

Do Test Score Gaps Grow Before, During, or Between the School Years? Measurement Artifacts and What We Can Know in Spite of Them

Paul T. von Hippel,^a Caitlin Hamrock^b

a) University of Texas at Austin; b) E3 Alliance

Abstract: Do test score gaps between advantaged and disadvantaged children originate inside or outside schools? One approach to this classic question is to ask (1) How large are gaps when children enter school? (2) How much do gaps grow later on? (3) Do gaps grow faster during school or during summer? Confusingly, past research has given discrepant answers to these basic questions.

We show that many results about gap growth have been distorted by measurement artifacts. One artifact relates to scaling: Gaps appear to grow faster if measurement scales spread with age. Another artifact relates to changes in test form: Summer gap growth is hard to estimate if children take different tests in spring than in fall.

Net of artifacts, the most replicable finding is that gaps form mainly in early childhood, before schooling begins. After school begins, most gaps grow little, and some gaps shrink. Evidence is inconsistent regarding whether gaps grow faster during school or during summer. We substantiate these conclusions using new data from the Growth Research Database and two data sets used in previous studies of gap growth: the Beginning School Study and the Early Childhood Longitudinal Study, Kindergarten Cohort of 1998–1999.

Keywords: achievement gap; summer learning loss; summer setback; summer slide; early childhood; inequality

Citation: von Hippel, Paul T., and Caitlin Hamrock. 2019. "Do Test Score Gaps Grow Before, During, or Between the School Years? Measurement Artifacts and What We Can Know in Spite of Them." *Sociological Science* 6: 43–80.


Received: February 20, 2018

Accepted: July 23, 2018

Published: January 24, 2019

Editor(s): Jesper Sørensen, Stephen Morgan

DOI: 10.15195/v6.a3

Copyright: © 2019 The Author(s). This open-access article has been published under a Creative Commons Attribution License, which allows unrestricted use, distribution and reproduction, in any form, as long as the original author and source have been credited. 

GIVE me a child until the age of seven, and I will give you the man.

—attributed to the Jesuits

A classic question in the social and behavioral sciences is whether inequality in academic achievement comes primarily from inside or outside of schools (Coleman et al. 1966; Downey and Condron 2016; Downey, von Hippel, and Broh 2004; Jencks 1972; Jennings et al. 2015; Raudenbush and Eschmann 2015). The question is fundamental to both research and policy. If inequality grows mainly inside schools, then identifying and eliminating school-based mechanisms that exacerbate inequality should be a major priority. But if inequality grows mainly outside of schools, then research and policy should focus more on reducing nonschool inequality and on evaluating ways that schools and other institutions can compensate for inequality that originates elsewhere.

One way to assess the relative importance of school and nonschool inequality is with a multiyear longitudinal study that tests children repeatedly, starting near

the beginning of their school career. The idea is that test score gaps that are present when children enter school can only be due to nonschool influences, whereas later gap growth is due to a mix of school and nonschool factors (e.g., Phillips, Crouse, and Ralph 1998).

The relative influence of school and nonschool effects can be separated further if a longitudinal study uses a *seasonal* design that tests children twice per year—in fall and spring—so that learning during the school year can be separated from learning during the summer. The idea behind a seasonal design is that summer learning is due primarily to nonschool factors, whereas school-year learning is due to a mix of school and nonschool influences. If score gaps grow fastest during the school year, then it would appear that schools are the primary source of test score inequality, at least after the age of five. But if score gaps grow fastest during summer vacations, then it would appear that, even after school begins, the major sources of inequality lie elsewhere (e.g., Alexander, Entwisle, and Olson 2001; Downey et al. 2004; Hayes and Grether 1969, 1983; Heyns 1978; Murnane 1975).

Confusingly, estimates of gap growth have been inconsistent from one study to another. Some studies have reported that test score gaps grow very little after children enter school (Duncan and Magnuson 2011; Heckman and Masterov 2007). Other studies have reported that gaps grow dramatically between first and eighth grade, with practically all the growth occurring during summer vacations (Alexander et al. 2001; Hayes and Grether 1969, 1983). And many studies attribute gaps and gap growth to schools, teachers, or specific educational practices (e.g., Condrón 2009; Hanushek and Rivkin 2009; Heck 2007).

Results have been inconsistent for different types of gaps as well. The gap between children of high and low socioeconomic status (SES) has been reported to grow fastest during summer (Downey et al. 2004; Entwisle and Alexander 1992), but the gap between black and white children has been reported to grow fastest during school (Condrón 2009; Downey et al. 2004; Entwisle and Alexander 1994).

From these disparate results, one might conclude that the relative importance of in-school and out-of-school inequality varies according to the place, the time, and the groups being compared. And those contextual factors may play some role.

In this article, though, we present evidence that discrepancies between different estimates of gap growth often stem not from context but from artifacts of testing and measurement. Although measurement is often viewed as subordinate to substantive research questions, the way that tests are designed, administered, and scaled can have profound implications for the apparent magnitude and timing of growth in test score gaps. As we will show, some widely cited studies of gap growth suffered from measurement artifacts. In some research, the artifacts were much larger than the true gap growth that researchers were trying to measure.

If we discard studies with serious measurement artifacts—or if we accept those studies and try to compensate for artifacts—we come to more consistent conclusions about the growth of test score gaps. Net of artifacts, results consistently show that gaps are already substantial when children enter school and grow relatively little later on.

Results are not as consistent on the question of whether gaps grow faster during summer or during school. In most analyses, though, it appears that the cumulative

total of gap growth, across every school year and summer from kindergarten through eighth grade, is considerably smaller than the gaps that have already opened before kindergarten begins. Results that suggest otherwise typically suffer from measurement artifacts.

In short, the bulk of results support the view that score gaps emerge primarily in early childhood, before children enter school. But the results also suggest caution because so many conclusions are sensitive to measurement artifacts.

Questions and Artifacts in Studies of Gap Growth

In our empirical analyses, we will ask the question in the title: “Do test score gaps grow fastest before, during, or between the school years?” We ask this question about several test score gaps: the gaps between boys and girls; the gaps between black, white, and Hispanic children; the gaps between the children of more- and less-educated mothers; the gaps between children in poor and nonpoor families; and the gaps between high-poverty and low-poverty schools.

We highlight two measurement artifacts, which we now introduce in connection with two well-known seasonal learning studies.

Artifact 1: Test Score Scaling

The first artifact is test score *scaling*. Scaling is the mathematical method by which a pattern of right and wrong answers is transformed into a test score. A test score is commonly interpreted as a measure of a child’s *ability*, which in psychometrics is defined as the child’s level of knowledge or skill at the time of the test. In sociology and economics, the word “ability” is sometimes used to represent a fixed or permanent trait, but under the psychometric definition, ability can change as children learn and grow.

When we use a score to compare gaps in ability, we implicitly assume that the score uses an *interval* scale of children’s ability—so that, say, a one-point score gap represents the same difference in ability at the bottom of the scale as at the top. We also assume that the score is *vertically aligned* across ages—so that, say, a one-point gap in kindergarten represents the same difference in ability as a one-point gap in eighth grade. But not all scaling methods try to produce vertical interval measures of ability, and methods that try do not always succeed.

Because not all scales are created equal, the choice of scaling method can have profound implications for whether and when score gaps appear to grow. Psychometricians have known this for a long time, but social scientists have not fully appreciated the implications for the literature on gap growth.

One of the best-known studies of gap growth is the Beginning School Study (BSS) of students who started first grade in Baltimore City Public Schools in 1982 (Alexander and Entwisle 2003). Analyses of the BSS concluded that the gap in reading scores between children of higher and lower socioeconomic status (SES) approximately tripled between first grade and ninth grade (Alexander, Entwisle, and Olson 2007a).

But the BSS used a test—the California Achievement Test (CAT), Form C—whose scaling method did not ensure a vertical interval measure of student ability. In fact, the CAT Form C and its scaling method were being phased out even as the BSS was underway. The CAT Form C was the last version of the CAT to use Thurstone scaling (Gulliksen 2013; Thurstone 1925, 1938). Thurstone scaling had been popular in the 1960s and 1970s, but by the early 1980s, when the BSS began, test publishers had come to recognize problems with Thurstone scaling and were phasing it out in favor of item response theory (IRT), which is the scaling method that is most commonly used today.

The new IRT scales did not agree with the old Thurstone scales on the question of whether children's abilities grew more or less dispersed with age. According to the Thurstone-scaled CAT Form C, between first and eighth grade, the standard deviation (SD) of scores doubled in reading and nearly tripled in math. But according to the very next version of the CAT—the IRT-scaled CAT Forms E and F—the SDs of reading and math abilities did not grow but actually shrank between first grade and eighth (Clemans 1995; Yen, Burket, and Fitzpatrick 1995a, 1995b). In between the two versions of the CAT, the CAT's publisher released the Comprehensive Test of Basic Skills, Form U, which also used IRT scaling and also showed SDs shrinking with age (Yen 1986).

An observer used to the old CAT scores and their spreading SDs expressed dismay at the new IRT scores. He opined that "something's awry in the state of test mark" (Clemans 1993) and argued that the old Thurstone scales must be better than the new IRT scales. These arguments were refuted by psychometricians representing the CAT's publisher, who argued that the new IRT scales were better than the old Thurstone scales (Yen et al. 1995a, 1995b). The psychometricians were correct. Thurstone scaling cannot be better than IRT scaling because the model used to define Thurstone scales is just a special case of the IRT model with some unrealistic restrictions imposed. We will demonstrate this later, in our section on test scaling.

