement level, but less so in terms of growth, and that the high rate of exits and entrances from special education during the elementary school years distorts the size of the achievement gap when data are reported cross-sectionally rather than longitudinally.

Expectations for growth cannot rest solely on policy intent but must also be contextualized by empirical findings that describe the growth that occurs for students as a function of their exceptionality and background. We believe the results presented here provide rich additional context for characterizing and interpreting student growth for SWD and Non-SWD students through the use of effect size and normative interpretations of student growth. Empirical information about growth like that presented here can be used to guide instructional decision-making and policy formation. It is not only important to express how much growth we would like to see from a pedagogical or policy perspective but also how much growth typically occurs and how that growth is conditioned on student exceptionality, characteristics, and background. Richer contextualization of empirical growth results can provide an important foundation to temper decisions about how much growth typically occurs and what can be realistically expected for future growth. Expectations for growth must also consider what additional resources may be needed to close achievement gaps and whether we must recognize different expectations for students with different exceptionalities and needs.

Limitations

A number of limitations in the present study should be noted and considered. First, we were unable to adequately represent and model a number of exceptionality categories because the sample sizes were too small. We also excluded a number of students from analytic samples due to missing data, grade retention or advancement and school mobility. The present study also only includes those students that took the North Carolina EOG test and does not address those who took the North Carolina alternate assessment. As a result, the findings reported here may not fully generalize to student groups other than those studied.

We also know that the method used here to identify special education status and student exceptionality, while convenient is flawed. Student classifications change from grade to grade and student exits and entrances into special education are correlated with achievement. This results in biased cross-sectional or longitudinal reports of performance gains and losses (Schulte, 2010; Ysseldyke & Bielinski, 2002). In addition, SWDs are more likely to be retained in grade and/or perform at the bottom of grade-level score distributions. Both retention (because of a lowered standard relative to promoted peers) and low scores (because of measurement error and regression toward the mean) increase the chances of invalid inferences about the improvement of this subgroup of students when scores are compared across years. In ongoing research (including another paper in this session) we are exploring the patterns of change in student classification as well as alternative methods for representing student exceptionality in longitudinal models.

Another substantial weakness in the present study is the listwise deletion of students whose school enrollment changed during the years studied. Research has shown that listwise deletion in longitudinal models likely introduces significant bias in estimates of student and school performance (Lockwood et al., 2005; Zvoch & Stevens, 2005). While not applied in the present study we are now working on a study with the North Carolina data that compares results depending on whether listwise deletion, cross-classified HLM, or HLM multiple membership models are employed.

References

Abedi, J., Bailey, A., Butler, F., Castellon-Wellington, M., Leon, S., & Mirocha, J. (2005). *The validity of administering large-scale content assessments to English language learners: An investigation from three perspectives*. UCLA Center for the Study of Evaluation/ National Center for Research on Evaluation, Standards, and Student Testing (CSE Technical Report 663): Los Angeles.

Anderson, R. C., Wilson, P. T., & Fielding, L. G. (1988). Growth in reading and how children spend their time outside of school. *Reading Research Quarterly, 23*, 285–303.

Baker, L. A., Decker, S.N., & DeFries, J. C. (1984). Cognitive abilities in reading-disabled children: A longitudinal study. *Journal of Child Psychology and Psychiatry, 25*, 111–117.

Bast, J. W., & Reitsma, P. (1997). Matthew effects in reading: A comparison of latent growth curve models and simplex models with structured means. *Multivariate Behavioral Research, 32*, 135–167.

Bast, J., & Reitsma, P. (1998). Analyzing the development of individual differences in terms of Matthew effects in reading: Result from a Dutch longitudinal study. *Developmental Psychology, 34*(6), 1373–1399.

Blackorby, J., Wagner, M., Cameto, R., Levine, P., Davies, E., et al. (2005). *Engagement, Academic, Social Adjustment, and Independence: The Achievements of Elementary and Middle School Students With Disabilities*. Menlo Park, CA: SRI International.

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. doi: 10.1080/ 19345740802400072

Bodovski & Farkas, 2007

Butler, F. A., & Castellon-Wellington, M. (2005). *Students' concurrent performance on tests of English language proficiency and academic achievement. Validity of administering large-scale content assessments to English language learners: An investigation from three perspectives (final deliverable to OERI/OBEMLA)* (pp. 47–83). Los Angeles: University of California, National Center for Research on Evaluation, Standards, and Student Testing.

Carlberg, C., & Kavale, K. (1980). The efficacy of special versus regular class placement for exceptional children: A meta-analysis. *Journal of Special Education, 14*, 295-309.

Carlson, E., Jenkins, F., Bitterman, A., and Keller, B. (2011). *A Longitudinal View of Receptive Vocabulary and Math Achievement of Young Children with Disabilities*, (NCSER 2011-3006). U.S. Department of Education. Washington, DC: National Center for Special Education Research.

Carraher, D. W., & Schliemann, A. D. (2007). Early algebra. In E. K. Lester (Ed.), *Second handbook of research on mathematics teaching and learning* (pp. 669–706). Charlotte, NC: Information Age.

Center on Educational Policy. (2009). from http://www.cep-dc.org

Chang, M., Singh, K., & Filer, K. (2009). Language factors associated with achievement grouping in mathematics classrooms: A cross-sectional and longitudinal study. *School Effectiveness and School Improvement, 20*(1), 27–45.

Chatterji, M. (2005). Achievement gaps and correlates of early mathematics achievement: Evidence from the ECLS K–First Grade sample. *Education Policy Analysis Archives, 13*(46).

Clements, D. H. (2004). Major Themes and Recommendations. In D. H. Clements, J. Sarama, and A.-M. DiBiase (Eds.), *Engaging Young Children in Mathematics: Standards for Early Childhood Mathematics Education* (pp. 7-72). Mahwah, NJ: Erlbaum.

CTB/McGraw-Hill (2003). *TerraNova technical report*. Monterey, CA: Author.

