Achievement Gaps for Students with Disabilities: Stable, Widening, or Narrowing on a State-wide Reading Comprehension Test?

*Journal of Educational Psychology, 16*, 925-942.

© 2016 American Psychological Association  
This paper is not the copy of record and may not exactly replicate the authoritative document published in the APA journal. Please do not copy or cite without the author's permission. The final article is available at [http://dx.doi.org/10.1037/edu0000107](http://dx.doi.org/10.1037/edu0000107).

This research was funded in part by a Cooperative Service Agreement from the Institute of Education Sciences (IES) establishing the National Center on Assessment and Accountability for Special Education (NCAASE; PR/Award Number R324C110004). The findings and conclusions do not necessarily represent the views or opinions of the U.S. Department of Education.

Correspondence concerning this article should be addressed to Ann C. Schulte, Sanford School of Social and Family Development, Arizona State University, 308J Payne Hall, PO Box 873701, Tempe, AZ 85287-3701.  
E-mail: ann.schulte@asu.edu
Abstract
Reading comprehension growth trajectories from third to seventh grade were estimated for 99,919 students on a state reading comprehension assessment. We examined whether differences between students in general education (GE) and groups of students identified as exceptional learners were best characterized as stable, widening, or narrowing. The groups included students with disabilities (SWD) from eight exceptionality groups and two groups of academically gifted students (AG). Initial reading comprehension achievement differed for all exceptionalities. Controlling for socio-demographic variables, small, but statistically significant differences in growth rate were observed, with SWD groups growing more rapidly and AG groups growing more slowly than GE students. Given that differences in growth for SWD were small relative to the magnitude of the initial achievement gaps, the observed pattern of growth was one of stable differences. There was evidence of some narrowing of the achievement gap for students identified with learning disabilities in reading. The findings were interpreted within the simple view of reading where increases in word recognition skills for SWD in the grade range examined may have accounted for their more rapid growth in reading comprehension relative to GE students. The findings suggest that similar expectations for rate of reading growth for GE students and SWD might be incorporated into growth-based accountability models, but they also suggest that reading comprehension growth sufficient to have an impact on SWD achievement gaps does not routinely occur in typical educational practice.
Keywords: reading comprehension, achievement gaps, hierarchical linear modeling, students with disabilities, Matthew effect, gifted students, longitudinal research
Achievement Gaps for Students with Disabilities: Stable, Widening, or Narrowing on a State-wide Reading Comprehension Test?

The purpose of this study was to examine reading comprehension achievement growth and gaps across Grades 3 to 7 for students with disabilities (SWD) in comparison to students in general education (GE) and students identified as academically/intellectually gifted (AG). A particular focus was the developmental pattern of individual differences observed for reading comprehension achievement by exceptionality, and whether the differences between the focal groups and a comparison group of GE students were best characterized as increasing, decreasing, or remaining stable across the grade span examined.

Reading comprehension is widely considered "the essence of reading" (Durkin, 1993) and a critical outcome of schooling (National Institute of Child Health and Human Development, 2000). SWD often encounter difficulty acquiring reading skills (Blackorby et al., 2005). For example, on the 2013 National Assessment of Educational Progress (NAEP; U.S. Department of Education, 2014), in comparison to students without disabilities (SWoD), much lower percentages of SWD reached the "proficient" level or above in reading in Grade 4 (10% vs. 38%), Grade 8 (7% vs. 39%), or Grade 12 (8% vs. 40%).

Historically, obtaining a comprehensive picture of SWD reading achievement growth and gaps across grades has been problematic because many of these students have been excluded from large scale achievement testing programs (Koretz & Hamilton, 2006; McDonnell, McLaughlin, & Morison, 1996); excluded from longitudinal studies of reading growth (e.g., Huang, Moon, & Boren, 2014); or included in longitudinal studies, but without their growth examined separately (e.g., Rescorla & Rosenthal, 2004). Although cross sectional depictions of the achievement gap for SWD indicate that the gap widens across grades (e.g., Chudowsky, Chudowsky, & Kober, 2009; U.S. Department of Education, 2014), such depictions are unlikely to represent the gaps that would be observed if students were followed longitudinally. Students’ entrances and exits from special education are related to their achievement with lower achieving students in general education entering special education and higher achieving students in special education exiting, and this pattern affects the size of observed SWD achievement gaps (Schulte & Stevens, 2015; Ysseldyke & Bielinski, 2002). A second issue in obtaining accurate information about longitudinal achievement gaps for SWD is that the group encompasses students who are receiving special education in 13 disability categories where the effect of the disability on reading achievement varies considerably (Blackorby et al., 2005; Wei, Blackorby & Schiller, 2011). Treating SWD as a single group may mask differences in reading achievement trajectories that have implications for intervention as well as for accountability policies (Buzick & Laitusis, 2010; Temple-Harvey & Vannest, 2012; Wei et al., 2011).

Although students who are AG are not at risk for low achievement, obtaining a comprehensive picture of reading achievement growth for this group of students also presents challenges. Unlike SWD, AG students generally have been included in longitudinal studies of achievement growth, but only a few studies have examined growth for these students separately (e.g., Rambo-Hernandez & McCoach, 2015). Advocacy groups have expressed concern that the current focus on grade level proficiency in the No Child Left Behind Act of 2001 (NCLB, 2002) may result in less focus on enhancing achievement outcomes for students who are AG because they are likely to score above proficiency standards in each grade (Council for Exceptional Children, 2010; National Association for Gifted Children, 2014). These advocacy groups have argued for an increased focus on achievement growth rather than status in accountability models. A similar argument has been made for a focus on achievement growth for SWD because a
substantial number of students in this group may score far enough below the cut-point for grade-
level proficiency that improved outcomes may go unrecognized when the accountability focus is
on the single cutpoint for grade level proficiency (Buzick & Latusis, 2010).

Given the increasing interest in using achievement growth as a key outcome in school
accountability models (e.g., Hoffer et al., 2011), lack of information about reading achievement
growth for children with exceptionalities is problematic, whether they are SWD or students who
are AG. The potential of achievement growth to offer more fair and valid measures of
achievement progress depends on a normative understanding of student achievement growth
including the nature and likely range of inter-individual differences in observed growth and how
these inter-individual differences change across grades. When the growth expectations
incorporated into school accountability policies are not based on empirical evidence, policy
validity is threatened (Harris, 2009; Lee, 2004) given that inferences about teacher and school
performance may be inaccurate.

Although an accurate picture of reading achievement growth across grades and how it
differs among groups of children has implications for school accountability models, it also can
inform models of the development of reading. With the wide-scale implementation by U. S.
states of annual testing in reading and mathematics across Grades 3 to 8, datasets are now
available where the achievement of large numbers of students can be tracked across grades, often
using vertical scales (Dadey & Briggs, 2012). One potential use of these datasets is an
examination of how different theoretical predictions of individual and group performance (e.g.,
Baumert, Nagy, & Lehman, 2012; Stanovich, 1986) fit with the observed growth trajectories for
children participating in annual testing programs. Such datasets also can provide descriptions of
achievement growth for groups of students where longitudinal studies have been scarce and often
conducted on small samples (e.g., Scarborough & Parker, 2003).

**Theory and Research on Reading Comprehension Growth**

The "simple view of reading" is the premise that reading comprehension is the joint product
of word identification and language (listening) comprehension (Gough & Tunmer, 1986).
Although it is not a complete account of the skills underlying reading comprehension (Perfetti,
Landi & Oakhill, 2005; Vellutino, Tunmer, Jaccard, & Chen, 2007), as a general framework it is
useful for characterizing reading development and the nature of reading difficulties, as well as
identifying key areas for instruction and remedial interventions (e.g., Compton, Miller, Ellemem,
& Steacy, 2014; García & Cain, 2013). Within the simple view of reading, word identification
and language comprehension are viewed as largely independent contributors to reading
comprehension, determined by underlying skills (e.g., phonological awareness and rapid
decoding for word identification; semantic and syntactic knowledge for language
comprehension) that have limited overlap (Vellutino et al., 2007). The relative contribution of
the two components varies at different stages of reading acquisition (Tighe & Schatschneider,
2014). Language comprehension plays a smaller role until readers develop enough facility in
word identification to be able to fluently decode text at or near their ability to understand spoken
language--at age 9 or 10 for most students (García & Cain, 2013; Vellutino et al., 2007).

The simple view of reading has two important implications for the study of reading
achievement growth and gaps for SWD. First, it follows from the model that there are two major
sources of reading difficulties for children, deficits in word identification and deficits in language
comprehension, which can occur separately or together (Compton et al., 2014). These
difficulties are likely to be distributed differently among children in the various exceptionalities
served in special education, depending on the nature and severity of the disability. For example,
most students with learning disabilities (LD) in reading have marked difficulties in word recognition, but adequate language comprehension skills (Vellutino, Fletcher, Snowling, & Scanlon, 2004). In contrast, students with intellectual disabilities show impairments in both areas (Wise, Sevcik, Romske, & Morris, 2010).

Second, as the relative influence of word identification and language comprehension on reading comprehension changes, it is likely to affect the rate of reading comprehension growth. When children acquire foundational skills in decoding, their sight word vocabularies expand rapidly (Ehri, 2005; Ehri & Snowling, 2004), removing word identification as an initial bottleneck and allowing rapid growth in reading comprehension (Scarborough & Parker, 2003). However, when word identification skills have reached a level commensurate with listening comprehension skills, reading comprehension growth is likely to slow as skills such as inference making, comprehension monitoring, and lexical knowledge become important determinants of further reading growth (Oakhill, Cain, & Bryant, 2003; Perfetti et al., 2005). With regard to SWD, if the disability slows initial acquisition of decoding skills, but students eventually acquire them, rapid growth may occur, but at a later point than for SWoD. To the extent that the disability also affects language comprehension skills, if students' language comprehension skills are compromised, then growth may slow again.