Some critics appealed to "common-sense" arguments, claiming that children's abilities simply must grow more dispersed with age. But psychometricians defending the new IRT scores parried these arguments effectively (Yen 1986:312):

From one point of view it might seem intuitively appealing that the fast learners move farther away from the slow learners as grade increases, so that the variation in reading ability should increase as grade increases. On the other hand, it could be seen as a tremendous revolutionary change for a young child to go from not reading at all to having an "Aha!" experience and starting to read. The changes that go on in high school in terms of polishing comprehension abilities could be seen as minor refinements compared to what typically occurs in first grade. According to this point of view, a valid reading scale should show much more variance in reading ability in first grade than in 10th grade.

The history of scaling in the CAT has implications for the BSS. The BSS started in 1982 with the Thurstone-scaled CAT Form C, and the BSS continued to use Form C even after the IRT-scaled Forms E and F were released in 1985. So the BSS finding that SES gaps tripled between first and sixth grade was likely an artifact

of CAT Form C; it could not have been replicated using Forms E and F or another IRT-scaled test. Likewise, the BSS finding that “a large portion of the [eighth-grade] achievement gap originates over the summer” (Alexander, Entwisle, and Olson 2007b) would be harder to support using an IRT-scaled test that showed little or no gap growth after children began school.

Scaling artifacts have continued to vex studies of gap growth. In the 2000s, several influential studies used data from the Early Childhood Longitudinal Study, Kindergarten Cohort of 1998–1999 (ECLS-K). These studies reported that socioeconomic and racial/ethnic gaps in reading and math grew substantially during the first two years of school (Condrón 2009; Downey et al. 2004; Fryer and Levitt 2006). At the time of these studies, though, the scores that were available for the ECLS-K were not vertical interval measures of ability (Koretz 2009; Reardon 2008). Later releases of the ECLS-K included new ability scores that had a stronger claim to vertical interval scaling (Najarian, Pollack, and Sorongon 2009). When researchers used the new scores, the apparent timing and extent of gap growth changed, and key results that had been obtained using the old scores could not be replicated (Koretz 2009; Reardon 2008).

Artifact 2: Changes of Test Form

In addition to concluding that the reading gap between high- and low-SES children tripled between first and eighth grade, the BSS concluded that all of that gap growth occurred during summer vacations. The SES gap did not grow at all during the school years (Alexander et al. 2007a).

But this finding, too, may be an artifact of measurement. The artifact comes from changes in test form. The BSS used CAT Form C, which was a “fixed-form” paper test. In first grade, all BSS children filled in a paper form that contained a fixed or unvarying set of questions in fall and spring. Then, after the summer break, students started second grade and switched to a new form with harder questions. This pattern continued into later years: Children used the same form in fall and spring, then switched forms after the summer break, when a new school year began.

What this means is that estimates of summer learning were confounded with changes in test form. School-year learning was estimated by comparing fall and spring scores on the *same* test form, but summer learning was estimated by comparing spring and fall scores on *different* test forms. The fact that scores grew more dispersed between one grade and another could mean that children’s true abilities diverged over the summer—or it could mean that true *abilities* did not diverge, but *scores* diverged because the forms used in different grades were different.

In the BSS, scores on the Thurstone-scaled CAT Form C grew more dispersed with age, and practically all the dispersion took place over the summer. But scores on the next release of the CAT, the IRT-scaled Forms E and F, showed the opposite pattern. Gaps shrank with age, and some of the shrinkage occurred over the summer (Clemans 1993:334–6):

Students at the 98th percentile . . . show some growth over each school year. However, except for the first two grades, their scores decrease over the summer vacation . . . Pupils at the second percentile, on the other hand . . . show

significant gains in achievement during each of the 11 vacation periods (or from one test form to another), and from the fifth grade on, show either no gains or an actual loss during the school year, when the same test forms are used.

In other words, on the CAT Forms E and F, the gaps between high- and low-achieving children grew during the school year and shrank over the summer. This pattern was exactly the opposite of the summer gap growth that the BSS observed on the CAT Form C.

The BSS was not the only seasonal learning study to use fixed forms that changed after the summer. Practically all seasonal studies did so from the 1960s into the 1990s. The entire summer learning literature reviewed in a 1996 meta-analysis was potentially vulnerable to artifacts related to scaling and changes of test form (Cooper et al. 1996).

Although fixed-form tests are still widely used, modern studies increasingly use adaptive tests (Gershon 2005), which are less vulnerable to artifacts that might affect estimates of summer learning. Adaptive tests do not ask the same questions of all children; instead, adaptive tests estimate children's ability and adapt to it, asking harder questions of children whose earlier answers suggested that they have greater ability. So adaptive tests do not need to change abruptly at the start of a new school year. Using adaptive tests, summer learning can be estimated in the same way as school-year learning, with no change of test form and less concern about artifacts.

Preview

In the rest of this article, we will estimate score gaps at the start of elementary school and estimate the growth and shrinkage of score gaps through the end of eighth grade. We will ask whether gap growth and shrinkage occur mainly during the school years or mainly during summer vacations. To evaluate the replicability of our conclusions, we will compare results from three different data sets. These data sets include children from different cohorts and different areas of the United States. Different data sets also take different approaches to measurement—some more valid than others.

We will highlight which results are sensitive to measurements artifacts and which results are replicable. The results will show that many conclusions—about which gaps grow or shrink as children get older, and about whether gaps change more during the school years or during summer vacations—are not replicable across different data sets and scores. Many, though not all, of these nonreplicable results can be traced to measurement artifacts.

Once measurement artifacts are accounted for, the most-replicable result is that gaps are almost full fledged by the time children begin elementary school. Any changes in gaps during later school years and summers are relatively small.

Data

We compared three longitudinal data sets: the Beginning School Study (BSS), the Early Childhood Longitudinal Study, Kindergarten Cohort of 1998-99 (ECLS-K),

Table 1: Test characteristics.

	BSS	ECLS-K	GRD
Test	CAT Form C	Custom	MAP
Domains	Reading comprehension Math concepts	Reading Math	Reading Math
Scaling	Thurstone	IRT (three-parameter logistic)	IRT (one-parameter logistic)
Scales	Number-right score	Ability score θ Scale score (number right)	Ability score $10\theta+200$
Medium	Paper	Paper in eighth grade, computer earlier	Computer
Assessment method	Fixed-form test	Two-stage adaptive test	Continuously adaptive test

and the Growth Research Database (GRD) maintained by the Northwest Evaluation Association (NWEA). The BSS and ECLS-K have been used in widely cited studies of gap growth over the past 15 to 25 years. The GRD is newer.

The three data sets drew from different populations of children. The data sets also used different tests with different scaling methods, different approaches to administration (fixed form vs. adaptive), and different testing schedules.

Tests

Features of the tests used in the three data sets are summarized in Table 1. The BSS used sections from Form C of the California Achievement Test (CAT). The GRD used the Measures of Academic Progress (MAP) tests. The ECLS-K used custom tests developed by psychometricians at the Educational Testing Service (Najarian et al. 2009; Rock and Pollack 2002). The tests differed in several ways.

Fixed-form versus adaptive testing. One difference between the tests is that the BSS used fixed-form testing, whereas the ECLS-K and GRD used adaptive testing. As discussed earlier, fixed-form testing can bias comparisons of school-year and summer learning because school-year learning is estimated by comparing scores from the same test form, whereas summer learning is estimated by comparing scores from different test forms used before and after the summer.

The GRD and ECLS-K were not vulnerable to change-of-form artifacts because they used adaptive testing. In adaptive testing, different students answered different items drawn from a large item pool. The difficulty of the items presented to a particular student was calibrated according to an estimate of the student's ability. The ECLS-K used two-stage adaptive testing, in which a first-stage routing test provided an initial estimate of ability, which was used to assign students to a second-stage test of appropriate difficulty (Najarian et al. 2009). The GRD used continuously adaptive testing, which revised each student's ability estimate after each response and used the revised ability estimate to choose the difficulty of the next question (Northwest Evaluation Association 2010). Either way, adaptive testing ensured that students got more difficult questions as they grew older and more capable, with no need to change test forms abruptly at the beginning of a new school year.

Test scaling. The tests also differed with respect to scaling. The BSS started in 1982 with the CAT Form C, which, along with its contemporary Form D,¹ was the last CAT form to use Thurstone scaling. The BSS continued to use Form C even after the IRT-scaled Forms E and F were published in 1985 (Entwisle, Alexander, and Olson 1997; Yen 1986). The ECLS-K and GRD both used IRT scaling, but not exactly the same type. Specifically, the GRD used a one-parameter logistic IRT model, also known as a Rasch model (Northwest Evaluation Association 2010), whereas the ECLS-K used a three-parameter IRT model (Najarian et al. 2009).

To understand the differences between these scales, let's review the underlying models and assumptions. In a given domain (e.g., reading or math), scaling tries to estimate the current ability θ_s of student s from the student's answers to a set of items i . Each item has a difficulty d_i , which is defined on the same scale as θ_s , so that we can compare the difficulty of an item to the ability of the student who is trying to answer it. A student's probability of responding correctly to an item is modeled by the item response function (IRF), which is a function of student ability, item difficulty, and possibly other parameters.