Davison, M. L., Seok Seo, Y., Davenport, E. C., Jr., Butterbaugh, D., & Davison, L. J. (2004). When do children fall behind? What can be done? *Phi Delta Kappan, 85*(10), 752–761.

Denton, K., & West, J. (2002). *Children's reading and mathematics achievement in kindergarten and first grade*. Washington, DC: National Center for Education Statistics.

Denton, K., West, J., & Walston, J. (2003). *Reading – Young Children's Achievement and Classroom Experiences*. Washington, DC: U.S. Department of Education.

Ding, C. S., Davison, M. L., & Petersen, A. C. (2005). Multidimensional scaling analysis of growth and change. *Journal of Educational Measurement*, *42*, 171–191.

Eckes, S. E., & Swando, J. (2009). Special education subgroups under NCLB: Issues to consider. *Teachers College Record, 111*(11), 2479-2504.

Francis, D. J., Shaywitz, S. E., Stuebing, K. K., Shaywitz, B. A., & Fletcher, J. M. (1996). Developmental lag versus deficit models of reading disability: A longitudinal, individual growth curves analysis. *Journal of Educational Psychology, 88*, 3–17. doi:10.1037/0022-0663.88.1.3

Galindo, C. (2010). English language learners' mathematics and reading achievement trajectories in the elementary grades. In E. García, & E. Frede (Eds.), *Developing the research agenda for young English language learners* (pp. 42–58). NYC: Teachers College Press.

Grimm, K. J. (2008). Longitudinal associations between reading and mathematics achievement. *Developmental Neuropsychology, 33*(3), 410–426.

Han, 2008

Hanushek, E., & Raymond, M. (2005). Does school accountability lead to improved student performance? *Journal of Policy Analysis and Management, 24*(2), 297-327.

Harcourt Educational Measurement (2002). *Metropolitan 8 Form V technical manual*. San Antonio, TX: Author.

Hart, B. H., & Risley, T. R. (1995). *Meaningful differences in the everyday experience of young American children*. Baltimore, MD: Paul H. Brookes.

Heck, R. (2006). Assessing school achievement progress: Comparing alternative approaches. *Educational Administration Quarterly, 42*, 667-699.

Hemphill, F. C., Vanneman, A., & Rahman, T. (2011). *Achievement gaps: How Hispanic and White students in public schools perform in mathematics and reading on the National Assessment of Educational Progress* (NCES 2011-459). Washington, DC: National Center for Education Statistics, Institute of Education Sciences, U.S. Department of Education.

Holman, L. J. (1995, April). *Impact of ethnicity, class, and gender on achievement of border area students on a high-stakes examination*. Paper presented at the annual meeting of the American Educational Research Association, San Francisco, CA.

Hyde, J. S., Fennema, E., & Lamon, S. J. (1998). Gender differences in mathematics performance: A meta-analysis. *Psychological Bulletin, 107*, 139–155. doi:10.1037/0033-2909.107.2.139

Jordan, N. C., Kaplan, D., & Hanich, L. B. (2002). Achievement growth in children with learning difficulties in mathematics: Findings of a two year longitudinal study. *Journal of Educational Psychology, 94*, 586–597.

Juel, C. (1988). Learning to read and write: A longitudinal study of 54 children from first through fourth grades. *Journal of Educational Psychology, 80*, 437–447.

Kieffer, M. J. (2008). Catching up or falling behind? Initial English proficiency, concentrated poverty, and the reading growth of language minority learners in the United States. *Journal of Educational Psychology, 100*, 851–868.

Kieran, C. (2007). Learning and teaching of algebra at the middle school through college levels. In F. K. Lester (Ed.), *Second handbook of research on mathematics teaching and learning* (pp. 707–762). Charlotte, NC: Information Age.

Lee, J. (2010). Tripartite growth trajectories of reading and mathematics achievement: Tracking national academic progress at primary, middle and high school levels. *American Educational Research Journal, 47*(4), 800–832.

Lee, V. E., & Burkham, D. T. (2002). *Inequality at the starting gate*. Washington, D.C.: Economic Policy Institute.

Leppanen, U., Niemi, P., Aunola, K., & Nurmi, J.-E. (2004). Development of reading skills among preschool and primary school pupils. *Reading Research Quarterly, 39*, 72–93.

Linn, R. (2008). Educational accountability systems. In K. Ryan & L. Shepard (Eds.), *The future of test-based educational accountability* (pp. 3-24). New York: Routledge.

Linn, R., & Haug, C. (2002). Stability of school-building accountability scores and gains. *Educational Evaluation and Policy Analysis, 24*, 29-36.

Lockwood, J. R. (2005). The (sometimes harsh) reality of longitudinal student achievement modeling. In R. Lissitz (Ed.), *Value-added models in education: Theory and applications*. Maple Grove, MN: JAM Press.

LoGerfo, L., Nichols, A., & Reardon, S. (2006). *Achievement Gains in Elementary and High School*. Washington, DC: Urban Institute.

Manna, P., & Ryan, L. L. (2011). Competitive grants and educational federalism: President Obama's Race to the Top program in theory and practice. *Publius: The Journal of Federalism, 41*(3), 522-546. doi:10.1093/publius/pjr021

Mazzocco, M. M. M., & Thompson, R. E. (2005). Kindergarten predictors of math learning disability. *Learning Disabilities Research & Practice, 20*,142–,155. doi:10.1111/j.1540-5826.2005.00129.x

McCoach, B. D., O'Connell, A. A., Reis, S. M., & Levitt, H. A. (2006). Growing readers: A hierarchical linear model of children's reading growth during the first 2 years of school. *Journal of Educational Psychology, 98*(1), 14–28.

McDonnell, L. M., McLaughlin, M. J., & Morison, P. (Eds.). (1997). *Educating one and all: Students with disabilities and standards-based reform*. Washington, DC: National Academy Press.