**Reading growth across grades.** Several studies have made use of large national datasets to investigate the nature of reading growth across the school years (e.g., Bloom, Hill, Black, & Lipsey, 2008; Lee, 2010). The general finding is that reading growth is curvilinear with large increases in the early grades that decelerate as students progress through school. For example, Lee (2010) made use of long-term NAEP data, several national datasets, and national norms from standardized achievement tests to examine the nature of student growth in reading and mathematics. He characterized national achievement growth trajectories as having "remarkable consistency and stability across different tests and cohorts over the long term" (Lee, 2010, p. 825).

**Individual differences in reading growth and developmental patterns.** In individual differences research, investigators seek to understand the sources of group and individual variation that occur within a general developmental pattern (Pennington, 2002). In terms of individual differences in reading growth, three general developmental patterns have been described (Pfost et al., 2014). The first is a pattern where growth in students with higher initial literacy levels outpaces growth for students with lower initial literacy levels, resulting in widening achievement gaps over time. This pattern is often termed a Matthew effect, or a cumulative growth or fan-spread pattern (Morgan, Farkas, & Hibell, 2008; Stanovich, 1986). The second pattern, termed a compensatory growth or fan-close pattern (Francis, Shaywitz, Stuebing, Shaywitz, & Fletcher, 1996), is one where students with lower initial literacy levels show more growth than students with higher levels of initial literacy, resulting in narrowing of the achievement gap over time. The final pattern is one of stable differences among students where growth for students at different initial achievement levels remains parallel over time.

Although different mechanisms are thought to produce each of these developmental patterns (Pfost et al., 2014), there are also multiple mechanisms that might underlie each of the patterns. In terms of the fan-spread pattern, Stanovich (1986) proposed that reciprocal relationships between reading development and the factors enhancing it result in widening individual differences in reading achievement over time. Many of the self-reinforcing relationships he described were the result of differences in children's emergent literacy skills altering their motivation to read and opportunity to practice reading skills, ultimately affecting the development of new vocabulary and other reading skills, and resulting in accelerated or slowing
growth (Pfost et al., 2014). However, other mediating mechanisms for the fan-spread pattern are also possible (Baumert et al., 2012; Morgan et al., 2008; Scarborough & Parker, 2003). For example, Baumert et al. (2012) distinguished between individual- and status-driven fan-spread effects, with cognitive or behavioral characteristics producing cumulative effects at the individual level, but status-driven effects resulting from advantages or disadvantages afforded different groups (e.g., an association between student poverty and lower quality school environments that has a cumulative impact on reading achievement).

Possible explanatory mechanisms for fan-close effects in reading include (a) developmental lags that narrow individual differences as the initially lower group catches up and initially higher achieving students’ growth plateaus (Francis et al., 1996); (b) insufficient opportunities to learn for initially higher achieving students resulting in slowing growth (Rambo-Hernandez & McCoach, 2015); or (c) strong compensatory or remedial education services for initially low achieving students (Baumert et al., 2012). Finally, for the stable differences pattern, Baumert et al. proposed two possible mechanisms. One was that growth at each time period was not a cumulative result of all previous reading experiences but simply a function of achievement at the immediately previous time point. The second possibility was that multiple growth influences could be simultaneously operative and moving in opposite directions, such as a fan-close pattern influencing individual development and a status-driven pattern producing a fan-spread effect, together producing a pattern of stable differences across time.

In a meta-analysis that examined the research evidence for Matthew effects in reading across 25 years, Pfost et al. (2014) found no support for the existence of a single developmental pattern that characterized reading skill growth. Instead, there was evidence that developmental patterns varied by the type of reading skill assessed. For example, constrained skills, such as letter knowledge, phonics, and concepts of print, were more likely to show a fan-close developmental pattern. Decoding speed was more likely to show a fan-spread or stable differences pattern, and reading comprehension was more often associated with a stable differences or fan-close pattern. Pfost et al. (2014) also found that methodological features of studies affected the developmental pattern observed. Use of reading measures with lower reliability (< .90) or ceiling or floor effects were more likely to result in a fan-close pattern, an indication that measurement error and regression toward the mean could be contributory factors when this pattern is observed.

**Methodological considerations in examining achievement growth patterns.** Pfost et al. (2014) specifically excluded studies focusing on SWD from their meta-analysis; however, their study has several implications for studies of SWD growth patterns and the design of the present study. First, the finding that developmental patterns were related to the reading skill studied highlights the importance of examining reading growth with measures tapping individual reading skills (i.e., reading vocabulary, decoding, reading comprehension) rather than a composite measure. This is particularly important given that individual component skills have different relationships to socio-demographic and home literacy variables (Hecht, Burgess, Torgesen, Wagner, & Rashotte, 2000) and the relative weighting of component skills within composite measures changes across grades. Use of a composite reading measure may explain why studies of longitudinal achievement gaps for different socio-demographic groups sometimes show abrupt shifts in the pattern of individual differences across grades (e.g., Chatterji, 2006; Kieffer, 2012).

Second, use of a reading measure with high reliability and adequate floor and ceiling is important to prevent misattributing testing artifacts to developmental patterns in the distribution of individual differences over time. Although not specifically mentioned by Pfost et al. (2014), a related complication in understanding patterns of academic growth is the dependence of interpretations on the adequacy of the developmental scale (Briggs & Weeks, 2009). There are a
number of technically sound approaches to creating vertical scales that provide a basis for scale linkage—usually a common set of examinees over grades or embedded linking items at two or more grade levels (see Kolen & Brennan, 2004). Nonetheless, interpretation of patterns of growth can easily be influenced by details of the scaling model applied, the design and number of linking items, differences in content specification over time, or lack of correspondence between the IRT scale and the underlying theoretical ability dimension represented (Bolt, Deng, & Lee, 2014; Zwick, 1992).

Third, as discussed by Baumert et al. (2012), different student characteristics may be associated with different developmental patterns operative at the same time. Therefore, controlling for covariates is important when examining developmental patterns of reading growth associated with particular student characteristics. A number of student characteristics associated with lower reading comprehension achievement or growth, such as being male; economically disadvantaged; lacking English proficiency; or being of Black, Hispanic, or American Indian race/ethnicity also vary by SWD status and by specific exceptionality (Wei et al., 2011). Including socio-demographic variables as controls when these characteristics are not the primary focus of study is important to avoid confounding group differences due to the characteristic of interest versus group differences that result from differential socio-demographic composition among groups (Morgan et al., 2011).

**Reading growth and longitudinal achievement gaps for SWD.** Although a number of investigators have examined reading comprehension achievement growth trajectories for children as a function of demographic characteristics (Keiffer, 2012; Huang et al., 2014), or lower and higher initial reading achievement (e.g., Protopapas Sideridis, Mouzaki, & Simos, 2011), surprisingly few investigators have examined reading comprehension growth and gaps across grades for SWD. To date, only three published studies have (a) examined growth in overall reading or reading comprehension achievement, (b) included SWD and SWoD, and (c) also controlled for socio-demographic differences (Francis et al., 1996; Judge & Bell, 2010; Morgan et al. 2011). Across these three studies, only two exceptionality groups were examined: students with LD, included in all three studies; and students with speech-language impairments (SLI), included in one study (Judge & Bell, 2010). Each of the studies used composite reading measures (although Francis et al. reported their results remained the same when decoding and reading were examined separately). Two of the studies, Judge and Bell, and Morgan et al., used school-identified students with LD, included students with LD in any academic area, and examined growth from Grades K to 5. Francis et al. used researcher-identified students with LD in reading only, based on an IQ/achievement discrepancy at third grade, and examined reading growth across Grades 1 to 9.

Although all three studies found students with LD had lower initial reading achievement compared to SWoD, only Judge and Bell (2010) found the fan-spread pattern predicted by Stanovich (1986). Morgan et al. (2011) and Francis et al. (1996) both found stable difference patterns with achievement gaps remaining similar across the grade spans studied. In the one study including students with SLI, Morgan et al. (2011) found that students with SLI showed lower kindergarten reading achievement than the reference group and fell further behind across grades. A notable feature of the Morgan et al. study was the use of additional control variables beyond student demographic characteristics in successive models. When a teacher rating of students' "approach to learning" (i.e., attentiveness, task persistence, eagerness to learn, adaptability, and organization) was included in the model, the differences in intercepts between the two disability groups and the reference group dropped substantially. Teacher ratings on the
approaches to learning measure also were positively related to reading achievement intercept and linear growth.

A small number of additional studies also examined longitudinal reading achievement gaps for students with LD or SLI, but did not control for differences in socio-demographic variables. Findings from these studies have been mixed, with two reporting substantial reductions in the achievement gap across grades; Scarborough and Parker (2003) on both a reading composite and comprehension measure for students with LD, and Skibbe et al. (2008) using a composite reading measure and tracking students with language difficulties identified prior to school entry. One study reported stable achievement gaps for both word recognition and reading comprehension for students with language impairments (Catts, Bridges, Little, & Tomblin, 2008), and one reported a widening achievement gap in reading comprehension for students who were LD (McKinney & Feagans, 1984). In sum, studies of longitudinal achievement gaps in reading for SWD are quite limited, have been restricted to only two specific exceptionality, and do not consistently report a fan-spread pattern.

One additional study by Wei et al. (2011) is relevant to the present study because it addressed reading growth for students in a much broader range of exceptionality categories; however, it did not include a comparison group of SWoD. Using students with LD as the reference group, Wei et al. examined level of reading achievement and curvilinear growth for students in 10 of the 12 remaining exceptionality categories recognized in the Individuals with Disabilities Education Act (IDEA, 2004). In terms of reading comprehension, all exceptionality groups showed curvilinear growth in reading comprehension achievement from age 7 to 17. Level of reading comprehension achievement (at the average age of 12.67) differed for nine of the ten exceptionalities, with four groups scoring significantly lower than students with LD (students with intellectual disabilities, multiple disabilities, autism, and hearing impairment), and five significantly higher (students with orthopedic impairment, emotional disturbance, other health impairment, visual impairment, and SLI). Linear change coefficients were largely comparable across exceptionality groups, although some groups had small but significantly lower linear slope coefficients. None of the exceptionalities differed significantly from students with LD in terms of quadratic curvature.