Figure 1 illustrates the IRFs for the Thurstone, one-parameter logistic, and three-parameter logistic models by plotting the probability of a correct response to an easy item and a hard item. Under the Thurstone IRF, the probability of a correct response steps up from 0 to 1 as soon as a student's ability exceeds the difficulty of the item (Thurstone 1925):

$$P(\text{correct}) = \begin{cases} 1 & \text{if } \theta_s > d_i \\ 0 & \text{otherwise} \end{cases} \quad (\text{Thurstone IRF.})$$

Under the one-parameter logistic IRF, the probability of a correct response increases more gradually, following a logistic function (also known as inverse logit function) that rises as a student's ability increases, approaches, and then exceeds the difficulty of an item (DeMars 2010):

$$P(\text{correct}) = \text{logit}^{-1}(\theta_s - d_i) \quad (\text{Rasch or one-parameter logistic IRF.})$$

Under a three-parameter logistic IRF, items differ not just in difficulty but in discrimination a_i and guessability c_i (DeMars 2010):

$$P(\text{correct}) = c_i + (1 - c_i) \text{logit}^{-1}(a_i(\theta_s - d_i)) \quad (\text{three-parameter logistic IRF.})$$

Guessability (c_i) is the probability that a student will give a correct response if they don't actually know the answer. For example, on a multiple-choice item with four plausible options, guessability would be $c_i = 1/4$, which is the value assumed for the illustration in Figure 1. Discrimination a_i is a slope that reflects how quickly the probability of a correct response increases with student ability. If an item discriminates poorly, then the probability of a correct response rises slowly with ability, but if an item discriminates well, then the probability of a correct answer rises quickly with ability. The illustration in Figure 1 assumes that the hard item discriminates twice as well as the easy one ($a_i = 2$ vs. 1).

Probability of answering item correctly

according to 3 different item response functions (IRFs)

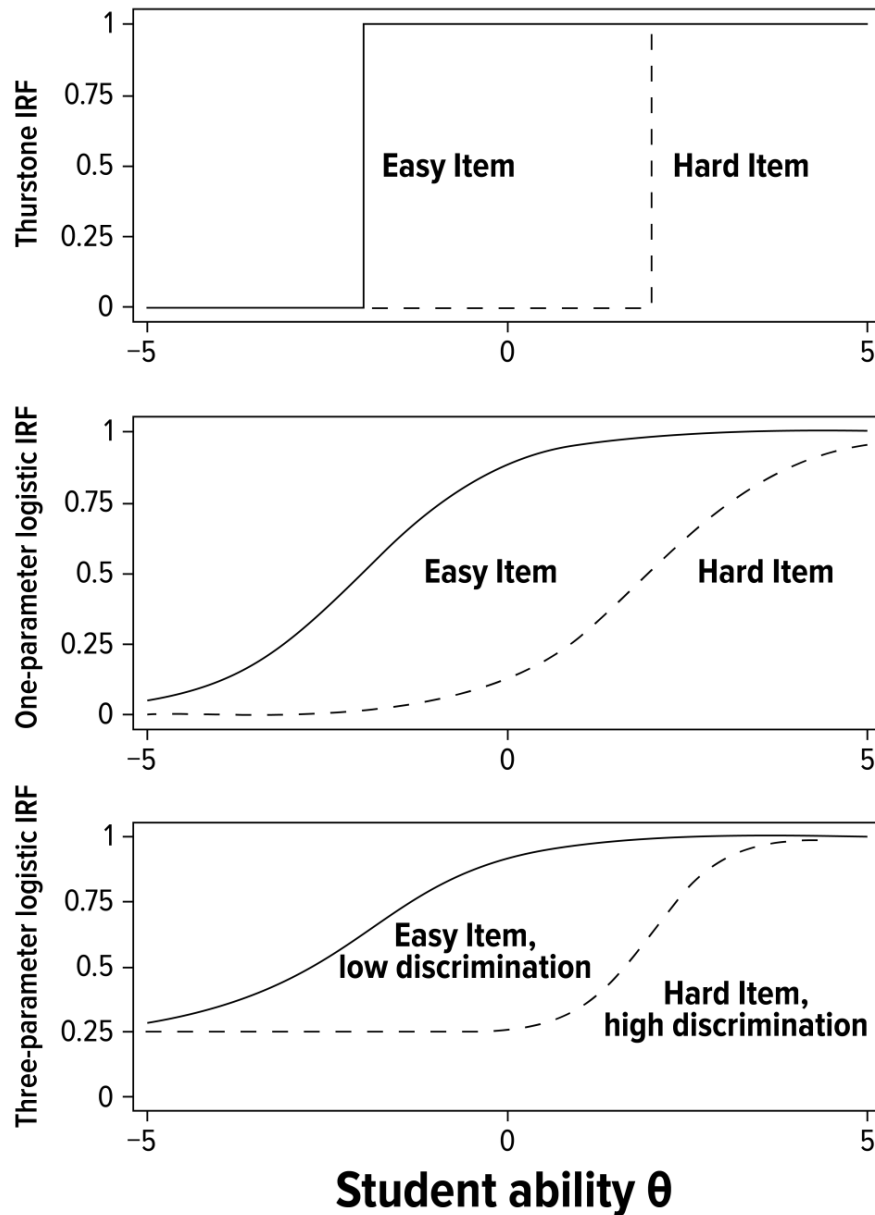


Figure 1: Item response functions (IRFs) for three scaling methods: Thurstone scaling (used by the BSS), one-parameter logistic scaling (used by the GRD), and three-parameter logistic scaling (used by the ECLS-K). For each item response function, the figure illustrates the probability of a correct answer to an easy item and a hard item ($d_i = -2$ and $+2$) for students with abilities θ ranging from low to high (-5 to $+5$). The figure for the three-parameter IRF assumes that the hard item is twice as discriminating as the easy item and that a student who doesn't know the correct answer has a 1 in 4 chance of guessing correctly.

Comparing the three IRFs, we see that the one-parameter logistic model is like the three-parameter logistic model if no items were guessable ($c_i = 0$) and all items discriminated equally ($a_i = 1$). The Thurstone model is like the three-parameter logistic model if no items were guessable and all items discriminated perfectly ($a_i \rightarrow \infty$). When a one-parameter logistic model is used, test developers try to choose items that come close to satisfying its assumption of equal discrimination, but when a Thurstone model is used, there is no way to satisfy its assumption of perfect discrimination. That assumption is unrealistic.

Violations of a model's assumptions can bias the estimated distribution of student ability. For example, suppose that the tests items given in second grade discriminate better than items given in first grade. The three-parameter logistic model can model this, but the Thurstone and one-parameter logistic models cannot because they assume that all items discriminate equally. So the Thurstone and one-parameter logistic models will effectively mistake the increase in item discrimination for an increase in the dispersion of student ability.² This may help to explain why SDs grew with age on the Thurstone-scaled CAT Form C, but SDs did not necessarily grow with age on IRT-scaled tests.

Ability scores versus number-right scores. Later releases of the ECLS-K provided IRT ability estimates $\hat{\theta}_s$, but early releases provided only a number-right score \hat{R}_s (which the ECLS-K somewhat confusingly called a “scale score”³ [Rock and Pollack 2002]). The number-right score estimated how many correct answers the child would have given had they been presented with every item in the full item pool—instead of just the sample of items drawn adaptively from the item pool for the first- and second-stage assessments.

Unfortunately, a number-right score is not, in general, a valid estimate of student ability. The number-right score is a function of ability, but it is also a function of the difficulty of questions in the item pool.⁴ Figure 2 shows the relationship between ability and number-right scores in the ECLS-K (cf. Reardon 2008). The relationship is S shaped, reflecting the cumulative distribution of item difficulty.⁵ As a child's ability grows, their number-right score grows slowly at first because the item pool has few easy questions. Then the number-right score reaches an inflection point and grows more quickly as students reach the moderate ability levels, where most questions are pitched. Because the number-right score is not a linear function of ability, the number-right score cannot be an interval measure of ability.

Several artifacts are possible because of the S-shaped relationship in Figure 2. One artifact is that the gaps between advantaged and disadvantaged students can grow on the number-right scale even when they do not grow on the ability scale. Figure 2 illustrates this graphically by showing the gap between students who do and do not qualify for meal subsidies. The gap is shown at two time points: the fall of kindergarten and the spring of first grade. On the number-right score, it appears that the gap grew over this period, but on the ability score, it appears that the gap shrank.