McGee, R., Williams, S., Share, D. L., Anderson, J., & Silva, P. A. (1986). The relationship between specific reading retardation, general reading backwardness and behavioral problems in a large sample of Dunedin boys: A longitudinal study from five to eleven years. *Journal of Child Psychology and Psychiatry, 27*, 597–610.

McGuinn, P. (2005). The national schoolmarm: No Child Left Behind and the new educational federalism. *Publius: The Journal of Federalism*, *35*(1), 41–68.

Molfese, V. J., Modglin, A.,&Molfese, D. L. (2003). The role of environment in the development of reading skills. *Journal of Learning Disabilities, 36,* 59–67.

Morgan, P. L., Farkas, G., & Wu, Q. (2009). Five-year growth trajectories of kindergarten children with learning difficulties in mathematics. *Journal of Learning Disabilities, 42,* 306–321.doi:10.1177/0022219408331037

Morgan, P. L., Farkas, G., & Wu, Q. (2011). Kindergarten children's growth trajectories in reading and mathematics: Who falls increasingly behind? *Journal of Learning Disabilities,44*(5) 472–488.

Muthén, B., & Khoo, S. T. (1998). Longitudinal studies of achievement growth using latent variable modeling. *Learning and Individual Differences, Special issue: latent growth curve analysis*, *10,* 73–101.

No Child Left Behind (NCLB) Act of 2001, Pub. L. No. 107–110, § 115, Stat. 1425 (2002).

North Carolina Department of Public Instruction [NCDPI]. (1997). Research on end-of-grade testing: Assessment of mathematics. *Assessment Brief*, *3*(4), 1–2.

Parrila, R., Aunola, K., Leskinen, E., Nurmi, J., & Kriby, J. R. (2005). Development of individual differences in reading: results from longitudinal studies in English and Finnish. *Journal of Educational Psychology*, *97*(3), 299–319.

Phillips, L. M., Norris, S. P., Osmond, W. C., & Maynard, A. M. (2002). Relative reading achievement: A longitudinal study of 187 children from first through sixth grades. *Journal of Educational Psychology*, *94,* 3–13.

Princiotta, D., Flanagan, K. D., and Germino Hausken, E. (2006). *Fifth Grade: Findings from the Fifth Grade Follow-up of the Early Childhood Longitudinal Study, Kindergarten Class of 1998-99 (ECLS-K)*. Washington, DC: U.S. Department of Education.

Public Schools of North Carolina. (1996a). North Carolina End-of-Grade Tests (Tech. Rep. No. 1). Raleigh: Public Schools of North Carolina, Department of Public Instruction, Division of Accountability Services.

Public Schools of North Carolina. (1996b). Setting annual growth standards: "The formula" (Accountability Brief No. 1–1). Raleigh: Public Schools of North Carolina, Department of Public Instruction, Division of Accountability Services.

Raudenbush, S. W., Bryk, A. S., Cheong, Y. F., Congdon, R. T., & du Toit, M. (2011). *HLM 7: Hierarchical Linear and Nonlinear Modeling*. Lincolnwood, IL: Scientific Software International.

Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis methods* (2nd ed.). Thousand Oaks, CA: SAGE.

Reardon, S. F., & Galindo, C. (2009). The Hispanic-White achievement gap in math and reading in the elementary grades. *American Educational Research Journal, 46*, 853–891. doi:10.3102/0002831209333184

Rescorla, L., & Rosenthal, A. (2004). Growth in standardized ability and achievement test scores from third to tenth grade. *Journal of Educational Psychology, 96*, 85–96.

Rogosa, D. (1979). Causal models in longitudinal research: Rationale, formulation, and interpretation. In J. R. Nesselroade & P. B. Baltes (Eds.). *Longitudinal research in the study of behavior and development* (pp. 263-302). New York, NY: Academic Press.

Sabornie, E. J., Cullinan, D., Osborne, S. S., & Brock, L. B. (2005). Intellectual, academic, and behavioral functioning of students with high-incidence disabilities: A cross-categorical meta-analysis. *Exceptional Children, 72,* 47–63.

Scarborough, H. S. (1998). Predicting the future achievement of second graders with reading disabilities: Contributions of phonemic awareness, verbal memory, rapid naming, and IQ. *Annals of Dyslexia, 48,* 115–136.

Scarborough, H. S., & Parker, J. D. (2003). Matthew effects in children with learning disabilities: Development of reading, IQ, and psychosocial problems from Grade 2 to Grade 8. *Annals of Dyslexia, 53,* 47–71.

Schulte, A. C. (2010). *Assessing growth for students with disabilities on large scale assessments.* Paper presented at the National Conference on Student Assessment, Detroit, MI.

Schulte, A. C. (2012, April). *Critical Issues for Examining Special Education Outcomes in Status and Growth Accountability Models*. Paper presented at the annual meeting of the National Council for Measurement in Education, Vancouver, BC, Canada.

Schulte, A. C., Osborne, S. S., & Erchul, W. P. (1998). Effective special education: A United States dilemma. *School Psychology Review, 27,* 66–76.

Schulte, A., & Villwock, D. (2004). Using high-stakes tests to derive school-level measures of special education efficacy. *Exceptionality, 12*, 107-127.

Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). *Experimental and quasi-experimental designs for generalized causal inference.* . Boston, MA: Houghton Mifflin Company.

Shaywitz, B. A., Holford, T. R., Fletcher, J. M., Stuebing, K. K., Francis, D. J., & Shaywitz, S. E. (1995). A Matthew effect for IQ but not for reading: Results from a longitudinal study. *Reading Research Quarterly, 30,* 387–388.

Shin, T., Davison, M. L., Long, J. D., Chan, C-K, & Heistad, D. (2013). Exploring gains in reading and mathematics achievement among regular and exceptional students using growth curve modeling. *Learning and Individual Differences, 23*, 92–100.

Stevens, J. (2005). The study of school effectiveness as a problem in research design. In R. Lissitz (Ed.), *Value-added models in education: Theory and applications*. Maple Grove, MN: JAM Press.