**Reading growth for students who are AG.** As noted earlier, research on achievement growth for the AG exceptionality group is quite limited. Warne (2014) used above-level achievement testing to examine academic growth in gifted students. Compared to annual expected growth based on the reading test norms for students who were two grades higher than the students in the study, he found that male students who were AG made less or similar growth on a reading composite test and females who were AG made more growth. In another study, Rambo-Hernandez and McCoach (2015) examined reading growth from third to sixth grade for students who had scored in the top 2% of students nationally on a composite measure of reading achievement in kindergarten. Compared to students who had average reading achievement in kindergarten, AG students grew more slowly during the academic year in Grades 3 to 6, but more rapidly than the average-achieving students during the summer. Rambo-Hernandez and McCoach also found that gifted students’ reading growth in Grades 4 and 5 increased relative to their growth rate in third grade, a pattern different than the curvilinear growth pattern observed for other student populations. This result suggests that the pattern of reading growth for gifted students may differ in functional form from the pattern observed for most students (e.g., Bloom et al., 2008).
Study Purpose and Research Questions

In summary, there is increasing interest in incorporating measures of student growth into programs that report student achievement outcomes for monitoring or accountability purposes. However, there are only a limited number of studies examining reading growth and gaps, with no published studies that have (a) examined a broad range of exceptionality groups identified in federal policy (IDEA, 2004), (b) included a comparison group without disabilities, and (c) controlled for student socio-demographic characteristics. The component skills that account for growth in reading comprehension differ across grades, with growth in decoding or word identification skills likely to be a more important factor in reading comprehension growth in the early grades, and growth in language comprehension and higher level cognitive skills more important in later grades. This developmental change makes the use of composite measures in studying reading growth problematic, but may also affect the apparent pattern of observed growth in reading comprehension when student groups differ markedly in their word identification skills and as a result what component skills account for growth during that age span. Lastly, few reading growth studies have used states’ large-scale reading assessments, the primary outcome measure for student achievement in the NCLB (2002) school accountability framework. The purpose of this study was to address two fundamental questions concerning achievement growth in reading comprehension for SWD:

1. Controlling for demographic differences between groups, what is the developmental progress in reading comprehension for GE students and students in specific exceptionality groups (including AG students) on a statewide achievement test used for accountability purposes?

2. Do SWD and AG students show a fan-spread, fan-close, or stable growth pattern in reading comprehension from third to seventh grade or any changes in reading achievement gaps relative to GE students?

Method

Sample

The initial sample for this study was all North Carolina students who were in the third grade in the 2002-2003 school year, had participated in end of grade achievement testing, and had not been retained in third grade from the previous year (N = 101,885; see first column of Table 1 labeled “Total sample”). The analytic sample was created by excluding students who did not have a unique identifier in 2003 (N = 5, < 0.1%); (b) did not have complete demographic information in Grade 3 (N = 27, < 0.1%); or (c) had never participated in the large scale reading test in Grades 3 to 7 (N = 1,772, 1.7%). After all these exclusions had been applied, the number of students in some exceptionality categories (i.e., multiple disabilities, orthopedic impairment, traumatic brain injury, and visual impairment was 100 or less and too small to ensure stable statistical estimation, and students from these categories were excluded (N = 162, < 0.2%). When all students meeting one or more of these exclusion criteria had been eliminated, the analytic sample consisted of 99,919 students (98.1% of the students in the state test data file). Characteristics for these students are provided in Table 1 under “Analytic sample”). The percent of SWD in the total sample (14.4%) was slightly higher than the percent of students served in public education nationally (12.9%), and the proportions of SWD served within the different exceptionality categories were generally comparable to national figures except that the proportion of students identified with intellectual disabilities was somewhat higher and the

To create the longitudinal sample, the students who were present in the database in 2002-2003 in Grade 3 were matched to all succeeding years of test data through Grade 7 (2006-2007), when the state introduced a new edition of the reading test. Of the 99,919 students in the analytic sample, 80.5% had reading scores in all five years, 6.7% had scores in four years, 5.1% had scores in three years, 3.6% had scores in two years, and 4.1% had one reading score during the five-year study period. Reasons for missing achievement data included student absence, administration of an alternate assessment in that year, or leaving the state school system.

We examined the extent to which the analytic sample differed from the total sample in terms of representation of the different exceptionalities using z-tests of the difference between two proportions. Given the sample size, even quite small differences between the total and analytic sample were statistically significant ($p < .05$) and a measure of effect size (ES), Cohen's $h$, was therefore used as a means to interpret differences. Cohen (1988) suggested that in the absence of knowledge of the range of typical ES values found in an area of study, $h$ values of .20 be considered small, .50 as medium, and .80 and greater as large. Given these guidelines, all differences in proportions of SWD were quite small, ranging from 0.01 to 0.07. The largest was the ES for the proportion of students with intellectual disabilities represented in the analytic versus total sample ($ES = .07$). Just over 40% of the students with intellectual disabilities never participated in the large-scale reading achievement testing, instead taking an alternate assessment.

We also examined how the total sample differed from the analytic sample in terms of socio-demographic characteristics, and how the SWoD and SWD groups within the analytic sample differed from each other (see Table 2). The SWD group had a higher proportion of males than the SWoD group (67% versus 49%, $h = 0.38$) and more students participating in the free lunch program, an indicator of economic disadvantage (48% versus 37%, $h = .24$). These ES were not small and reflect the higher likelihood that males and children in poverty are placed in special education (U.S. Department of Education, 2007).

**Measures**

For all analyses, the outcome measure was student developmental scale score on the standardized, second edition North Carolina End of Grade Reading Comprehension Tests (EOG-RC) at the grade level in which the students were placed that year. A technical manual for the second edition was published by the North Carolina Department of Public Instruction (NCDPI, 2004a) and provides information on the test construction process, test reliability and validity, and the procedures used to construct the developmental scale. At each grade level, there were three alternate forms of the test, each consisting of 50 to 56 multiple-choice items intended to measure the four strands in the state English/Language Arts curriculum: (a) cognition, (b) interpretation, (c) critical stance, and (d) connections. Average internal consistency reliability estimates across forms were above .90 for Grades 3 to 8, with standard errors of measurement of 2 to 3 developmental scale score points for the majority of respondents scoring within two standard deviations of the grade level mean, and as large as 6 points for respondents at the extremes of the score distribution. The EOG-RC developmental scale range for the grade span examined in this study was a low score of 216 in third grade and a high score of 287 in seventh grade. Examination of score distributions for our analytic sample indicated adequate test score ranges within each grade and no evidence of floor or ceiling effects overall or for specific student groups. Validity evidence provided in the technical manual (NCDPI, 2004a) included high
teacher ratings of item alignment with the reading curriculum and correlations of the developmental scale scores with teacher ratings of students' expected grades in English/Language Arts (median correlation of .58 across the Grades 3 to 8) and judgment of student achievement (median correlation of .63 across Grades 3 to 8). Drawing from other sources for validity evidence, the EOG-RC have been found to correlate highly with other group administered reading tests, including the STAR reading assessments ($r$ range of .74-.80 by grade; Renaissance Learning, 2012) and the Measures of Academic Progress ($r$ range of .77-.82 by grade; Northwest Evaluation Association, 2014).

The developmental scale scores were created based on a vertical linking study using a common items design (Nicewander et al., 2013; NCDPI, 2004a). As discussed earlier, creation of vertical scales requires a number of procedures and design features to support interpretability. Patz (2007) noted that the North Carolina Language Arts test content is clearly more amenable to vertical scaling than content standards in some other states with North Carolina content standards based on common goals that provide “…continuity of language study and increasing language skill development” (NCDPI, 2004a, p. 11). The North Carolina linking design entailed the use of 12 linking item sets administered to each adjacent pair of grades for Grades 3 through 8 (Nicewander et al., 2013). In addition, "triplet forms" were used to examine linkages of scores across spans of three grades. Creation of the vertical scale employed a number of procedures recommended for vertical linking (Briggs & Weeks, 2009; Kolen & Brennan, 2004) including a separate linking approach, use of item response theory (3PL) to place forms on the common scale, and use of a maximum likelihood ability estimator. The linking process resulted in a scale that displays desirable patterns of grade to grade growth that support the vertical scale including mean score increases across grades, relatively flat grade to grade variability, and even separation of grade distributions (Kolen & Brennan, 2004). Finally, we examined the ratio of each grade’s standard deviation (SD) to the Grade 3 SD to evaluate scale variability and found no evidence of scale shrinkage (Dadey & Briggs, 2012) with ratios of -0.01, -0.04, 0.08, and 0.05 for each adjacent pair of grades from Grade 3 to 7.

Procedures

The North Carolina EOG-RC tests are administered to students the last three weeks of each school year as part of the state's educational accountability program. Most students take the test in their general education classrooms in a single session of 130 minutes (including breaks and instructions), with classroom teachers administering and proctoring the examination. Students may be exempted from testing or take an alternate assessment for a variety of reasons including medical issues, limited English proficiency, or determination by an IEP team that a student with a disability should participate in the alternate rather than the general assessment (NCDPI, 2004). Test accommodations are available for SWD and students with limited English proficiency and the percent of students in the analytical sample receiving accommodations each year ranged from 12.2 to 15.4. Across the five years of the study, the most common test accommodations for the EOG-RC were extended time (12.14% to 14.4% of the sample by year), testing in a separate room (9.3% to 13.2%), and marking in the test booklet rather than an answer sheet (4.4% to 9.4%).