The S-shaped relationship can also change the apparent *timing* of gap growth. Advantaged students start with higher scores, so they will be the first to reach the inflection point in the S. When advantaged students reach the inflection point, they will appear to pull away from disadvantaged students on the number-right score,

Relationship between ability and number-right scores in the ECLS-K

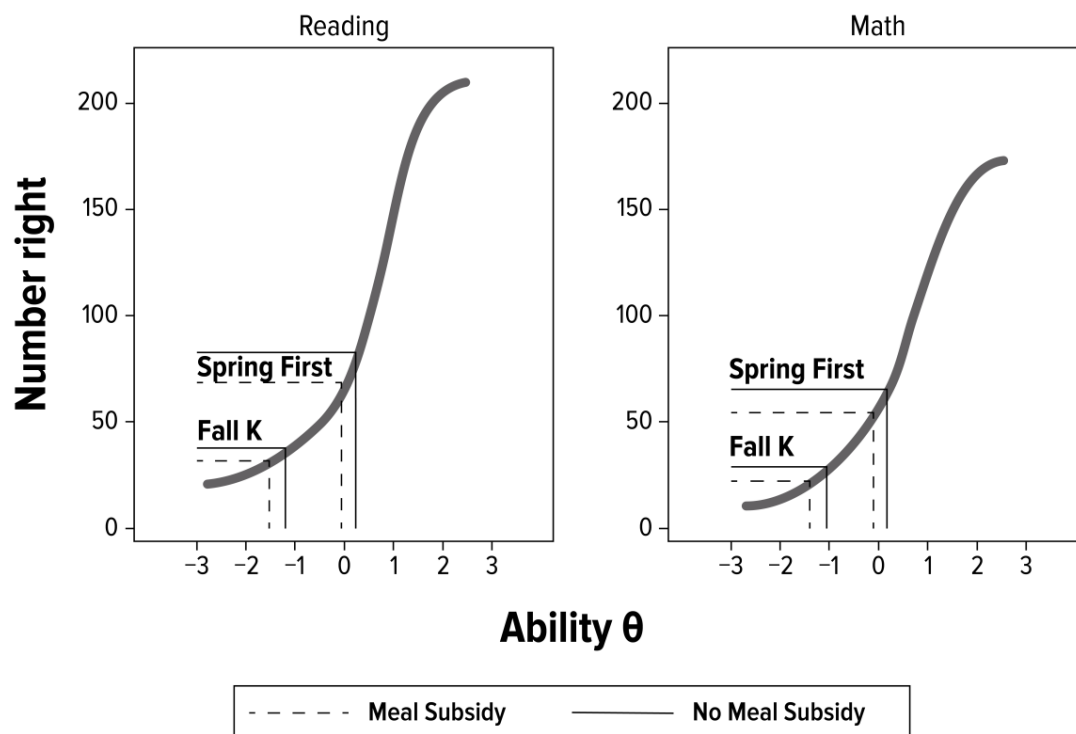


Figure 2: Curved relationship between ability and number-right scores in the ECLS-K. Between the fall of kindergarten and the spring of first grade, the gaps between children with and without meal subsidies appears to grow on the number-right score even though it shrinks on the ability scale.

even if they are not pulling away on the ability score. If advantaged students reach the inflection point near the start of summer vacation, then gaps on the number-right score will open more quickly during summer than during the previous school year. But if advantaged students reach the inflection point near the *end* of summer vacation, then gaps on the number-right score will open more quickly during the school year than during the previous summer. Again, this can happen even if gaps in ability are not changing at all.

Notice that scaling artifacts may affect different comparisons differently. The moment when advantaged students reach the inflection point may be different for reading than for math. It may be different for different groups of advantaged students (e.g., different for white students than for students of college-educated mothers), depending on where exactly each group starts out in the test score distribution. So it may appear that the reading gap opens during summer and the math gap opens during school. Or it may appear that the white-black gap opens during school, but the gap between the children of more- and less-educated mothers opens during summer. Yet all these differences may be illusory, artifacts of the

number-right score and its nonlinear relationship to ability. Fewer artifacts are possible if we use the ability score.

Despite its shortcomings, the number-right score was widely used in early longitudinal analyses of the ECLS-K (Condrón 2009; Downey et al. 2004; Downey, von Hippel, and Hughes 2008; Fryer and Levitt 2006; Reardon 2003; von Hippel 2009). The shortcomings of the number-right score were not widely appreciated, and the ability scores had not yet been released. One study reverse engineered a version of the ability scores (Reardon 2008), and shortly after that, the ability scores became part of the standard ECLS-K release (Najarian et al. 2009). Yet even after the ability scores were released, some investigators continued to use the number-right scores. In our analyses, we will highlight which results change when we use ability scores instead of number-right scores.

The GRD MAP tests scored children using “Rasch units” (RITs), which are a linear transformation of ability: $RIT = 10\theta + 200$ (Northwest Evaluation Association 2010). Because any linear transformation of ability is still an interval measure of ability,⁶ the use of the RIT metric did not change any substantive conclusions about the growth or shrinkage of ability and ability gaps. We call the RIT score the ability or 10θ score to emphasize its similarity to the ability score θ used by the ECLS-K. The major advantage of the RIT scale is cosmetic. The RIT score looks like other scores that parents and teachers are used to: It takes positive three-digit values and increases by five to 20 points per year, whereas raw ability scores θ can take negative values and typically increase by less than one point per year.

Different scales suggested different patterns of dispersion with age. Some scales spread substantially with age, some spread only a little, and some did not spread at all. Specifically, the SDs of the BSS Thurstone scores and the ECLS-K number-right scores more than doubled between first and eighth grade, whereas the SDs of the GRD ability scores grew by less than half, and the SDs of the ECLS-K ability scores actually shrank. For details, see Table A1 in the supplement, which gives the means and SDs of every score on every measurement occasion from the fall of kindergarten through the spring of eighth grade.

Test content. The tests from the BSS, ECLS-K, and GRD also differed in content. The BSS reading scores were limited to the Reading Comprehension section of the CAT and did not include CAT sections on Reading Vocabulary or (in grades 1–3) Phonic Analysis and Structural Analysis. Likewise, the BSS math scores were limited to the CAT section on Mathematics Concepts and Applications section and did not include the CAT section on Mathematics Computation (compare CTB/McGraw-Hill [1979] to Entwistle et al. [1997]).

The contents of the ECLS-K and GRD MAP tests were broader. The ECLS-K test content was derived from the framework used for the National Assessment of Educational Progress. Tested reading skills ranged from basic skills and vocabulary to understanding, interpretation, reflection, and “critical stance.” Tested math skills ranged from number sense, properties, and operations through measurement; geometry and spatial sense; data analysis, statistics, and probability; and patterns, algebra, and functions (Najarian et al. 2009).

The content of the GRD MAP tests was aligned with state curricula standards, many of which were aligned with the common core. MAP items were screened

for validity and bias through a multistage process, including review by a “Sensitivity and Fairness panel with educators from culturally diverse backgrounds” (Northwest Evaluation Association 2010).

Testing schedule. Any study that seeks to separate school learning and summer learning must measure children near the beginning and end of at least one school year and one summer. Our three data sets differed in the number of school years and summers that they covered.

The BSS tested students twice per year, in the fall and spring of first grade through sixth grade, and then once per year, in the spring of seventh grade and eighth grade. This schedule let us estimate learning rates for every school year and summer from first grade to sixth grade. After sixth grade, we could not separate summer learning from school-year learning, but we could still estimate average learning rates over periods of one to two full years.

The ECLS-K tested students in the fall and spring of kindergarten and first grade, and then every two to three years in the spring of third grade, fifth grade, and eighth grade. (The fall first-grade test was limited to a random 30 percent subsample of schools.) This schedule let us separate school-year from summer learning rates from the start of kindergarten through the end of first grade. After first grade, we could not separate summer learning from school-year learning, but we could still estimate average learning rates over periods of two or three years.

Our GRD extract followed an accelerated longitudinal design (Raudenbush and Chan 1992). During two consecutive school years (2008–2009 and 2009–2010), we followed eight cohorts of students: We followed one from kindergarten through first grade, one from first grade through second grade, and so on through the oldest cohort, which we followed from seventh grade through eighth grade. The advantage of this accelerated design was that we could estimate learning over every school year and summer from kindergarten through eighth grade, in less time and with less attrition than if we followed a single cohort for nine years.

Another advantage of the GRD is that many students were tested in the winter as well as in the fall and spring. The fall, winter, and spring average test scores fit a straight line within each school year, which confirmed the linear assumption of the growth model that we will specify later.

Ideally, tests would be given on the first and last day of the school year, but none of the studies followed that schedule. The GRD and ECLS-K recorded the exact date on which each student took each test; most commonly, students took the fall test in October and the spring test in May, but this varied across schools. The BSS also gave tests in October and May (Entwisle et al. 1997), but the exact test dates did not appear in the data. In analyses where we needed exact test dates for the BSS, we plugged in October 15 and May 15. Plugging in other dates from October and May did not materially affect our results.

To separate summer learning from school-year learning, we needed the dates on which each school year started and ended. The ECLS-K made school start and end dates available as restricted data. The GRD did not include school start and end dates, but we found them by looking up the calendars of participating schools on the Internet. The BSS did not include school start and end dates either, but staff at the Baltimore City Public School System told us that during 1982 through 1990,

when the BSS was running, Baltimore's public schools opened on the day after Labor Day. We assumed that Baltimore schools closed 284 days later, in mid-June, because 284 is the average number of calendar days between school start and end dates in the ECLS-K.

Missing values. All three data sets had missing test scores and covariates. We filled in missing values using a multiple imputation model, which we describe in the supplement. Imputed values helped to smooth our graphs of average scores by age, which would be more jagged if different children were missing scores on different occasions.

Children and Schools

Tables 2a–2d report the school, child, and family characteristics of the BSS, ECLS-K, and GRD, highlighting some of the strengths and limitations of each data set.