Tate, W. F. (1997). Race-ethnicity, SES, gender, and language proficiency trends in mathematics achievement: An update. *Journal for Research in Mathematics Education, 28*, 652–679.

Teddlie, C., & Reynolds, D. (2000). *The International handbook of school effectiveness research*. New York: Falmer Press.

The North Carolina Mathematics Tests Technical Report March, 2006 Mildred Bazemore, Section Chief, Test Development

Thissen, D., & Orlando, M. (2001). Item response theory for items scored in two categories. In D. Thissen & H. Wainer (Eds), *Test Scoring* (Pp. 73-140). Mahwah, NJ: Lawrence Erlbaum Associates.

Thissen, D., Pommerich, M., Billeaud, K., & Williams, V.S.L. (1995). Item response theory for scores on tests including polytomous items with ordered responses. *Applied Psychological Measurement, 19,* 39-49.

Thurlow, M., Altman, J., Cormier, D., & Moen, R. (2008). *Annual performance report: 2005-2006 state assessment data*. Minneapolis, MN: University of Minnesota, NCEO.

U. S. Department of Education. (2009). *Transcript of Race to the Top Technical Assistance Workshop in Baltimore, Maryland*. Washington, DC: Author.

U. S. Department of Education. (2010). *Race to the Top Program Guidance and Frequently Asked Questions*. Washington, D.C.: Author. http://www2.ed.gov/programs/racetothetop/faq.pdf

Wagner, M., Newman, L., Cameto, R., & Levine, P. (2006). *The academic achievement and functional performance of youth with disabilities: A report from the national longitudinal transition study-2 (NLTS2)* (NCSER 2006–3000). U.S. Department of Education, National Center for special education research. Washington, DC: U.S. Government Printing Office.

Wei, X., Blackorby, J., & Schiller, E. (2011). Growth in reading achievement in a national sample of students with disabilities ages 7 to 17. *Exceptional Children, 78,* 89–106.

Wei, X., Lenz, K. B., & Blackorby, J. (2012). Math Growth Trajectories of Students With Disabilities: Disability Category, Gender, Racial, and Socioeconomic Status Differences From Ages 7 to 17. *Remedial and Special Education*, published online 16 July 2012 DOI: 10.1177/0741932512448253

Willett, J. B., Singer, J. D., & Martin, N. C. (1998). The design and analysis of longitudinal studies of development and psychopathology in context: Statistical models and methodological recommendations, *Development and Psychopathology, 10*, 395-426.

Williams, V.S.L., Pommerich, M., & Thissen, D. (1998). A comparison of developmental scales based on Thurstone methods and item response theory. *Journal of Educational Measurement, 35*, 93-107.

Williamson, G. L., Appelbaum, M., & Epanchin, A. (1991). Longitudinal analysis of academic achievement. *Journal of Educational Measurement, 28,* 61–76.

Willms, J. D. (1992). *Monitoring school performance: A guide for educators*. London: Falmer Press.

Wright, W. E., & Li, X. (2008). High-stakes mathematics tests: How No Child Left Behind leaves newcomer English language learners behind. *Language Policy, 7*, 201–216.

Ysseldyke, J., & Bielinski, J. (2002). Effect of different methods of reporting and reclassification on trends in test scores for students with disabilities. *Exceptional Children, 68,* 189–200.

Zvoch, K., & Stevens, J. (2005). Sample exclusion and student attrition effects in the longitudinal study of middle school mathematics performance. *Educational Assessment, 10*(2), 105-123.

Authors' Notes

Joseph J. Stevens, Educational Methodology, Policy, and Leadership, University of Oregon; Ann Schulte, Learning Sciences Institute, Arizona State University.

This work was supported by the Institute of Education Sciences, U.S. Department of Education, through grant R32C110004 awarded to the University of Oregon. The opinions expressed are those of the authors and do not necessarily represent views of the Institute or the U.S. Department of Education.

Address all correspondence to the first author, Joseph J. Stevens, stevensj@uoregon.edu,
102 Education, 5267 University of Oregon, Eugene, OR 97403.

Table 1

Student Characteristics by Sample at Wave 1

	Total		Analysis Sample	
Characteristic	*N*	%	*N*	%
Student Group:				
Non-SWD	88,393	85.7	81,179	88.2
Regular Education	81,478	79.0	74,333	80.8
Academically Gifted	6,951	6.7	6,846	7.4
Autism	332	0.3	141	0.2 *
Emotional Disturbance	829	0.8	634	0.7 *
Intellectual Disability	2,017	0.2	1,244	1.4 *
Hearing Impairment	170	0.2	131	0.1 *
Multiple Disability	100	0.1	--	
Orthopedic Impairment	80	0.1	--	
Other Health Impairment	1,502	1.5	1,171	1.3
Specific Learning Disability	6,377	6.2	5,221	5.7
Speech-language Impairment	2,660	2.6	2,324	2.5
Traumatic Brain Injury	29	< 0.1	--	
Visual Impairment	59	< 0.1	--	
Unidentified	314	0.3	--	
Demographic Characteristic:				

Female	50,463	48.9	46,364	50.4
American Indian	1,549	1.9	1,353	1.5
Asian	1,958	1.9	1,791	1.9
Black	31,190	30.2	26,096	28.4
Hispanic	5,555	5.4	4,555	4.9
Multi-racial	1,818	1.8	1,639	1.8
White	61,005	59.2	56,611	61.5
Unidentified	17	< 0.1	--	
Limited English	3,553	3.4	2,724	3.0
Title I Student	4,661	4.5	3,827	4.2
Free/Reduced Lunch	40,189	39.1	37,266	40.5
Parental Education:				
< High School	12,158	11.8	9,532	10.4
High School	47,247	45.8	41,756	45.4
High School +	4,148	4.0	3,848	4.2
Community College Graduate	13,368	13.0	12,599	13.7
Trade/Business School Graduate	20,771	20.1	20,179	21.9
College Graduate	4,225	4.1	4,131	4.5
Total Sample Size	103,123		92,045	

* Difference between samples is statistically significant, *p* < 0.05.