For the first year of this study, 2002-2003, the third grade participation rate in the EOG-RC was 98.8% overall, and 84.0% for SWD. Within the special education population, the participation rate by disability varied from 5.2% for students identified as multi-handicapped to 99.2% for students with SLI.
**Determination of student exceptionality.** Students' primary exceptionality classifications in the third grade were used to define the exceptionality groups in this study. North Carolina identified students for exceptional children's services using category names that differed slightly from the present IDEA categories for SWD groups. We mapped the North Carolina categories in use at the time of testing onto the federal categories (e.g., students identified as "mentally handicapped" in North Carolina's testing database were described as students with intellectual disabilities). The NC identification criteria for LD identification required an ability/achievement discrepancy, and students had to show a standard score discrepancy on individually administered IQ and achievement tests of 15 standard score units, or (in rare cases) provide classroom documentation that a severe discrepancy existed in the absence of an ability/achievement discrepancy on standardized tests. We divided the LD group into students who were identified in reading (LD-R) versus other academic areas (LD-O). North Carolina also explicitly recognizes gifted and talented students in reading and mathematics as subgroups in their accountability reporting system. We created separate groups of students within those identified as academically/intellectually gifted, for children identified in reading (AG-R) and other areas (AG-O). All other students were classified as GE students.

**Construction of the longitudinal file.** The study dataset was constructed from the annual test and annual student membership electronic files available from the North Carolina Educational Research Data Center (NCERDC). These files were available for each student who attended a North Carolina school during the school year in question, even if the student was absent or exempt from testing. In addition to test scores, these records contained demographic information including students' exceptional children classification as coded by the classroom teacher. The NCERDC added a unique identifier for each student to the annual records to allow matching of student records across years. To create the longitudinal records, we first merged the test and membership files for each year, conducted data quality checks on these files, and then merged the annual files by student identification number to create the longitudinal dataset.

**Analytic Method**

We used two-level hierarchical linear models (HLM; Raudenbush & Bryk, 2002) to examine the effects of student exceptionality and demographic characteristics on third grade reading comprehension and subsequent growth across grades. We used full information maximum likelihood estimation, specified model parameters as random effects, and used data from any available time point for a student. The multilevel analyses were completed using HLM 7.0 (Raudenbush, Bryk, Cheong, Congdon, du Toit, 2011). Given our interest in developmental change, we centered the intercept at the first testing occasion (Grade 3). We did not include school as a third level in the growth models because our research questions did not pertain to school-level differences and use of a third level would have resulted in overly complex models and sample attrition due to transitions from elementary to middle school and student mobility across schools.

In model building, we first applied an unconditional growth model that served as a basis for comparison with more complex models. In the next model, we added dummy-coded predictors for student demographic variables. In the final model, we added dummy coded predictors representing student exceptionality status in Grade 3. We evaluated differences between models using deviance tests and calculation of pseudo-$R^2$ statistics.

In each conditional model, the level-1 model specified student EOG-RC scores predicted by a quadratic function of time of measurement. The level-2 models were composed of the prediction of level-1 model parameters as a function of student demographic characteristics in...
the second model and student exceptionality categories and demographic characteristics in the third model. The choice of a quadratic model to examine growth was based on previous research characterizing achievement growth as curvilinear (e.g., Lee, 2010) and preliminary analyses that indicated inclusion of the quadratic term accounted for statistically significant additional variance over a linear model. The initial level-1 model was as follows:

\[ Y_{it} = \pi_{0i} + \pi_{1i}(\text{time}) + \pi_{2i}(\text{time}^2) + r_{it} \]  

(1)

where \( Y \) was the EOG-RC developmental scale score for student \( i \) at time \( t \) and \( \pi_{0i} \) was the initial status or intercept for student \( i \) at time 0 (Grade 3), \( \pi_{1i} \) was the linear rate of change, \( \pi_{2i} \) was the quadratic curvature representing the acceleration or deceleration in each student's growth trajectory, and \( r_{it} \) was the residual for each student.

At level-2, the individual student intercept and slope estimates became the criterion variables predicted by the level-2 student characteristics. All predictors were dichotomous and uncentered, with the coefficient representing the effect for the group coded one. The level-2 equations for the reading initial status and growth parameters were as follows:

**Initial Status**, \( \pi_{0i} = \beta_{00} + \sum \beta_{0k} (\text{Predictor}_k) + u_{0i} \)  

(2)

**Linear Change**, \( \pi_{1i} = \beta_{10} + \sum \beta_{1k} (\text{Predictor}_k) + u_{1i} \)  

(3)

**Curvature**, \( \pi_{2i} = \beta_{20} + \sum \beta_{2k} (\text{Predictor}_k) + u_{2i} \)  

(4)

where \( \beta_{00} \) was the reading score intercept at Grade 3 for all students, each \( \beta_{0k} \) represents the average partial regression coefficient relating the predictor of interest to student's initial status, and \( u_{0i} \) is the residual between the fitted predicted value for each student and the student's observed reading score. For each rate of change parameter (i.e., \( \pi_{1i} \), linear growth and \( \pi_{2i} \), curvilinear change), each individual’s rate of change, \( \pi_{pi} \), was modeled as a function of the average reading comprehension rate of change, \( \beta_{p0} \). Each \( \beta_{pk} \) represents the average partial regression coefficient relating the predictor of interest to students’ rate of change, and \( u_{pi} \) is the residual between each student's fitted growth parameter of interest and the average parameter across all students.

The final step in our overall analysis strategy was the calculation of empirical Bayes (EB) estimated means and achievement gap ES at each grade. For ES, we calculated a model-based ES by subtracting the EB estimated mean for the GE students from the EB estimated means for each exceptionality group obtained from our final HLM model and dividing the estimated group difference by the square root of the sum of the level-1 and level-2 model variance components (Spybrook, Raudenbush, Liu, Congdon, & Martinez, 2008). To provide a comparison with the more descriptive, model independent type of ES more often reported in the literature for disadvantaged groups, we also calculated ES by subtracting observed means for each exceptionality group from the mean for SWoD (the combination of students in general education and students identified as AG) and dividing by the observed standard deviation of the scores for all students in that grade (Bloom et al., 2008).

**Results**

**Multilevel Growth Models**

**Unconditional and longitudinal level-1 models.** We first applied a fully unconditional random effects model, estimating only grand means and variance components. We then estimated a two-level linear longitudinal model, followed by a quadratic growth model. We allowed each growth trajectory parameter to vary randomly across students. We found that the quadratic model resulted in a statistically significant improvement in model fit over a linear model \((p < .001)\) and a multi-parameter variance component test indicated that random effects provided a better fit to the data than a fixed effects model, \( \chi^2(5) = 9434.34, p < 0.001 \).
Results of the unconditional quadratic model are shown in first columns of Table 3. For all students, the estimated mean reading comprehension score in Grade 3 was 247.80. The average initial linear change was 5.21 scale score points, which differed significantly from 0 ($z = 357.86, SE = 0.01, p < .001$). The curvature in the growth function was -0.44 scale score points, a value that also differed significantly from 0 ($z = -128.89, SE = 0.003, p < .001$). The model parameter intercorrelations between intercept and linear, intercept and curvilinear, and linear and curvilinear parameters were -.54, .50, and -.88, respectively. Multilevel model parameter reliabilities were .86 for the intercept, .16 for the linear slope, and .10 for quadratic change.

**Model with socio-demographic characteristics.** The next model added dummy coded predictors reflecting students’ socio-demographic characteristics, with students who were male, White, not receiving free lunch, and not classified as having limited proficiency in English comprising the reference group. With the introduction of the demographic variables, the intercept increased to 250.49, linear slope decreased slightly to 5.12, and curvature remained the same at -.44. As expected based on previous research (e.g., Morgan et al., 2011), students' growth trajectories differed on the basis of socio-demographic characteristics, with the greatest differences occurring in terms of initial reading achievement. Students who had limited English proficiency, or were of Black, Hispanic, American Indian, or Multiracial ethnicity had significantly lower initial intercepts than the reference group. Females and students of Asian ethnicity had significantly higher initial reading comprehension scores. In terms of linear growth and curvature, most socio-demographic characteristics were associated with significantly higher linear slope coefficients and significantly greater deceleration in growth. Compared to the unconditional growth model, the addition of socio-demographic predictors resulted in a statistically significant reduction in unexplained variance, $\chi^2 (24) = 27,638.93, p < 0.001$. The demographic only model accounted for 22.39% of the variance in the intercept, 3.18% of the variance in linear slope, and 4.13% of the variance in curvature. Intercorrelations among model parameters were as follows: -.55 between intercept and linear, .48 between intercept and curvature, and -.88 between linear and curvature.

**Model with socio-demographics and exceptionalities.** In the final model, we added predictors for student exceptionality to the previous model. In this model, the values for the reference group (the first row of the right-most columns of Table 3) represent the average intercept, slope, and quadratic curvature parameters for students who were in general education, male, White, not classified as limited English proficient, and not eligible for free lunch. Adding exceptionality groups to the model resulted in little change in intercept (250.49 to 250.82), and slight decreases in linear slope (5.12 to 5.03) and curvature (-0.44 to -0.42) for the reference group. Each exceptionality group differed significantly from the reference group for the intercept, with students who were AG-R and AG-O scoring almost 7 or 8 scale score points above the reference group and the different SWD exceptionality groups scoring from 2.64 scale score points below the reference group for students with SLI to almost 15 scale points below the reference group for students with intellectual disabilities. In terms of linear slope and curvature, only two exceptionality groups (i.e., students with autism and hearing impairments) did not differ significantly from the reference group. Students who were AG-RD and AG-O had linear slope coefficients that were negative but the AG-RD group also showed curvature that was accelerated over that seen in the reference group. In all cases where a SWD subgroup had a linear slope that differed significantly from the reference group, the coefficient was positive but there was also significant deceleration for the group.

Compared to the model that only included student socio-demographic characteristics, the addition of the exceptionality predictors resulted in a statistically significant reduction in
unexplained variance, $\chi^2 (30) = 23598.93, p < 0.001$; with the explained variance in intercept increasing about 17% to 39.44%, linear slope about 5% to 8.45%, and the variance in curvature accounted for doubling to 8.26%. Intercorrelations among model parameters in the final model were as follows: -.52 between intercept and linear, .46 between intercept and curvilinear, and -.88 between linear and curvilinear.