The BSS is the oldest data set. It began in the fall of 1982 with 790 first graders sampled from 20 public schools in the Baltimore City Public School System (Alexander and Entwisle 2003). The 20 schools were sampled within strata defined by school racial and socioeconomic composition. BSS students were approximately representative of the 1982 enrollment of the Baltimore City Public Schools, which meant that they were, on average, somewhat poorer than the U.S. population. The BSS was limited to black and white children; only 2 percent of Baltimore residents were Hispanic or Asian in 1982.

The BSS is important to the summer learning literature, and it provided rich data on participating schools and families. It is somewhat dated, however, and its sample was narrow, containing only black and white children from a single high-poverty urban district where more than half of first graders were black, nearly two-thirds qualified for meal subsidies, half had a single parent, and more than two-fifths had a mother who had not completed high school. The BSS sample was also somewhat small, so even large differences between school and summer gap growth sometimes failed to achieve statistical significance.

The ECLS-K is a much larger and more diverse sample. It began with a U.S. probability sample of children who attended kindergarten in the fall of 1998 (National Center for Education Statistics 2009). The ECLS-K was a cluster sample that clustered children within schools, and oversampled private schools and areas with high enrollments of Asian Americans. (Sampling weights compensate for the oversampling.) Our analytic ECLS-K sample excluded 27 year-round schools, where children attended school much of the summer (von Hippel 2016). After this exclusion, our analytic sample totaled 17,779 children in 977 schools.

Our GRD extract was even larger, containing 177,549 students in 389 schools, 25 school districts, and 14 U.S. states (Alaska, Colorado, Indiana, Kansas, Kentucky, Minnesota, New Mexico, Nevada, Ohio, South Carolina, Texas, Washington, Wisconsin, and Wyoming). Although the sample was not nationally representative (it is limited to districts that use NWEA's MAP tests), the sample districts were quite diverse. Across GRD districts, the percentage of students who were white ranged from 16 percent to 98 percent, the percentage of students who received subsidized meals ranged from 12 percent to 73 percent, and the average math percentile score

Table 2: Characteristics of the three samples.

	BSS	ECLS-K	GRD
a. Periods and sample sizes.			
Years of testing	1982–1990	1998–2007	2008–2010
Students	825	17,825	177,549
Schools	20	977	419
Districts	1	438	25
States	1	41	14
b. Child characteristics.			
Race/Ethnicity			
White (non-Hispanic)	45%	58%	53%
Black	55%	16%	18%
Hispanic (any race)	0%	18%	19%
Asian, Native Hawaiian, Pacific Islander	0%	4%	6%
American Indian	0%	2%	2%
Multiethnic	0%	2%	3%
Female	50%	49%	49%
c. Family characteristics.			
Mother's education			
Less than a high school diploma (or equivalent)	42%	14%	—
At least a high school diploma (or equivalent)	58%	86%	—
At least a bachelor's degree	—	27%	—
Number of parents in the home	1.49	1.73	—
Poor (free or reduced-price lunch)	65%	45%	—
d. School characteristics.			
Average % subsidized lunch (free or reduced price)	51%	37%	52%
High-poverty schools (>40% subsidized lunch)	60%	43%	68%
Title I	—	57%	78%
Private	0%	30%	0%
City	100%	37%	40%
Suburb or town	0%	37%	36%
Rural	0%	26%	24%

Note: Missing variables are indicated by dashes (—). Sample weights were used in the ECLS-K, but the other data sets did not have sample weights because they were not probability samples.

ranged from 22 to 70. Some districts were in large cities, whereas others were in suburbs, small towns, or rural areas.

At the school level, the GRD was supplemented with information on school characteristics and demographics from the federal Common Core of Data. At the child level, unfortunately, the GRD was limited; it contained no child or family variables except for race/ethnicity and gender. We requested data on meal subsidies, which school districts collect routinely, but NWEA staff told us that meal subsidy status was not integrated into the GRD.

Methods

We used all three data sets to estimate the size and growth or shrinkage of test score gaps between advantaged and disadvantaged children through school years and summers from the start of kindergarten (first grade in the BSS) until the end of eighth grade. Our overarching questions were these:

1. How large were gaps at the start of kindergarten (first grade in the BSS)?
2. How much did gaps grow or shrink by the end of eighth grade?
3. To the degree that gaps grew, was the growth faster during the summers or during the school years?

We evaluated whether the answers to these questions were consistent across different data sets, different populations, different tests, different subjects, different scales, and different measures of advantage.

Measures of Advantage

We estimated gaps between advantaged and disadvantaged children. Advantage and disadvantage were defined in different ways. Because each data set has somewhat different measures of advantage, not every data set can be used to estimate every gap.

Past studies of the BSS and ECLS-K measured SES by using custom indices that combine measures of parental education, income, employment, and occupation (Alexander et al. 2007b; Downey et al. 2004). We could not use these indices here because the measures that they required were survey-specific and could not be replicated across different surveys. For example, the ECLS-K used an SES index that we cannot replicate in the BSS, and the BSS used an SES index that we cannot replicate in the ECLS-K.

Instead of using composite SES indices, we used three simple SES components to estimate different gaps:

1. We estimated the gap between poor and nonpoor children, where poor children were defined as children eligible for meal subsidies (free or reduced lunch). Note that this gap was not available in the GRD, which lacked data on individual students' meal subsidy status.
2. We estimated the gap between low-poverty and high-poverty schools. Our original intention was to use Title I status, but that was not available in the BSS, so instead we defined high-poverty schools as schools where at least 40 percent of students qualified for meal subsidies. The percentage of students with meal subsidies was available in all three data sets.
3. We estimated the gap between the children of more- and less-educated mothers. We defined three levels of maternal education at the time of the first tests: mothers who had not completed high school, mothers who had at least a high school diploma or equivalent, and mothers who had completed at least

a bachelor's degree. Maternal education was unavailable in the GRD. It was available in both the ECLS-K and the BSS, but in the BSS, only a couple of mothers reported having a bachelor's degree, so we grouped them with other high school graduates.

All three of these SES measures were available in the BSS and ECLS-K, but only school poverty was available in the GRD.

In addition to gaps related to SES, we estimated gaps related to race and ethnicity:

1. We estimated the gap between white and black students. This gap could be estimated in all three data sets.
2. We estimated the gap between Hispanic students and non-Hispanic white students. This gap could be estimated in the ECLS-K and the GRD but not in the BSS, where there were no Hispanics.

We coded each gap so that it was positive. For example, instead of the black–white gap, we estimated the white–black gap, defined as the average number of points by which white students led black students. Likewise, we defined the gap between nonpoor and poor students as the average number of points by which students who did not receive meal subsidies led students who did. When gaps were defined in this way, gap growth was always a positive number, and gap shrinkage was always negative. This convention simplified interpretation.

Standardized Gaps, Adjusted for Reliability

We graphed the average scores of advantaged and disadvantaged students against average test dates from the fall of kindergarten (or first grade) through the spring of eighth grade. On these graphs, we could eyeball the size of achievement gaps and identify periods when gaps were growing or shrinking rapidly.

Although it was important to see patterns on each test's native scale, it was also desirable to have a common, less scale-dependent metric on which gaps from different data sets could be compared and some artifacts would be reduced. The simplest try would be to standardize each score with respect to the mean and SD of scores for the same test and occasion (e.g., MAP scores in the fall of third grade). The result would be a *standard score Z*, which on each occasion would have a mean of 0 and an SD of 1 regardless of the original scale of the test.

But standard scores are vulnerable to changes in *reliability*. Reliability is classically defined the fraction of variance in test scores that is due to true ability rather than measurement error (Lord and Novick 1968). Reliability is important when scores are standardized because standardized gaps are larger on a more reliable test. To see this simply, imagine the extreme situation in which a kindergarten test is 0 percent reliable, so that all of the variance in test scores is due to measurement error and none is due to true ability. Then students' kindergarten scores will be completely random, and scores will show no average gaps between high- and low-ability students.

Estimates of gap growth can be distorted if the reliability of a test changes with age. For example, between first grade and second grade, the reliability of the BSS reading score increased from 68 percent to 89 percent. Because of this change in reliability, standardized reading gaps would have grown by 14 percent even if gaps in true ability did not grow at all. The BSS tests were less reliable than the ECLS-K and NWEA tests, and all three data sets used tests that increased in reliability after kindergarten or first grade.⁷ Table A2 in the supplement gives reliability estimates for each test on each occasion.

To correct for differences in reliability, we use a standardized and reliability-adjusted (SRA) score $Z_{SRA} = Z / \sqrt{\hat{\rho}}$, where $\hat{\rho}$ was an estimate of reliability (Ho 2009; Reardon 2011). Reliability adjustment spreads out scores from less-reliable tests, so that two groups that differ by one SD in true ability will differ by one unit in Z_{SRA} . Reliability adjustment noticeably increases estimated gaps if reliability is low but not if reliability is high. For example, if a test is 68 percent reliable—like the first reading test in the BSS—then reliability adjustment will increase gaps by 21 percent. But if a test is 90 percent reliable—like the ECLS-K and GRD tests—then reliability adjustment will increase gaps by just 5 percent.