Note. Non-SWD group is composed of the combination of Regular Education and Academically Gifted students.

Table 2

Mathematics Scale Score Means and Standard Deviations by Student Group (N = 92,045)

	Grade		
Student Group	3	4	5
All Students	251.44 (7.50)	257.39 (8.21)	263.22 (8.68)
Non-SWD	251.99 (7.31)	257.95 (8.07)	263.98 (8.33)
Regular Education	251.17 (6.91)	257.04 (7.61)	263.06 (7.91)
Academically Gifted	260.80 (5.55)	267.64 (6.30)	273.76 (6.12)
Autism	246.30 (7.76)	252.39 (8.61)	256.43 (10.59)
Emotional Disturbance	243.98 (7.06)	249.91 (7.13)	254.19 (8.59)
Intellectual Disability	237.56 (4.79)	243.25 (4.77)	246.68 (5.26)
Hearing Impairment	247.79	253.28	259.65

	(7.10)	(7.46)	(8.27)
Other Health Impairment	246.27	251.49	255.79
	(7.00)	(7.50)	(8.27)
Specific Learning Disability	247.03	252.23	257.22
	(6.80)	(7.08)	(8.07)
Speech-language Impairment	249.74	255.85	261.80
	(7.46)	(8.16)	(8.80)

Note. Non-SWD group is composed of the combination of Regular Education and Academically Gifted students.

Table 3

Two-level Longitudinal HLM Regression Models, Grades 3-5

	Unconditional		Specific Exceptionality		Exceptionality & Demographics	
Predictor	Intercept	Linear	Intercept	Linear	Intercept	Linear
Grand Mean	251.28 (0.03)	5.89 (0.01)	251.09 (0.03)	5.92 (0.01)	253.27 (0.04)	5.98 (0.02)
Academically Gifted			9.82 (0.07)	0.59 (0.03)	6.89 (0.07)	0.32 (0.03)
Autism			-5.86 (0.70)	-0.66 (0.31)	-7.84 (0.65)	-0.72 (0.31)
Emotional Disturbance			-7.54 (0.29)	-0.67 (0.15)	-5.18 (0.27)	-0.42 (0.15)
Intellectual Disability			-13.77 (0.18)	-1.41 (0.10)	-10.48 (0.19)	-1.08 (0.10)
Hearing Impairment			-3.52 (0.61)	-0.02† (0.23)	-3.98 (0.52)	-0.02† (0.23)
Other Health Impairment			-5.20 (0.21)	-1.10 (0.09)	-5.52 (0.19)	-1.04 (0.09)
Specific Learning Disability			-4.40 (0.10)	-0.80 (0.04)	-4.12 (0.09)	-0.71 (0.04)

Speech-language Impairment		-1.41 (0.16)	0.05† (0.05)	-1.77 (0.13)	0.09† (0.05)
Sex				-0.43 (0.04)	0.03† (0.02)
Limited English				-2.66 (0.16)	-0.10† (0.07)
Parental Education				1.23 (0.02)	0.14 (0.01)
Free Lunch				-1.24 (0.05)	-0.23 (0.02)
Asian				0.24† (0.16)	1.11 (0.07)
Black				-4.09 (0.05)	-0.03† (0.02)
Hispanic				-0.91 (0.12)	0.63 (0.05)
American Indian				-1.89 (0.17)	-1.01 (0.08)
Model *df*	91,567	91,559		91,551	
Δ Deviance	--	22079.04		26,910.77	
χ^2 (*df*, *p*-value)	--	(16, < 0.001)		(16, < 0.001)	

† Not statistically significant, *p* > .05. *Note*. Standard errors shown in parentheses.

Table 4

Mathematics Growth Effect Size Over Time by Student Group

	Grade Transition	
Student Group	3-4	4-5
Non-SWD	0.78	0.74
Regular Education	0.81	0.78
Academically Gifted	1.15	0.99
Autism	0.75	0.42
Emotional Disturbance	0.84	0.54
Intellectual Disability	1.19	0.67
Hearing Impairment	0.75	0.81
Other Health Impairment	0.72	0.54
Specific Learning Disability	0.75	0.66
Speech Language Impairment	0.78	0.70

Note. Non-SWD group is composed of the combination of Regular Education and Academically Gifted students.

Table 5

Mathematics Achievement Gap Effect Size by Exceptionality Category in Comparison to All Non-SWD students and to Regular Education Students

	Grade		
Student Group	3	4	5
vs. Non-SWD Students:			
Autism	-0.76	-0.68	-0.87
Emotional Disturbance	-1.07	-.98	-1.13
Intellectual Disability	-1.93	-1.79	-1.99
Hearing Impairment	-0.56	-0.57	-0.50
Other Health Impairment	-0.76	-0.79	-0.94
Specific Learning Disability	-0.66	-0.70	-0.78
Speech Language Impairment	-0.30	-0.26	-0.25
vs. Regular Education Students:			
Academically Gifted	+1.28	+1.29	+1.23
Autism	-0.65	-0.57	-0.76
Emotional Disturbance	-0.96	-0.87	-1.02
Intellectual Disability	-1.82	-1.68	-1.89
Hearing Impairment	-0.45	-0.46	-0.39
Other Health Impairment	-0.65	-0.68	-0.84
Specific Learning Disability	-0.55	-0.59	-0.67
Speech Language Impairment	-0.19	-0.14	-0.15