The empirical Bayes (EB) estimated means and standard deviations from our final growth model by grade, and by exceptionality and grade are provided in Table 4 and depicted graphically in Figure 1. Although initial achievement by group varied substantially, in general the groups showed a similar curvilinear growth pattern across grades with more rapid growth up to Grade 5, followed by slower growth afterwards. The exceptionality groups tended to maintain their relative position across grades, with students who were AG-R and AG-O consistently outperforming the general education reference group and showing less deceleration across grades. Of the SWD groups, students with SLI performed most similarly to general education students followed by students with LD-O. Students with autism, other health impairment, emotional disturbance, and hearing impairment were similar in terms of initial reading achievement and growth across the grades. Students who were LD-R showed a somewhat different growth trajectory, characterized by low initial achievement in Grade 3 but more rapid growth than several SWD groups, surpassing students with emotional disturbance and hearing impairments by seventh grade.

Because linear and quadratic growth parameters separately do not denote the actual rate of change at a particular time point, we also calculated average rates of change for each group at each grade by combining slope and curvature coefficients (see Raudenbush & Bryk, 2002, p. 171). At third grade, the reference group had an initial growth rate of 5.03 scale score points. Students with LD-R had the highest initial growth rate (6.75 scale score points), followed by students with intellectual disabilities (6.41 scale score points) and students with emotional disturbance (6.02). Students who were AG-R and AG-O had the lowest initial grade growth rates (4.44 and 4.49 respectively). By Grade 7, the average growth rate for the reference group was 1.66 showing the deceleration of growth over grades. Students who were AG-R and AG-O now had the second and third highest growth rates (1.82 and 1.76), with students with emotional disturbance having the lowest average growth rate at Grade 7 (.69 scale score points). The growth rate for students who were LD-R had slowed to 1.23 scale score points by seventh grade.

**Supplemental analysis.** As noted above, students with LD-R had the highest initial growth rate of all SWD groups. To examine whether this growth rate differed significantly, we did a follow-up analysis substituting the LD-R subgroup for the GE students as the reference group, keeping all other reference group variables the same as those in the original model. All differences in intercept, linear slope, and quadratic curvature between the LD-R and the other SWD groups were statistically significant ($p < .05$) with a few exceptions: compared to LD-R students, students with hearing impairments did not differ on intercept or quadratic curvature, students with intellectual disabilities did not differ in slope or quadratic curvature, and students with emotional disturbance did not differ on curvature.

**Growth Patterns and Achievement Gaps**

Our second research question concerned whether the pattern of individual differences related to students’ disability status (controlling for socio-demographic differences among groups) showed a fan-spread, fan-close, or stable growth pattern, and whether reading comprehension gaps closed for any SWD groups. Bast and Reitsma (1998) proposed three facets of results from longitudinal growth studies that are relevant for characterizing developmental patterns in
individual differences: (a) stability of individual differences over time, (b) stability of inter-individual variance in the population, (c) direction of the correlation between baseline level and growth (Bast & Reitsma, 1998).

In terms of Bast and Reitsma's (1998) first criterion, the differences among exceptionality groups were quite stable across grades. The ordering of means for the exceptionality groups changed little from Grade 3 to 7, with the possible exception of the LD-R group, who in third grade was ranked second from bottom and who by seventh grade had surpassed students with emotional disturbance and hearing impairments. In terms of Bast and Reitsma's second criterion, the standard deviations across grades for all students as group showed a small fan-close pattern, with the standard deviation decreasing from 8.22 in third grade to 7.32 in sixth grade and 7.61 in seventh grade (see Table 4). With regard to the third criterion, the correlation between the student EB estimated intercepts and reading gains across Grades 3 to 7 (gain calculated by subtracting the estimated intercept at Grade 3 from the estimated score at Grade 7 for each student), the correlation was $r = -0.40$ ($p < .001$). These results are consistent with a fan-close effect.

To further examine how growth differed over time for students in the SWD, AG, and GE groups, we calculated achievement gaps at each grade for students in each exceptionality group using EB estimated means, and a second time using observed means, as detailed in the analysis section. As indicated in the top half of Table 5 and Figure 2, AG-R students scored about one standard deviation above the reference group of GE students at each grade, with students who were AG-O consistently scoring above the reference group, although lower than the AG-R group. The achievement gaps for students in the different SWD exceptionality groups varied; students with SLI consistently had the smallest gaps, under a quarter of a standard deviation, and students with intellectual disabilities consistently had the largest gaps, well over 1.5 standard deviations at each grade level.

In most cases, when achievement gaps for SWD are calculated (e.g., U.S. Department of Education, 2014), the comparison group is SWoD, with AG and GE students combined to form the comparison group, and means are not adjusted for differences in socio-demographic characteristics among groups. Therefore, we also calculated ESs across grades combining the GE and AG groups (see bottom half of Table 5). Achievement gaps for SWD were larger when the reference group was all SWoD, but the pattern of differences among exceptionalities groups remained the same.

With either comparison group, no SWD group closed the achievement gap appreciably by 7th grade. Students with LD-R, the highest prevalence SWD exceptionality group, showed the largest relative achievement gain across the grades, with the ES narrowing from -1.09 in third grade to -0.83 in seventh grade in comparison to GE students using EB estimated means, and from -1.16 to -0.92 in comparison to SWoD. However, for all SWD groups, the extent to which the achievement gap closed was small relative to the size of the initial achievement gap.

**Discussion**

Evaluating congruence between the results from descriptive studies of students' reading achievement growth and theory-based predictions is one means of advancing reading theory and understanding its implications for practice. In addition, research on achievement growth is limited for groups of students, such as SWD and AG students, whose learning needs may differ from those of the general population. The primary purpose of the present study was to add to the empirical knowledge about reading comprehension achievement growth for SWD. Innovative design features of the study included examination of growth for SWD by specific exceptionality
categories; examination of growth for AG students; inclusion of a GE reference group; use of an operational state test of reading comprehension; and the calculation of ES as an empirical means of examining changes in the SWD achievement gap across grades.

**Initial Differences and Reading Comprehension Achievement Growth Across Grades**

With regard to our first research question concerning the developmental progress in reading comprehension for students in specific exceptionality groups, we found that the pattern of growth over grades observed in previous studies of the general student population (e.g., Lee, 2010; Rescorla & Rosenthal, 2004) also characterized growth in the ten exceptionality groups examined. Developmental progress in reading comprehension across Grades 3 to 7 was best represented as a curvilinear function. Students made larger gains in the early grades that decelerated as students transitioned across grades. This finding is consistent with the simple view of reading or a convergent skills model of reading comprehension development (Gough & Tumner, 1986; Vellutino et al., 2007). As readers' word recognition skills become more fully developed, allowing them to decode text near or at their listening comprehension level, skills in language comprehension become primary contributors to reading comprehension growth (García & Cain, 2013; Tighe & Schatschneider, 2014; Vellutino et al., 2007), resulting in an overall slowing of reading comprehension growth. If we assume that students who are identified as AG-R in third grade already had well developed word recognition skills, this interpretation could also account for their lower initial growth rates in third grade and higher growth rates relative to most other exceptionality groups in seventh grade.

Although the overall shape of reading comprehension achievement growth across grades resembled that found in studies of the general student population, each SWD exceptionality group had significantly lower initial achievement in third grade compared to the reference group. Most SWD exceptionality groups also showed small, but statistically significant departures from the comparison group of GE students in both linear growth and quadratic curvature, with higher linear growth, but more deceleration across grades. Students identified as AG-R and AG-O had higher initial reading comprehension achievement, but lower linear growth and less deceleration across grades.

The lower initial reading achievement and curvilinear growth observed for SWD in the present study has been found in studies of reading growth for students with LD and SLI (e.g., Francis et al., 1996; Judge & Bell, 2010; Morgan et al., 2011). Similar to the one study that examined multiple exceptionalities other than LD and SLI (Wei et al., 2011), we observed considerable heterogeneity in intercept by exceptionality. On a test where the average GE student annual growth was approximately 5 scale score points in the third grade, differences in intercept between GE students and SWD in the present study ranged from less than 3 scale score points for students with SLI to almost 18 points for students with intellectual disabilities. Although Wei et al. examined differences in SWD intercepts when children were about three and a half years older than students in the present study (approximately 9.3 vs. 12.7 years of age for Wei et al.), the rank ordering of exceptionality groups was similar. Students with intellectual disabilities had the lowest reading achievement level and students with SLI had the highest level of the SWD groups. Students with autism, other health impairment, LD, or emotional disturbance showed initial achievement levels similar to one another that fell between the least and most impaired SWD groups.

Morgan et al. (2011) found that students with SLI ranked lower in reading achievement at the end of first grade than students with LD, a finding at odds with the present study and Wei et al. (2011). Longitudinal studies of changes in exceptionality classification across grades suggest
that many students initially identified with SLI in preschool are later identified as having LD (Delgado, 2009). It may be that students who are identified as SLI in preschool or kindergarten who later show significant impairment in reading are then reclassified as LD, resulting in a change in relative reading achievement for the two exceptionality groups. We were unable to obtain special education classifications for students prior to third grade to examine whether some of students in the LD-R group had previously been identified as SLI, but did confirm that the small number of students who were SLI in third grade that were later classified as LD-R had reading comprehension trajectories that more closely tracked the LD-R group.

The finding in the present study that students with autism ranked higher than students with LD-R is a reversal of the rankings reported by Wei et al. (2011). The difference in relative standing between the exceptionalities is likely due to differences in the subset of students with autism included in each study. In the present study, over half the students with autism had consistently participated in an alternate reading assessment rather than the general assessment, and were not included in the study. Although Wei et al. excluded some students with cognitive impairments from their sample, their reading assessment was designed for a wider range of ability and age (three to adulthood), and it is likely their study included students with more severe cognitive impairments than the students with autism in the present study.

Developmental Pattern for Reading Comprehension: Stable, Widening, or Narrowing?