Despite their advantages, reliability-adjusted Z_{SRA} scores do not necessarily offer an interval scale for estimating growth in score gaps. By construction, Z_{SRA} scores produce a scale on which the SD of ability does not change with age. If the SD of *true* ability does change with age, then the meaning of a one-SD score gap will be different for younger children than it is for older children. This difference will not be reflected in reliability-adjusted Z_{SRA} scores.

Multilevel Growth Model

We fit a multilevel growth model to estimate rates of test score growth during each school year and summer (Raudenbush and Bryk 2001; Singer and Willett 2002). We fit the model to test scores on their native scales and also to Z_{SRA} scores. Details of our model are given in the supplement. Briefly, the model adjusted for the fact that children were not tested on the first and last day of each school year. The model had a school random effect to account for the correlation between scores from the same school, as well as an autoregressive parameter to account for the serial correlation between successive tests taken by the same child. Note that because the model included a school random effect, it estimated a weighted average of within-school and between-school gaps (Greene 2011; Wooldridge 2001). This weighted average did not always agree with the simple gap between average scores that we graphed in our figures.

We used a linear combination of model parameters to estimate how much gaps grew between the start of first grade and the end of eighth grade. We also used linear combinations to ask whether gaps grew faster during the school years or during summer vacations. A detailed specification of these linear combinations appears in the supplement.

Results

We focused on two questions:

1. How much did reading and math gaps grow (or shrink) between first grade and eighth grade?
2. Did gaps grow faster (or shrink slower) during the summers or during the school years?

With respect to question 1, we found that gaps typically grew little on the IRT ability scales which were available in the ECLS-K and GRD. Gaps grew more on the BSS Thurstone scale and the ECLS-K number-right scale, but those scales were distorted by measurement artifacts.

With respect to question 2, in the GRD and ECLS-K, we found that summer gap growth was subtle when it was present at all, and seasonal patterns did not consistently replicate across data sets, subjects, grade levels, or measures of advantage. In the BSS, we found large summer learning gaps between high- and low-SES children, but those summer gaps were likely exaggerated by measurement artifacts. The BSS showed summer learning gaps between white and black children.

To save space, we only show figures for reading; figures for math appear in the supplement. Our tables provide a terse summary of estimates from our multilevel growth model; more detailed estimates appear in the supplement.

Gaps between Poor and Nonpoor Children

Figure 3 plots the mean reading scores of poor and nonpoor students—that is, children who received meal subsidies and children who did not. The figure shows results for the BSS and ECLS-K, but not for the GRD, which lacked child-level data on meal subsidy.

How much did gaps grow between first grade and eighth grade? Table 3a uses our multilevel growth model to estimate growth (or shrinkage) in reading and math gaps between the start of first grade and the end of eighth grade.

According to the ECLS-K ability scale, the gaps between poor and nonpoor children did not grow. Instead, gaps shrank by approximately one-quarter in both reading and math.

Less-valid scales exaggerated gap growth. On the ECLS-K number-right scale, which suffered from measurement artifacts, the gap nearly doubled in math and more than tripled in reading. On the BSS Thurstone scale, which also suffered from measurement artifacts, the gap grew more than fourfold in math and more than eightfold in reading.

Standardization and reliability adjustment reduced the appearance of gap growth on the flawed scales. On the best scale—the ECLS-K ability scale—standardization and reliability adjustment suggested that gaps grew by just 6 percent in reading and shrank by 15 percent in math.

Did gaps grow faster during school or during summer? Table 4a asks whether the gap between poor and nonpoor children grew faster (or shrank slower) during the school year or during summer vacation.

Gap between poor and nonpoor children: Reading

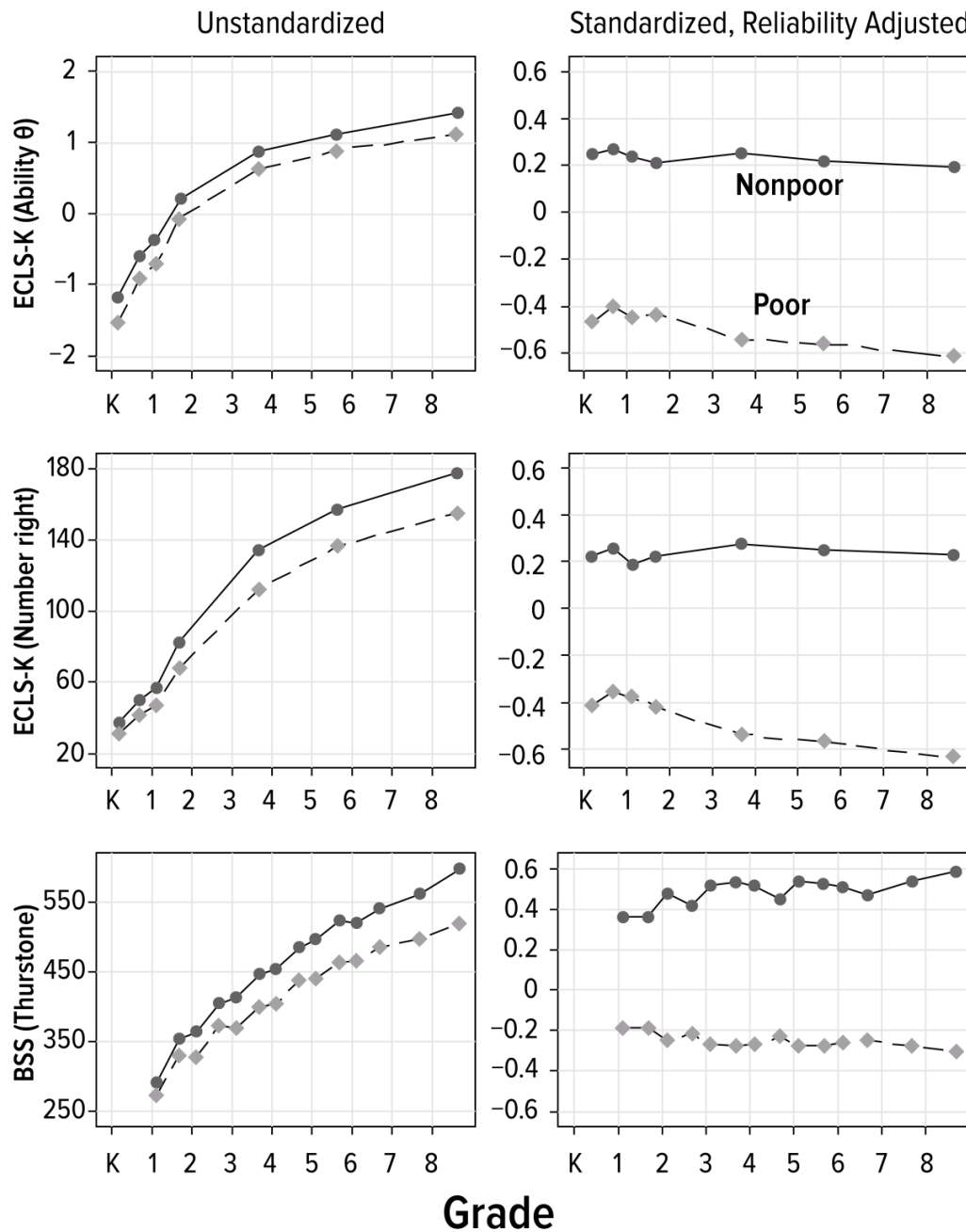


Figure 3: The reading gap between poor and nonpoor children grew substantially on the BSS Thurstone scale and the ECLS-K number-right scale, but those scales suffered from measurement artifacts. On the ECLS-K ability scale, which was a better measure, the gap changed very little from kindergarten through eighth grade. Summer gap growth was substantial only in the BSS, but this may have been an artifact of the BSS's changing test forms at the end of each summer.

Table 3: Percent gap growth (or shrinkage) from first through eighth grade.

Data set (scale)	Unstandardized		Standardized, reliability adjusted	
	Reading	Math	Reading	Math
a. Gap growth between nonpoor children and poor children				
ECLS-K (ability θ)	-26% [†]	-27% [†]	6%	-15% [†]
ECLS-K (number right)	224% [†]	83% [†]	48% [†]	-4%
GRD (ability 10 θ)	NA	NA	NA	NA
BSS (Thurstone)	712% [†]	369% [†]	68% [†]	24%
b. Gap growth between children whose mothers did or did not have a high school diploma (or equivalent).				
ECLS-K (ability θ)	-32% [†]	-34% [†]	-1%	-23% [†]
ECLS-K (number right)	313% [†]	103% [†]	57% [†]	-4%
GRD (ability 10 θ)	NA	NA	NA	NA
BSS (Thurstone)	488% [†]	358% [†]	22%	9%
c. Gap growth between low-poverty and high-poverty schools.				
ECLS-K (ability θ)	-13% [†]	-18% [†]	13% [†]	-9% [†]
ECLS-K (number right)	134% [†]	53% [†]	43% [†]	-2%
GRD (ability 10 θ)	31% [†]	97% [†]	10%	45% [†]
BSS (Thurstone)	244% [†]	218% [†]	13%	9%
d. Gap growth between white and black children.				
ECLS-K (ability θ)	22% [†]	6%	81% [†]	23% [†]
ECLS-K (number right)	536% [†]	137% [†]	173% [†]	35% [†]
GRD (ability 10 θ)	71% [†]	47% [†]	35% [†]	6%
BSS (Thurstone)	-556% [*]	172% [*]	-162% [*]	36%
e. Gap growth between white and Hispanic children.				
ECLS-K (ability θ)	-21% [†]	-38% [†]	2%	-34% [†]
ECLS-K (number right)	514% [†]	47% [†]	70% [†]	-25% [†]
GRD (ability 10 θ)	21% [†]	22% [†]	-4%	-18% [†]
BSS (Thurstone)	NA	NA	NA	NA

* $p < 0.05$, $^{\dagger}p < 0.01$. NA means that the variable or grade range is not available in that data set. A poor child is defined as one who received a school meal subsidy, and a high-poverty school is defined as a school where at least 40 percent of children are poor.