Table 6

Three-level Longitudinal HLM Regression Model, Grades 3-5

Fixed Effect	Coefficient	SE	*t*	*df*	*p*
Intercept, γ_{000}	253.44	0.076	3313.65	1110	<0.001
School Level Predictors					
Regular Education, γ_{001}	3.30	0.507	6.51	1110	<0.001
White, γ_{002}	0.70	0.357	1.95	1110	0.051
Free Lunch, γ_{003}	-0.30	0.560	-0.54	1110	0.591
Parental Education, γ_{004}	0.64	0.127	5.015	1110	<0.001
Limited English, γ_{005}	1.35	1.028	1.311	1110	0.190
School Size, γ_{006}	0.01	0.002	0.384	1110	0.701
Student Level Predictors					
Limited English, γ_{010}	-2.93	0.230	-12.771	55198	<0.001
Parental Education, γ_{020}	1.09	0.024	44.557	55198	<0.001
Sex, γ_{030}	-0.49	0.052	-9.515	55198	<0.001
Free Lunch, γ_{040}	-1.15	0.071	-16.187	55198	<0.001
Academically Gifted, γ_{050}	7.27	0.102	71.550	55198	<0.001
Emotional Disturbance, γ_{060}	-4.61	0.484	-9.527	55198	<0.001
Hearing Impairment, γ_{070}	-3.88	0.679	-5.714	55198	<0.001
Intellectual Disability, γ_{080}	-10.73	0.329	-32.622	55198	<0.001
Other Health Impairment, γ_{090}	-5.90	0.255	-23.166	55198	<0.001
Speech or Language Impairment, γ_{0100}	-1.58	0.176	-8.937	55198	<0.001

Specific Learning Disability, γ_{0110}	-4.26	0.128	-33.258	55198	<0.001
Autism, γ_{0120}	-7.80	0.795	-9.814	55198	<0.001
Asian, γ_{0130}	0.30	0.194	1.532	55198	0.125
Black, γ_{0140}	-3.76	0.081	-46.538	55198	<0.001
Hispanic, γ_{0150}	-0.73	0.170	-4.283	55198	<0.001
American Indian, γ_{0160}	-1.33	0.321	-4.158	55198	<0.001
Linear Slope, γ_{100}	6.04	0.033	183.488	1110	<0.001
School Level Predictors					
Regular Education, γ_{101}	-0.38	0.258	-1.306	1110	0.192
White, γ_{102}	-0.13	0.166	-0.789	1110	0.430
Free Lunch, γ_{103}	-0.52	0.256	-2.029	1110	0.043
Parental Education, γ_{104}	0.01	0.055	0.101	1110	0.919
Limited English, γ_{105}	0.63	0.602	1.041	1110	0.298
School Size, γ_{106}	-0.01	0.001	-2.558	1110	0.011
Student Level Predictors					
Limited English, γ_{110}	-0.02	0.084	-0.185	55198	0.853
Parental Education, γ_{120}	0.12	0.008	15.154	55198	<0.001
Sex, γ_{130}	0.02	0.019	0.872	55198	0.383
Free Lunch, γ_{140}	-0.14	0.027	-5.006	55198	<0.001
Academically Gifted, γ_{150}	0.25	0.043	5.845	55198	<0.001
Emotional Disturbance, γ_{160}	-0.16	0.218	-0.726	55198	0.468
Hearing Impairment, γ_{170}	-0.12	0.285	-0.413	55198	0.680

Intellectual Disability, γ_{180}	-0.98	0.149	-6.569	55198	<0.001
Other Health Impairment, γ_{190}	-0.99	0.111	-8.885	55198	<0.001
Speech or Language Impairment, γ_{1100}	0.05	0.061	0.848	55198	0.396
Specific Learning Disability, γ_{1110}	-0.66	0.051	-13.013	55198	<0.001
Autism, γ_{1120}	-0.72	0.359	-2.012	55198	0.044
Asian, γ_{1130}	0.94	0.075	12.632	55198	<0.001
Black, γ_{1140}	0.01	0.030	0.241	55198	0.810
Hispanic, γ_{1150}	0.56	0.060	9.429	55198	<0.001
American Indian, γ_{1160}	-0.25	0.131	-1.890	55198	0.059

Table 7

Correlation of School Level Characteristics with Estimated School Mathematics Intercept and Slope

	School Intercept	School Slope
Proportion Regular Education	-.155**	-.081**
Proportion White Students	.236**	.068*
Proportion Free Lunch Students	-.310**	-.119**
Parental Education	.309**	.091**
Proportion Limited English Students	-.003	.012
School Size	.172**	-.024

**. Correlation is significant at the 0.01 level (2-tailed).

*. Correlation is significant at the 0.05 level (2-tailed).

Figure Captions

Figure 1. Mean mathematics achievement by grade and exceptionality category.

Figure 2. School level estimated mathematics achievement trajectories by proportion of regular education students and level of parental education.

Figure 3. School level estimated mathematics achievement trajectories by proportion of free lunch students and school size.

Figure 4. Scatterplot of school average intercepts and slopes by proportion of regular education students.