Our second research question concerned the developmental growth pattern observed for reading comprehension skills and its impact on reading comprehension achievement gaps for SWD across grades. Taken in their totality, our findings were most consistent with a stable differences developmental pattern. For SWD, we found no evidence of a fan-spread pattern or Matthew effect relative to the criteria proposed by previous researchers (Bast & Reitsma, 1998; Pfost et al., 2014). Although some of our findings could be viewed as supportive of a fan-close pattern in the general population (i.e., the small decrease in total variance across time and moderate negative correlation between intercept and total gain), achievement gaps for SWD changed very little across grades, and group rankings for the exceptionality groups also remained stable. After four years, none of the SWD groups had "caught up" with the students in GE in their reading comprehension achievement. One possible exception to the stable differences pattern was the .26 SD unit decrease from Grade 3 to 7 in the achievement gap ES for students with LD-R. Taken as a percent of the initial achievement gap (ES = -1.09), this represented a 24% decrease in the achievement gap for students with LD-R.

It is difficult to evaluate the consistency of the present findings with previous research given the lack of longitudinal studies that have examined exceptionality groups other than LD and SLI, and the differences in methodology and age span studied within the existing corpus of studies. For SLI students, our results are consistent with Catt et al.’s (2008) characterization of differences for this group as stable across grades, but in conflict with the results of Morgan et al. (2011), who found an increasing achievement gap in reading for this group. Our results concerning changes in the achievement gap across grades for students who are LD-R were more positive than the widening achievement gap for students with LD reported by Judge and Bell (2010), and the stable gap reported by Morgan et al. (2011). However, neither of those studies separated out students with LD in reading versus other academic areas, and our results would more closely track Morgan et al.’s if our two LD groups had been combined. Our finding that students with LD-R differed from students with LD-O in terms of intercept and growth suggests that studies combining these two groups (e.g., Judge & Bell, 2010; Wei et al., 2011) may not
yield results that are representative of either group, and may cloud the possible differential response of the two groups to school-based reading interventions.

As noted earlier, several researchers have suggested that differences in findings across studies relative to developmental patterns of growth in reading may be attributable to the (a) grade at which initial status is determined, (b) the grade span examined, or (c) the reading skill assessed (Kieffer, 2012; Morgan et al., 2011; Pfost et al., 2014). For example, Judge and Bell (2010) and Morgan et al. (2011) used the same dataset (the Early Childhood Longitudinal Study-Kindergarten Cohort, ECLS-K), grade span, and composite reading measure, but placed their intercepts at kindergarten entry (Judge & Bell, 2010) and the end of first grade (Morgan et al., 2011). They found marked differences in the correlation between intercept and slope (.35 versus -.07). This suggests that the choice of intercept location is critical to the overall growth pattern observed, and our placement of the initial intercept at Grade 3 may explain our moderate and negative correlation between initial intercept and total gain. On the other hand, it is unlikely that reading comprehension was weighted heavily, if at all, at Grades K and 1 in the ECLS-K composite assessment, and the present study used a measure of reading comprehension. The differences in types of reading skills assessed may have contributed to discrepant findings across studies. It should also be noted that patterns of growth and the size of achievement differences may be affected by artifactual differences in the vertical scale from one assessment to another that can impact functional form or the equality of score intervals over time (Bolt et al., 2013; Briggs & Weeks, 2009).

**Potential Explanations for the Stable Differences Pattern**

The correlational nature of the present study, and lack of information about home and school literacy influences or cognitive and behavioral characteristics related to reading, except as they are represented by the different exceptionality classifications, limit any conclusions about the mechanisms underlying the observed stable differences growth pattern. One hypothesis that fits the study's overall pattern of results is that SWD in the present study were receiving instruction in general and special education that emphasized the development of word recognition skills—a likely culprit when most children experience early reading difficulties (Rayner, Foorman, Perfetti, Pesetsky & Seidenberg, 2001). If so, then that instruction may have produced sufficient gains in early word recognition skills to allow the rapid growth in reading comprehension seen as readers initially develop a corpus of sight words (Ehri, 2005), but it did not result in growth in the skills that underlie age-appropriate reading comprehension as language comprehension skills versus word identification skills become ascendant as determinants of reading comprehension for typically developing readers (Scarborough, 2001; Tighe & Schatschneider, 2014). As such, SWD showed an increased rate of growth as the initial bottleneck in their reading comprehension skill development was removed, but the instruction did not address needs that were "hidden" by the students' initial decoding deficits, such as deficits in background knowledge, making inferences, constructing meaning from text, or failure to develop sophisticated context-dependent word identification skills for word recognition (Compton et al., 2014; Oakhill & Cain, 2012; Perfetti & Stafura, 2014).

Supporting this view is the finding that the group of students with LD-R, the SWD group most likely to have a specific and marked deficit in word recognition, had an initial growth rate that differed significantly from all but one of the SWD groups, and was the only group where the achievement gap showed some narrowing. Other evidence supporting this view (although much less direct) is that the time period when students were receiving special education services in the present study corresponded with a state-wide professional development initiative in North
Carolina aimed at improving the teaching of reading foundations for SWD (NCDPI, 2005). This interpretation is also consistent with findings from a recent review of observational studies of special education instruction (McKenna, Shin, & Ciullo, 2015) that indicated that phonics instruction accounted for a substantial portion of time in special education settings with limited time spent teaching reading comprehension.

Baumert et al. (2012) postulated that a stable differences pattern for reading growth could result from competing mechanisms in reading development that produce simultaneous fan-close and fan-spread patterns. An explanation for the present study's results consistent with that hypothesis is that schools differentially allocate resources to assure that low-achieving children reach grade-level proficiency and these actions produce a fan-close pattern for schooling that is countered by many mechanisms within reading development that produce a fan-spread pattern (Stanovich, 1986). This interpretation is supported by research showing the cumulative impact of summer and after school experiences on achievement growth (Alexander, Entwisle, & Olson, 2007; Pfost, Dorfler, & Artelt, 2013) and the pattern of greater gains for students who are AG in summer compared to reading gains during the school year (Rambo-Hernandez & McCoach, 2015).

Study Limitations

Although this study has contributed new evidence on the reading achievement growth of SWD, these findings should be interpreted in light of several limitations. Specifically, the present study only included SWD who participated in the general reading assessment at some time during grades three to seven. SWD with severe cognitive impairments, most prevalent in the exceptionalities of intellectual disability and autism, did not take the general assessment and were not represented in the analytic sample. Therefore, our results should be considered representative of reading growth only for students participating in the general education assessment. A second important issue relative to the SWD group is that we based exceptionality category membership on students' third grade primary special education classification. Some students in special education are served in two or more exceptionality categories and we did not have information on how many students had multiple exceptionality classifications or comorbid conditions, such as attention-deficit/hyperactivity disorder, that are prevalent across categories (Blackorby et al., 2005). Reading achievement trajectories for students with comorbid disabilities may differ from the outcomes reported by primary disability only. Students enter and exit special education throughout their educational careers (Ysseldyke & Bielinski, 2002), change exceptionality classifications (Blackorby et al., 2005), and move in and out of the alternate and general assessment, and these changes were not represented in the present study. Depictions of achievement growth for students who were consistently in special education across the grade span of the study or who entered special education after third grade may be different (Schulte & Stevens, 2015).

We had no information about students' educational and home environments, or their motivation to read, independent reading, or attention and engagement in school, although each of these factors is related to reading comprehension growth (Guthrie et al., 2007; Morgan et al., 2011; Wei et al., 2011). We also lack external validation of the exceptionalities and how closely school identification of exceptionality matches with descriptions of different disabilities in the research literature. Another limitation of the present study is the use of large scale reading assessment data from only one state. Eligibility criteria, prevalence, and characteristics of children receiving special education differ by state (U.S. Department of Education, 2015), as do the content and format of state reading assessments (May, Perez-Johnson, Haimson, Sattar, &
Gleason, 2009). The NC state reading assessment focused exclusively on comprehension of connected text and did not include items specifically designed to assess reading vocabulary, decoding, or other reading skills which may show different patterns of individual differences in reading growth (Pfost et al., 2014).

Finally, although we had the advantage of using a database that permitted tracking of students who moved from one public school or school district within the state, we lost data for students who moved out of state or entered private schools. Our losses (about 4% per year) were less than observed in most longitudinal studies (e.g., Choi, Seltzer, Herman, & Yamashiro, 2007), but attrition still may have had a measurable effect on estimation of growth or ES.

**Implications and Conclusions**

The present study adds to the growing body of research about reading growth in SWD and the limited research on achievement growth for students who are AG. Cross sectional depictions of the SWD achievement gap have led to concerns that SWD are falling further behind SWoD at each grade (Morgan et al., 2008; Vaughn & Wanzek, 2014). Our results suggest that stable differences or a slight fan-close pattern more accurately describes the pattern of individual differences between SWD and AG students relative to GE students across third to seventh grade. Although this finding is more positive than some current portrayals of the SWD achievement gap, it should in no way minimize the implications of the substantial gaps present at third grade and beyond for SWD.

Why did the more rapid growth for SWD fail to result in a narrowing of the achievement gap, with the exception of a small amount of closure in the gap for students who were LD-R? We speculated that the more rapid growth observed for SWD could be attributed to an increased focus on instruction in word recognition in special education as a result of national and state policy initiatives. Removing the initial "bottleneck" in word recognition for children who were, on average, acquiring these skills somewhat later than SWoD, had a lasting impact only for the group where this deficit was most likely to be primary. If this interpretation is correct, it suggests that early assessment and intervention directed toward oral language comprehension and vocabulary for students with disabilities who show deficits in these areas may be a means of sustaining the more rapid initial growth in reading comprehension for SWD observed in the present study (e.g., Clarke, Snowling, Truelove, & Hulme, 2010).