According to the ECLS-K ability scale, reading and math gaps grew during the summer and shrank during the school year. The difference between summer gap growth and school-year gap shrinkage was visually subtle but statistically significant.

The ECLS-K number-right scale, which suffered from measurement artifacts, gave contradictory and inconsistent results. On the ECLS-K number-right scale, the reading gap grew significantly faster during school than during summer, and the math gap grew at approximately the same rate in both seasons.

Table 4: Do gaps grow faster (or shrink slower) during school or during summer vacation?

Grades	Data set (scale)	Unstandardized		Standardized, reliability adjusted	
		Reading	Math	Reading	Math
a. Season when gaps grow faster (shrink slower) between nonpoor children and poor children.					
K-1	ECLS-K (ability θ)	Summer [†]	Summer [†]	Summer [†]	Summer [†]
	ECLS-K (number right)	School [†]	NS	NS	Summer [†]
	GRD (ability 10 θ)	NA	NA	NA	NA
1-6	GRD (ability 10 θ)	NA	NA	NA	NA
	BSS (Thurstone)	Summer*	Summer*	Summer*	Summer [†]
K-8	GRD (ability 10 θ)	NA	NA	NA	NA
b. Season when gaps grow faster (shrink slower) between children whose mothers did or did not have a high school diploma (or equivalent).					
K-1	ECLS-K (ability θ)	Summer [†]	Summer [†]	NS	Summer [†]
	ECLS-K (number right)	NS	NS	NS	NS
	GRD (ability 10 θ)	NA	NA	NA	NA
1-6	GRD (ability 10 θ)	NA	NA	NA	NA
	BSS (Thurstone)	NS	NS	NS	NS
K-8	GRD (ability 10 θ)	NA	NA	NA	NA
c. Season when gaps grow faster (shrink slower) between low-poverty and high-poverty schools.					
K-1	ECLS-K (ability θ)	Summer [†]	Summer [†]	Summer [†]	Summer [†]
	ECLS-K (number right)	NS	NS	NS	Summer [†]
	GRD (ability 10 θ)	School*	NS	School [†]	NS
1-6	GRD (ability 10 θ)	Summer [†]	School [†]	School [†]	School [†]
	BSS (Thurstone)	Summer [†]	Summer	Summer [†]	Summer [†]
K-8	GRD (ability 10 θ)	School	School [†]	School [†]	School [†]
d. Season when gaps grow faster (shrink slower) between white and black children.					
K-1	ECLS-K (ability θ)	NS	NS	School*	School*
	ECLS-K (number right)	School [†]	School*	School [†]	School [†]
	GRD (ability 10 θ)	School [†]	NS	School [†]	NS
1-6	GRD (ability 10 θ)	School [†]	School [†]	School [†]	School [†]
	BSS (Thurstone)	NS	NS	NS	NS
K-8	GRD (ability 10 θ)	School [†]	School [†]	School [†]	School [†]
e. Season when gaps grow faster (shrink slower) between white and Hispanic children.					
K-1	ECLS-K (ability θ)	Summer [†]	Summer [†]	Summer [†]	Summer [†]
	ECLS-K (number right)	School*	NS	NS	Summer [†]
	GRD (ability 10 θ)	School*	Summer*	NS	Summer [†]
1-6	GRD (ability 10 θ)	NS	School [†]	NS	Summer*
	BSS (Thurstone)	NA	NA	NA	NA
K-8	GRD (ability 10 θ)	School*	School [†]	NS	Summer [†]

* $p < 0.05$, [†] $p < 0.01$. NS means nonsignificant ($p > 0.05$). NA means that the variable or grade range is not available in that data set. A poor child is defined as one who received a school meal subsidy, and a high-poverty school is defined as a school where at least 40 percent of children are poor.

Clearly, measurement artifacts in the number-right scale distorted conclusions about gap growth. Gaps that shrank on the ability scale could appear to grow on the number-right scale, and gaps that grew faster during summer on the ability scale could appear to grow faster during school on the number-right scale. The children were the same, and the test was the same, but the scale could change the conclusions. Standardization and reliability adjustment reduced but did not eliminate the discrepancies between scales.

The BSS Thurstone scale produced results that looked like a poster for summer learning loss (see Figure 3). Reading and math gaps grew substantially over the summer and held steady during the school year. But the BSS, too, suffered from measurement artifacts. It used a scale that exaggerated gap growth, and it changed test forms after each summer. In the introduction, we learned that the next version of the BSS test, which used different scaling, gave dramatically different impressions about gap growth.

Gaps between the Children of More- and Less-Educated Mothers

Figure 4 plots the mean reading gaps between children with more- and less-educated mothers. The figure shows results for the BSS and ECLS-K, but not for the GRD, which lacked data on maternal education.

How much did gaps grow between first grade and eighth grade? Table 3b uses our multilevel growth model to estimate growth (or shrinkage) in reading and math gaps between children whose mothers did not or did not have the equivalent of a high school diploma.

According to the ECLS-K ability scale, gaps shrank by one-third in both reading and math. By contrast, less-valid scales exaggerated gap growth. On the ECLS-K number-right scale, which suffered from artifacts, the gap doubled in math and more than quadrupled in reading. On the BSS Thurstone scale, which also suffered from artifacts, the gap grew more than fourfold in math and more than fivefold in reading.

Standardization and reliability adjustment reduced but did not eliminate the discrepancies between scales. After standardization and reliability adjustment, gaps grew relatively little even on flawed scales, and the best scale—the ECLS-K ability scale—suggested that the gap shrank by one-quarter in math and neither grew nor shrank in reading.

Did gaps grow faster during school or during summer? Table 4b asks whether the gap between mothers with and without a high school diploma grew faster (or shrank slower) during the school year or during summer vacation.

According to the ECLS-K ability scale, from kindergarten through first grade, the reading and math gaps grew during the summer and shrank during the school year. The difference between summer gap growth and school gap shrinkage was visually subtle but statistically significant.

Scores with measurement artifacts could give a different impression. On the ECLS-K number-right scale and the BSS Thurstone scale, there was no significant difference between school and summer learning gaps.