Figure 1

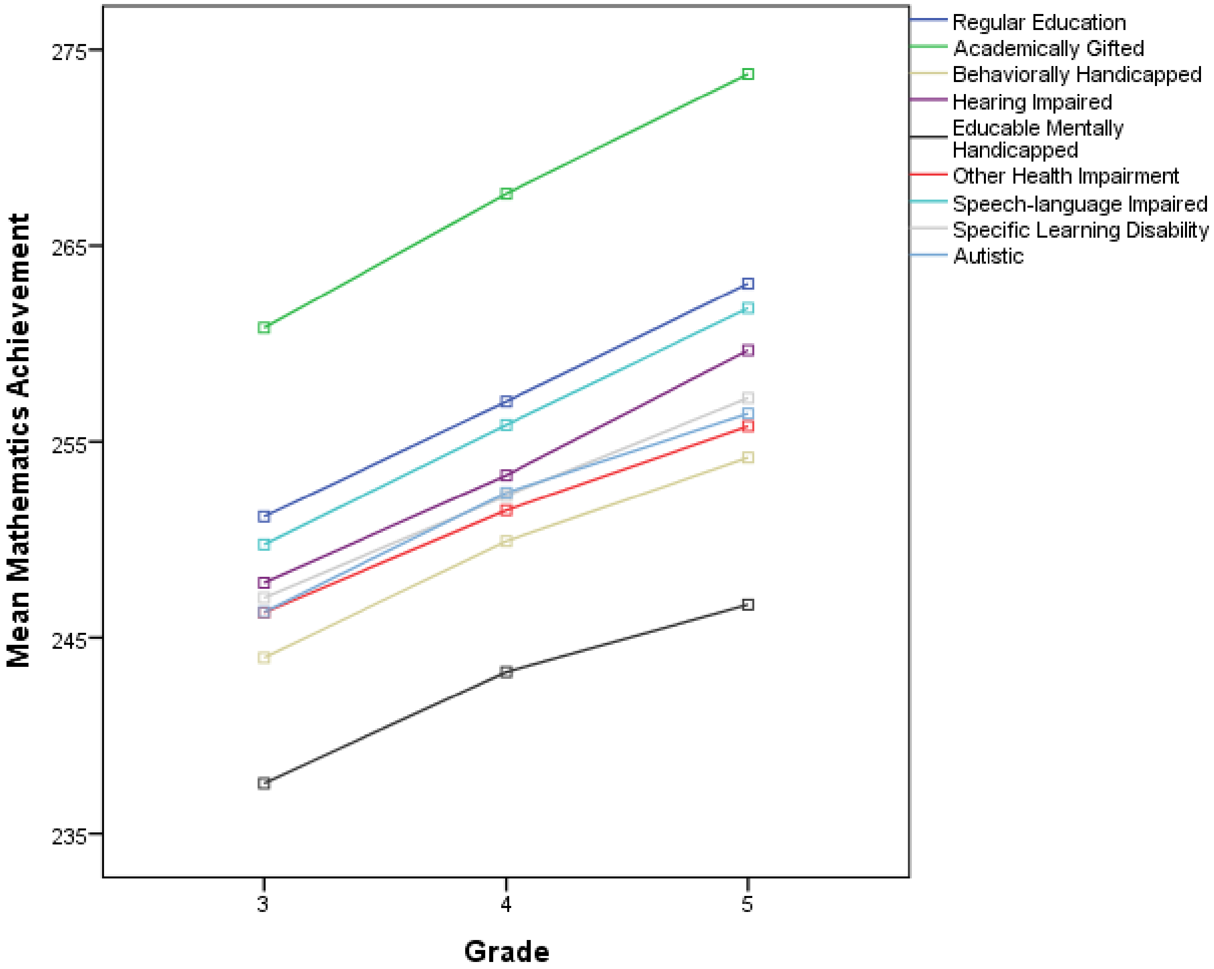

Mean Mathematics Achievement
275
265
255
245
235
3
4
5
Grade
Regular Education
Academically Gifted
Behaviorally Handicapped
Hearing Impaired
Educable Mentally Handicapped
Other Health Impairment
Speech-language Impaired
Specific Learning Disability
Autistic

Figure 2

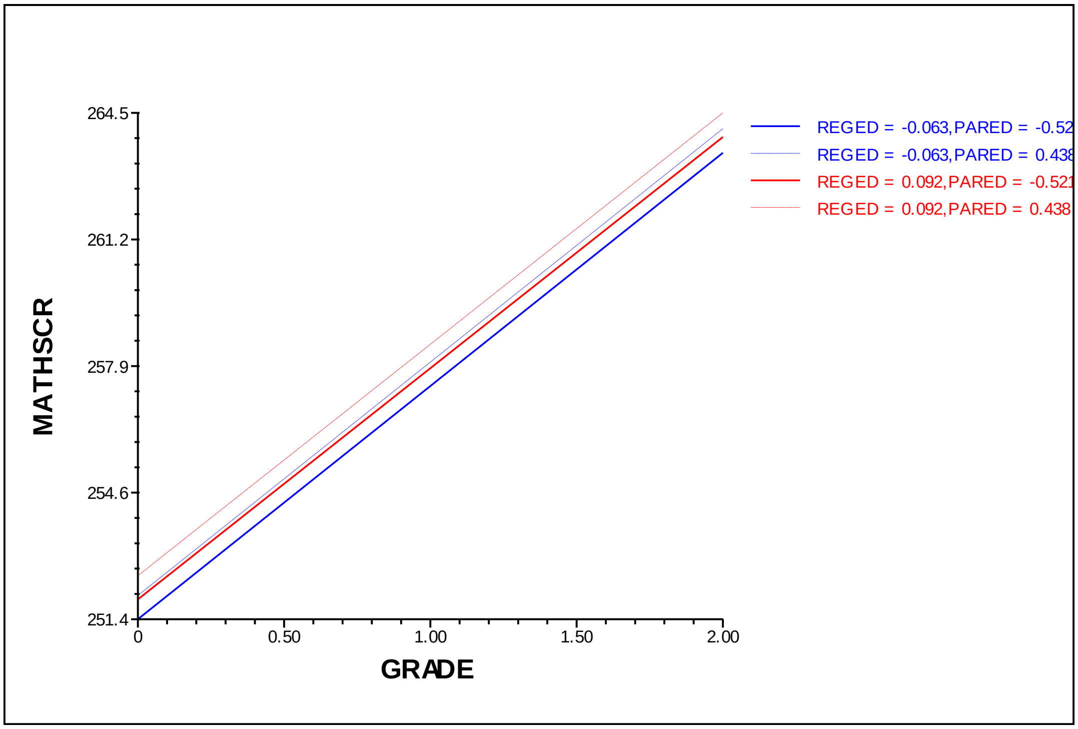

264.5
261.2
257.9
254.6
251.4
MATHSCR
0
0.50
1.00
1.50
2.00
GRADE
REGED = -0.063, PARED = -0.52
REGED = -0.063, PARED = 0.438
REGED = 0.092, PARED = -0.521
REGED = 0.092, PARED = 0.438