In terms of policy implications, our findings reinforce the concerns expressed by others (e.g., Wei et al., 2011) that one-size-fits-all achievement expectations ignore the magnitude and complexity of the differences in reading comprehension achievement for SWD already present in the third grade. Closing some of the larger SWD achievement gaps, even over the course of several grades, would require reading growth rates with much greater acceleration than those observed in the present study and higher than those found in studies with intensive, multi-year interventions for students with reading difficulties (e.g., Allor, Mathes, Roberts, Cheatham, & Otaiba, 2014; Roberts et al., 2013). Our findings do suggest that accountability models that set similar expectations for achievement growth for SWD and SWoD may more closely match data than growth to proficiency models that require much more accelerated growth for SWD.

NCLB’s (2002) goal of all students reaching grade-level proficiency within a uniform timeframe was set in the context of a limited body of research about students’ developmental patterns of achievement growth. This lack of information was particularly acute for SWD who had yet to be fully included in many national and state assessments, but also extended to other groups such as students who are gifted. The implementation of NCLB, with its annual assessments of students, has resulted in rich data sources for understanding student achievement
growth, but these data spotlight the challenges in achieving its ultimate goal. Although our results do not indicate increasing achievement gaps, the stable differences or slight fan-close pattern observed indicate that, on an assessment designed to monitor NCLB progress, the goal of closing the achievement gap for SWD has yet to be met.

References


Harris, D. N. (2009). The policy uses and “policy validity” of value-added and other teacher quality measures. In D. H. Gitomer (Ed.), *Measurement issues and assessment for teaching*


U.S. Department of Education, Office of Special Education Programs (2015). *Thirty-sixth annual report to Congress on the implementation of the Individuals with Disabilities*


### Table 1

*Student Disability Group for the Total and Analytic Sample at Wave 1*

<table>
<thead>
<tr>
<th>Characteristic</th>
<th>Total sample</th>
<th></th>
<th>Analytic sample</th>
<th></th>
<th>h</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>N</td>
<td>%</td>
<td>N</td>
<td>%</td>
<td></td>
</tr>
<tr>
<td>Students without disabilities</td>
<td>87,226</td>
<td>85.6</td>
<td>87,028</td>
<td>87.1</td>
<td></td>
</tr>
<tr>
<td>General education</td>
<td>80,182</td>
<td>78.7</td>
<td>79,984</td>
<td>80.0</td>
<td>.033</td>
</tr>
<tr>
<td>Academically gifted, reading</td>
<td>5,695</td>
<td>5.6</td>
<td>5,695</td>
<td>5.7</td>
<td>.005</td>
</tr>
<tr>
<td>Academically gifted, other</td>
<td>1,349</td>
<td>1.3</td>
<td>1,349</td>
<td>1.4</td>
<td>.002</td>
</tr>
<tr>
<td>Students with disabilities</td>
<td>14,642</td>
<td>14.4</td>
<td>12,891</td>
<td>12.9</td>
<td>.043</td>
</tr>
<tr>
<td>Autism</td>
<td>395</td>
<td>0.4</td>
<td>204</td>
<td>.2</td>
<td>.034</td>
</tr>
<tr>
<td>Deaf-blindness</td>
<td>--</td>
<td>--</td>
<td>--</td>
<td>--</td>
<td></td>
</tr>
<tr>
<td>Emotional disturbance</td>
<td>739</td>
<td>0.7</td>
<td>701</td>
<td>0.7</td>
<td>.003</td>
</tr>
<tr>
<td>Hearing impairment</td>
<td>181</td>
<td>0.2</td>
<td>159</td>
<td>0.2</td>
<td>.005</td>
</tr>
<tr>
<td>Intellectual disability</td>
<td>2,186</td>
<td>2.1</td>
<td>1,280</td>
<td>1.3</td>
<td>.067</td>
</tr>
<tr>
<td>Learning disability, reading</td>
<td>4,827</td>
<td>4.7</td>
<td>4,668</td>
<td>4.7</td>
<td>.003</td>
</tr>
<tr>
<td>Learning disability, other</td>
<td>1,151</td>
<td>1.1</td>
<td>1,140</td>
<td>1.1</td>
<td>.001</td>
</tr>
<tr>
<td>Multiple disabilities</td>
<td>115</td>
<td>0.1</td>
<td>--</td>
<td>--</td>
<td></td>
</tr>
<tr>
<td>Orthopedic impairment</td>
<td>83</td>
<td>0.1</td>
<td>--</td>
<td>--</td>
<td></td>
</tr>
<tr>
<td>Speech-language impairment</td>
<td>3,044</td>
<td>3.0</td>
<td>3,013</td>
<td>3.0</td>
<td>.002</td>
</tr>
<tr>
<td>Other health impairment</td>
<td>1,832</td>
<td>1.8</td>
<td>1,726</td>
<td>1.7</td>
<td>.005</td>
</tr>
<tr>
<td>Traumatic brain injury</td>
<td>22</td>
<td>&lt;0.1</td>
<td>--</td>
<td>--</td>
<td></td>
</tr>
<tr>
<td>Visual impairment</td>
<td>67</td>
<td>0.1</td>
<td>--</td>
<td>--</td>
<td></td>
</tr>
<tr>
<td>Missing exceptionality</td>
<td>17</td>
<td>&lt;0.2</td>
<td>--</td>
<td>--</td>
<td></td>
</tr>
<tr>
<td>Total sample size</td>
<td>101,885</td>
<td></td>
<td>99,919</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

*Note.* The cell frequency for students with deaf blindness was <10, so this group is not reported.
Table 2

*Student Demographic Characteristics by Sample, and by Student Group for the Analytic Sample at Wave 1*

<table>
<thead>
<tr>
<th>Characteristic</th>
<th>Total sample</th>
<th></th>
<th></th>
<th>SWoD</th>
<th></th>
<th></th>
<th>SWD</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>N</td>
<td>%</td>
<td>N</td>
<td>%</td>
<td>h</td>
<td>N</td>
<td>%</td>
<td>h</td>
</tr>
<tr>
<td>Female</td>
<td>49,564</td>
<td>48.6</td>
<td>48,875</td>
<td>48.9</td>
<td>.005</td>
<td>44,644</td>
<td>51.3</td>
<td>4,231</td>
</tr>
<tr>
<td>American Indian</td>
<td>1,515</td>
<td>1.5</td>
<td>1,479</td>
<td>1.2</td>
<td>.001</td>
<td>1,254</td>
<td>1.4</td>
<td>225</td>
</tr>
<tr>
<td>Asian</td>
<td>2,084</td>
<td>2.0</td>
<td>2,035</td>
<td>2.0</td>
<td>.001</td>
<td>1,919</td>
<td>2.2</td>
<td>116</td>
</tr>
<tr>
<td>Black</td>
<td>29,905</td>
<td>29.4</td>
<td>29,178</td>
<td>29.2</td>
<td>.003</td>
<td>25,059</td>
<td>28.8</td>
<td>4,119</td>
</tr>
<tr>
<td>Hispanic</td>
<td>7,140</td>
<td>7.0</td>
<td>6,911</td>
<td>6.9</td>
<td>.004</td>
<td>6,304</td>
<td>7.2</td>
<td>607</td>
</tr>
<tr>
<td>Multi-racial</td>
<td>2,474</td>
<td>2.4</td>
<td>2,439</td>
<td>2.4</td>
<td>.001</td>
<td>2,140</td>
<td>2.5</td>
<td>299</td>
</tr>
<tr>
<td>White</td>
<td>58,767</td>
<td>57.7</td>
<td>57,877</td>
<td>57.9</td>
<td>.005</td>
<td>50,352</td>
<td>59.9</td>
<td>7,525</td>
</tr>
<tr>
<td>Limited English proficiency</td>
<td>5,140</td>
<td>5.0</td>
<td>4,920</td>
<td>4.9</td>
<td>.006</td>
<td>4,464</td>
<td>5.1</td>
<td>456</td>
</tr>
<tr>
<td>Title I student</td>
<td>3,741</td>
<td>3.7</td>
<td>3,690</td>
<td>3.7</td>
<td>.001</td>
<td>2361</td>
<td>3.7</td>
<td>454</td>
</tr>
<tr>
<td>Free lunch</td>
<td>39,124</td>
<td>38.4</td>
<td>37,968</td>
<td>38.0</td>
<td>.008</td>
<td>31,754</td>
<td>36.5</td>
<td>6,214</td>
</tr>
<tr>
<td>Total sample size</td>
<td>101,885</td>
<td>99.919</td>
<td>87,028</td>
<td>87.0</td>
<td></td>
<td>12,891</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

*Note.* SWoD = students without disabilities; SWD = students with disabilities.
Table 3

Reading Results From Fixed and Random Effects Longitudinal HLM Regression Models (N = 99,919)