Figure 3

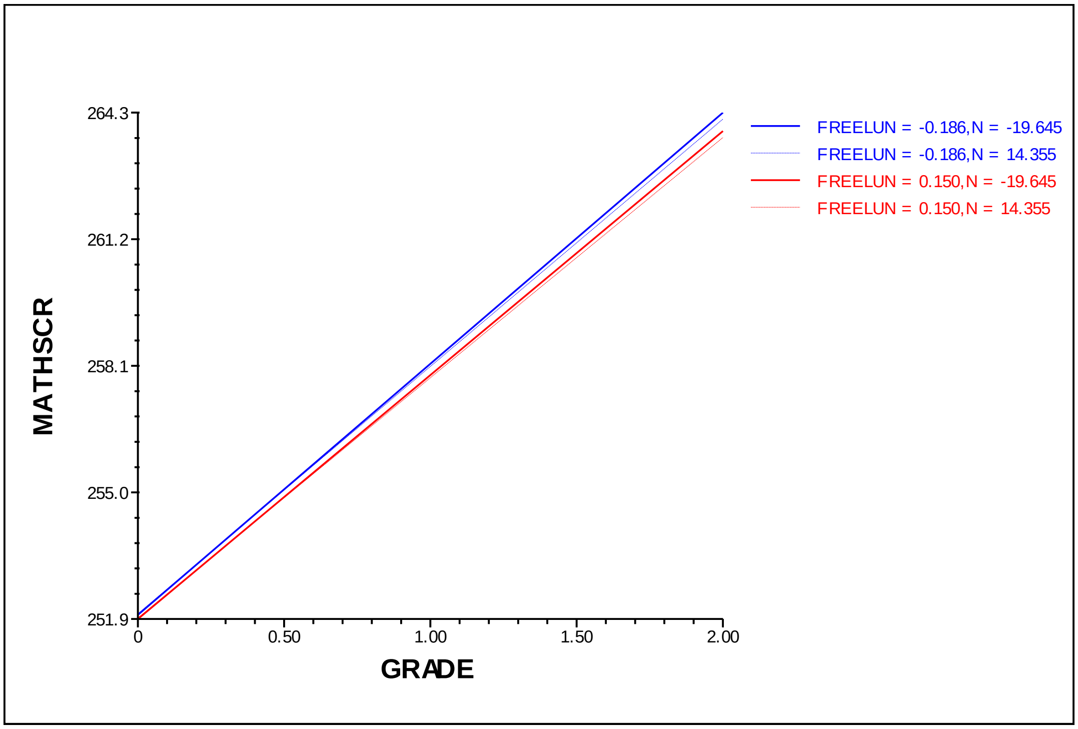
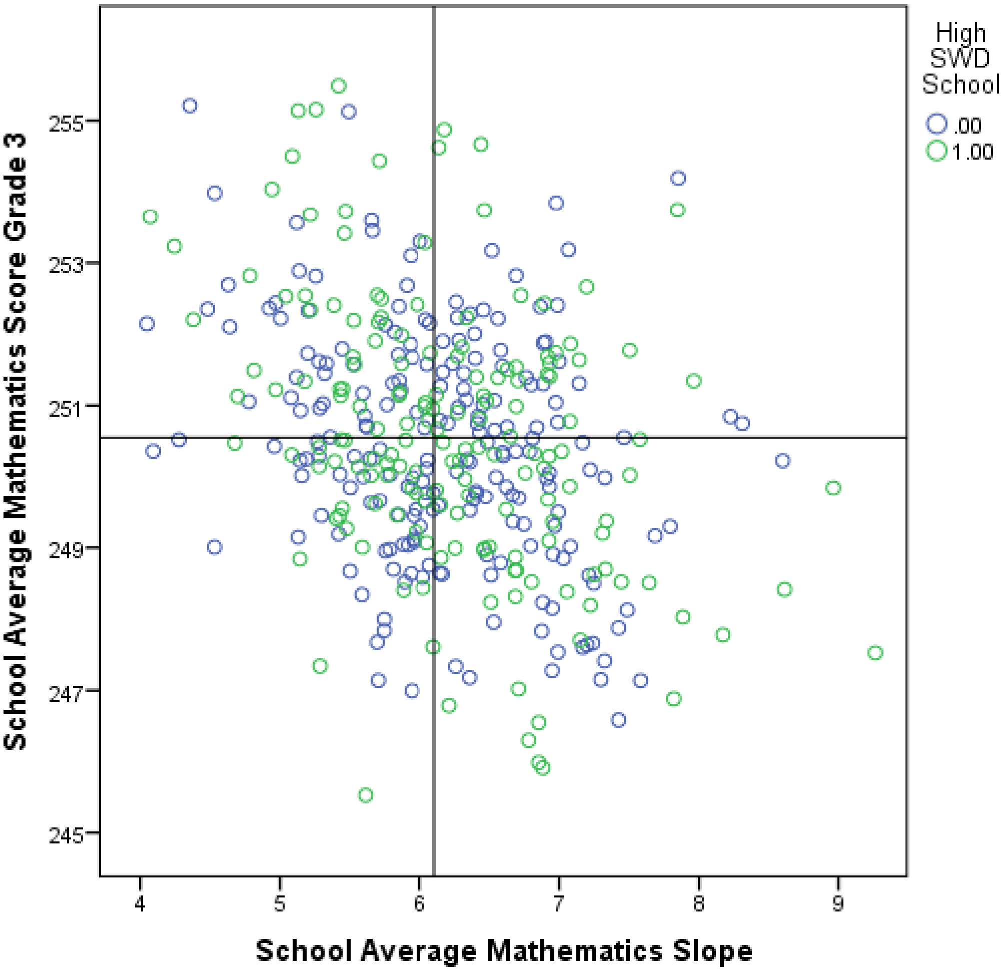

264.3
261.2
258.1
255.0
251.9
MATHSCR
0
0.50
1.00
1.50
2.00
GRADE
FREELUN = -0.186,N = -19.645
FREELUN = -0.186,N = 14.355
FREELUN = 0.150,N = -19.645
FREELUN = 0.150,N = 14.355

Figure 4


School Average Mathematics Score Grade 3
255
253
251
249
247
245
School Average Mathematics Slope
4
5
6
7
8
9
High SWD School
○ .00
○ 1.00