<table>
<thead>
<tr>
<th>Predictor</th>
<th>Unconditional</th>
<th>Demographics</th>
<th>Demographics &amp; Exceptionality</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Intercept</td>
<td>Linear</td>
<td>Quadratic</td>
</tr>
<tr>
<td>Mean</td>
<td>247.80**</td>
<td>5.21**</td>
<td>-0.44**</td>
</tr>
<tr>
<td></td>
<td>(0.03)</td>
<td>(0.01)</td>
<td>(0.00)</td>
</tr>
<tr>
<td>Sex</td>
<td>1.73**</td>
<td>-0.26**</td>
<td>0.08**</td>
</tr>
<tr>
<td></td>
<td>(0.05)</td>
<td>(0.03)</td>
<td>(0.01)</td>
</tr>
<tr>
<td>Free lunch</td>
<td>-4.76**</td>
<td>0.25**</td>
<td>-0.07**</td>
</tr>
<tr>
<td></td>
<td>(0.06)</td>
<td>(0.03)</td>
<td>(0.01)</td>
</tr>
<tr>
<td>Limited English</td>
<td>-5.95**</td>
<td>0.77**</td>
<td>-0.06*</td>
</tr>
<tr>
<td></td>
<td>(0.17)</td>
<td>(0.10)</td>
<td>(0.02)</td>
</tr>
<tr>
<td>Asian</td>
<td>1.32**</td>
<td>0.07</td>
<td>0.07*</td>
</tr>
<tr>
<td></td>
<td>(0.18)</td>
<td>(0.10)</td>
<td>(0.02)</td>
</tr>
<tr>
<td>American Indian</td>
<td>-3.02**</td>
<td>-0.61**</td>
<td>0.12**</td>
</tr>
<tr>
<td></td>
<td>(0.24)</td>
<td>(0.14)</td>
<td>(0.03)</td>
</tr>
<tr>
<td>Black</td>
<td>-4.56**</td>
<td>0.25**</td>
<td>-0.05**</td>
</tr>
<tr>
<td></td>
<td>(0.07)</td>
<td>(0.04)</td>
<td>(0.01)</td>
</tr>
<tr>
<td>Hispanic</td>
<td>-1.10**</td>
<td>0.34**</td>
<td>-0.04*</td>
</tr>
<tr>
<td></td>
<td>(0.14)</td>
<td>(0.08)</td>
<td>(0.02)</td>
</tr>
<tr>
<td>Multiracial</td>
<td>-0.95**</td>
<td>0.17</td>
<td>-0.02</td>
</tr>
<tr>
<td></td>
<td>(0.18)</td>
<td>(0.10)</td>
<td>(0.02)</td>
</tr>
<tr>
<td>Condition</td>
<td>Unconditional</td>
<td>Exceptionality classification</td>
<td>Exceptionality &amp; demographics</td>
</tr>
<tr>
<td>-----------------------------------</td>
<td>---------------</td>
<td>-------------------------------</td>
<td>-------------------------------</td>
</tr>
<tr>
<td></td>
<td>Intercept</td>
<td>Linear</td>
<td>Quadratic</td>
</tr>
<tr>
<td>Variance component</td>
<td>72.50**</td>
<td>3.33**</td>
<td>0.11**</td>
</tr>
<tr>
<td>Residual</td>
<td>12.97</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Pseudo- $R^2$ (as %)</td>
<td>--</td>
<td>--</td>
<td>--</td>
</tr>
</tbody>
</table>
### Model df

<table>
<thead>
<tr>
<th>Delta Deviance, $\chi^2$ (df, p-value)</th>
<th>99,918</th>
<th>99,910</th>
<th>99,900</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>27,638.93 (24, &lt;.001)</td>
<td>23,598.93 (30, &lt;.001)</td>
<td></td>
</tr>
</tbody>
</table>

*Note. Standard errors shown in parentheses.*

* $p < .05$

** $p < .001$
Table 4
*Empirical Bayes estimated means and standard deviations (in parentheses) from the final HLM regression model*

<table>
<thead>
<tr>
<th>Student group</th>
<th>Grade</th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>3</td>
<td>4</td>
<td>5</td>
<td>6</td>
<td>7</td>
</tr>
<tr>
<td>All students</td>
<td>247.77</td>
<td>252.56</td>
<td>256.46</td>
<td>259.48</td>
<td>261.62</td>
</tr>
<tr>
<td></td>
<td>(8.22 )</td>
<td>(7.61 )</td>
<td>(7.31 )</td>
<td>(7.32 )</td>
<td>(7.61 )</td>
</tr>
<tr>
<td>General education</td>
<td>248.14</td>
<td>252.85</td>
<td>256.71</td>
<td>259.73</td>
<td>261.90</td>
</tr>
<tr>
<td></td>
<td>(7.16 )</td>
<td>(6.65 )</td>
<td>(6.42 )</td>
<td>(6.45 )</td>
<td>(6.72 )</td>
</tr>
<tr>
<td>Academically gifted, reading</td>
<td>258.23</td>
<td>262.33</td>
<td>265.83</td>
<td>268.73</td>
<td>271.03</td>
</tr>
<tr>
<td></td>
<td>(4.28 )</td>
<td>(3.95 )</td>
<td>(3.84 )</td>
<td>(3.91 )</td>
<td>(4.12 )</td>
</tr>
<tr>
<td>Academically gifted, other</td>
<td>256.77</td>
<td>260.94</td>
<td>264.46</td>
<td>267.34</td>
<td>269.56</td>
</tr>
<tr>
<td></td>
<td>(4.80 )</td>
<td>(4.49 )</td>
<td>(4.41 )</td>
<td>(4.51 )</td>
<td>(4.76 )</td>
</tr>
<tr>
<td>Autism</td>
<td>241.37</td>
<td>246.41</td>
<td>250.55</td>
<td>253.80</td>
<td>256.17</td>
</tr>
<tr>
<td></td>
<td>(9.23 )</td>
<td>(8.58 )</td>
<td>(8.28 )</td>
<td>(8.29 )</td>
<td>(8.59 )</td>
</tr>
<tr>
<td>Emotional disturbance</td>
<td>238.97</td>
<td>244.48</td>
<td>248.55</td>
<td>251.20</td>
<td>252.41</td>
</tr>
<tr>
<td></td>
<td>(7.65 )</td>
<td>(7.11 )</td>
<td>(6.89 )</td>
<td>(6.95 )</td>
<td>(7.26 )</td>
</tr>
<tr>
<td>Hearing impairment</td>
<td>238.97</td>
<td>244.34</td>
<td>248.54</td>
<td>251.57</td>
<td>253.43</td>
</tr>
<tr>
<td></td>
<td>(8.00 )</td>
<td>(7.52 )</td>
<td>(7.33 )</td>
<td>(7.44 )</td>
<td>(7.81 )</td>
</tr>
<tr>
<td>Intellectual disability</td>
<td>230.81</td>
<td>236.71</td>
<td>241.19</td>
<td>244.25</td>
<td>245.88</td>
</tr>
<tr>
<td></td>
<td>(4.81 )</td>
<td>(4.48 )</td>
<td>(4.37 )</td>
<td>(4.43 )</td>
<td>(4.65 )</td>
</tr>
<tr>
<td>Learning disability, reading</td>
<td>237.86</td>
<td>244.05</td>
<td>248.81</td>
<td>252.15</td>
<td>254.07</td>
</tr>
<tr>
<td></td>
<td>(7.82 )</td>
<td>(7.27 )</td>
<td>(7.04 )</td>
<td>(7.08 )</td>
<td>(7.38 )</td>
</tr>
<tr>
<td>Learning disability, other</td>
<td>242.95</td>
<td>248.34</td>
<td>252.49</td>
<td>255.42</td>
<td>257.11</td>
</tr>
<tr>
<td></td>
<td>(7.85 )</td>
<td>(7.31 )</td>
<td>(7.07 )</td>
<td>(7.12 )</td>
<td>(7.41 )</td>
</tr>
<tr>
<td>Other health impairment</td>
<td>240.33</td>
<td>245.51</td>
<td>249.55</td>
<td>252.46</td>
<td>254.24</td>
</tr>
<tr>
<td></td>
<td>(7.69 )</td>
<td>(7.15 )</td>
<td>(6.93 )</td>
<td>(6.98 )</td>
<td>(7.28 )</td>
</tr>
<tr>
<td>Speech-language impairment</td>
<td>245.85</td>
<td>250.85</td>
<td>254.89</td>
<td>257.95</td>
<td>260.04</td>
</tr>
<tr>
<td></td>
<td>(7.94 )</td>
<td>(7.40 )</td>
<td>(7.16 )</td>
<td>(7.20 )</td>
<td>(7.49 )</td>
</tr>
</tbody>
</table>
Table 5
*Model-based and observed reading comprehension achievement gap effect sizes by exceptionality*

<table>
<thead>
<tr>
<th>Student group</th>
<th>Grade</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>3</td>
</tr>
<tr>
<td><strong>Using empirical Bayes (EB) estimated means</strong></td>
<td></td>
</tr>
<tr>
<td>Academically gifted, reading</td>
<td>+1.07</td>
</tr>
<tr>
<td>Academically gifted, other</td>
<td>+0.92</td>
</tr>
<tr>
<td>Autism</td>
<td>-0.72</td>
</tr>
<tr>
<td>Emotional disturbance</td>
<td>-0.97</td>
</tr>
<tr>
<td>Hearing impairment</td>
<td>-0.97</td>
</tr>
<tr>
<td>Intellectual disability</td>
<td>-1.84</td>
</tr>
<tr>
<td>Learning disability, reading</td>
<td>-1.09</td>
</tr>
<tr>
<td>Learning disability, other</td>
<td>-0.55</td>
</tr>
<tr>
<td>Other health impairment</td>
<td>-0.83</td>
</tr>
<tr>
<td>Speech-language impairment</td>
<td>-0.24</td>
</tr>
<tr>
<td><strong>Using obtained means from analytical sample</strong></td>
<td></td>
</tr>
<tr>
<td>Autism</td>
<td>-0.66</td>
</tr>
<tr>
<td>Emotional disturbance</td>
<td>-1.02</td>
</tr>
<tr>
<td>Hearing impairment</td>
<td>-1.01</td>
</tr>
<tr>
<td>Intellectual disability</td>
<td>-1.90</td>
</tr>
<tr>
<td>Learning disability, reading</td>
<td>-1.16</td>
</tr>
<tr>
<td>Learning disability, other</td>
<td>-0.64</td>
</tr>
<tr>
<td>Other health impairment</td>
<td>-0.90</td>
</tr>
<tr>
<td>Speech-language impairment</td>
<td>-0.34</td>
</tr>
</tbody>
</table>

Note. Effect sizes in top half of table were calculated by subtracting EB estimated mean for the GE students from the EB mean for each exceptionality group and dividing the estimated group difference by the square root of the sum of the level-1 and level-2 model variance components. Effect sizes in the bottom half of table calculated by subtracting the observed mean for all students without disabilities from each exceptionality group observed mean and dividing by the observed standard deviation of the scores for all students in that grade.
Figure 1. Empirical Bayes estimated means from final model by grade and student group.
**Figure 2.** Reading achievement gap effect sizes based on differences in empirical Bayes estimated means across grades for students in different exceptionality categories compared to students in general education.