Gap between children of more- and less-educated mothers: Reading

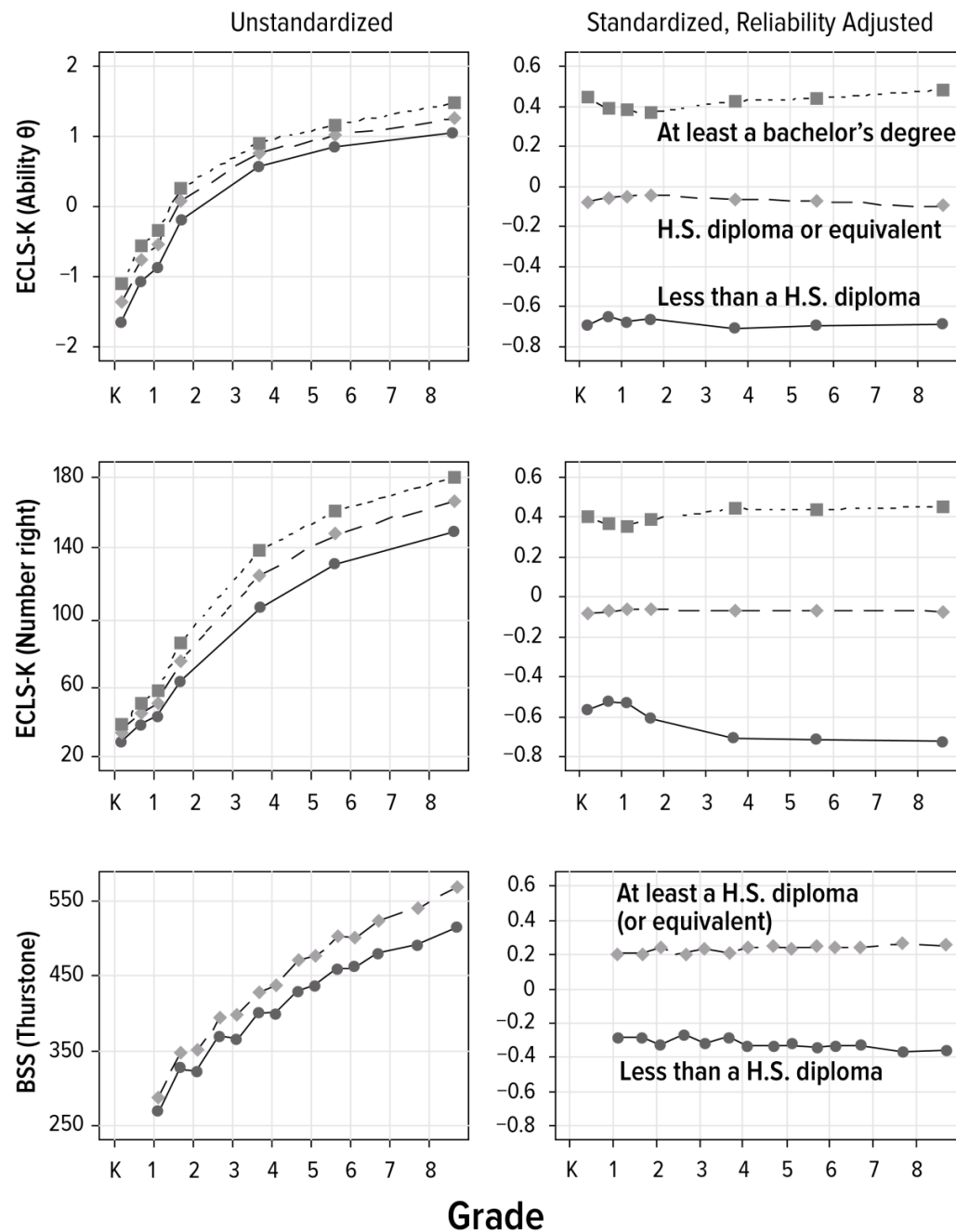


Figure 4: The reading gap between the children of more- and less-educated mothers grew substantially on the BSS Thurstone scale and the ECLS-K number-right scale, but those scales suffered from measurement artifacts. On the ECLS-K ability scale, which was a better measure, the gap changed very little from kindergarten through eighth grade. Summer gap growth was substantial only in the BSS, where summer gap growth may have been an artifact of the BSS's changing test forms at the end of each summer.

Standardization and reliability adjustment did little to change the seasonal patterns.

Gaps between Low- and High-Poverty Schools

The results so far suggest that gaps between socioeconomically advantaged and disadvantaged children changed little after school began. To the degree that gaps changed at all, the results suggest that gaps grew during summer and shrank or held steady during school.

But the evidence so far has been limited. We have only seen one score with a plausible claim to vertical interval scaling (the ECLS-K ability score), and that score provided evidence about only one summer: the summer between kindergarten and first grade.

To broaden the evidence, we now look at a gap that can be estimated in all three data sets: the gap between children in low-poverty and in high-poverty schools. Figure 5 plots those children's mean reading scores in the BSS, the ECLS-K, and the GRD.

How much did gaps grow between first grade and eighth grade? Table 3c uses our multilevel growth model to summarize gap growth (or shrinkage) between first grade and eighth grade.

According to the ECLS-K ability scale, reading and math gaps shrank by about 15 percent. But according to the GRD ability scale, gaps grew; the reading gap grew by one-third, and the math gap nearly doubled. The disagreement between the ECLS-K and GRD ability scales is a little disquieting because both are IRT ability scales with similar claims to vertical interval scaling.

Despite their differences, though, both ability scales typically suggested less gap growth than did scores that suffered from measurement artifacts. On the BSS Thurstone scale, which suffered from measurement artifacts, the reading and math gaps more than tripled. On the ECLS-K number-right scale, which also suffered from artifacts, the reading gap more than doubled, and the math gap grew by one-half.

Standardization and reliability adjustment reduced but did not eliminate the differences between scores.

Did gaps grow faster during school or during summer? Table 4c asks whether the gaps between high- and low-poverty schools grew faster (or shrank slower) during the school year or during summer vacation.

According to the ECLS-K ability scale, reading and math gaps grew during summer and shrank during school. But again, this conclusion was limited to the period from kindergarten through first grade. And again, the ECLS-K ability scale did not agree with the GRD ability scale.

Seasonal gap patterns were inconsistent in the GRD. From kindergarten through first grade, math gaps grew faster in summer, whereas reading gaps grew at about the same rate during summer as during school. But in grades one through six, the pattern changed: Reading gaps grew faster during school, whereas math gaps grew faster during summer. Overall, from kindergarten through eighth grade, math gaps

Gap between high- and low-poverty schools: Reading

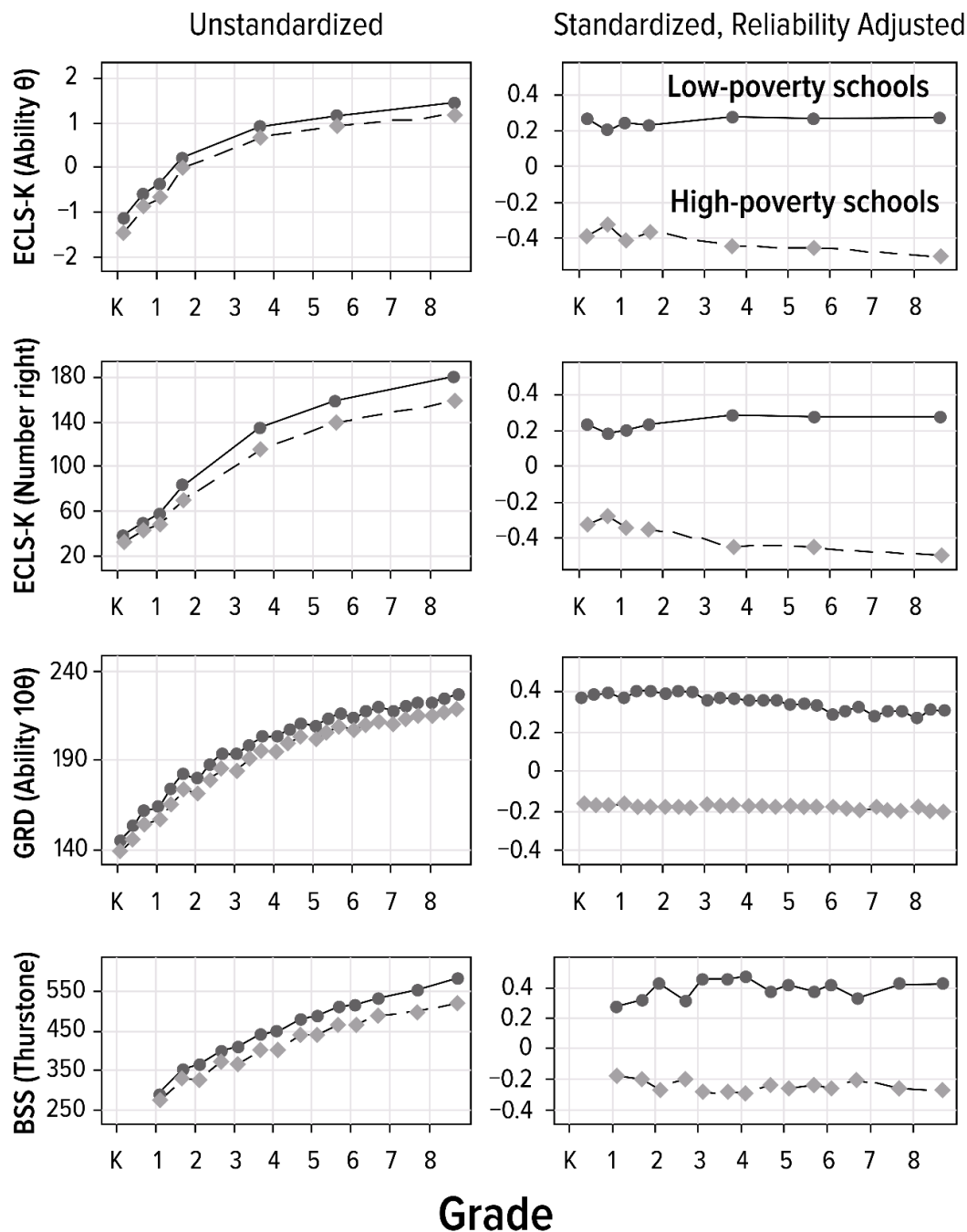


Figure 5: The reading gap between high- and low-poverty schools grew substantially on the BSS Thurstone scale and the ECLS-K number-right scale, but those scales suffered from measurement artifacts. On the ECLS-K and GRD ability scales, which were better measures, the gap changed little from kindergarten through eighth grade. Summer gap growth was substantial only in the BSS, but this may have been an artifact of the BSS's changing test forms at the end of each summer.

grew during the school year and shrank during summer. But reading gaps grew at about the same rate during school as during summer.

Scores with measurement artifacts could distort seasonal patterns. According to the ECLS-K number-right score, which suffered from artifacts, there was no significant difference between school and summer gap growth. According to the BSS Thurstone scale, which also suffered from artifacts, there was substantial gap growth during summer and none during school.

In short, the results for the gap between high- and low-poverty schools were inconsistent across data sets, scores, and subjects. Standardization and reliability adjustment did not resolve the discrepancies. Limiting the results to IRT ability scales did not resolve the discrepancies, either. In most cases, though, the ability scales suggested less overall gap growth than did the Thurstone or number-right scales.

Gaps between White and Black Students

Figure 6 compares the mean scores of black and white children in reading. Black and white children were present in all three data sets: the ECLS-K, the GRD, and the BSS.

How much did gaps grow between first grade and eighth grade? Table 3d summarizes gap growth (or shrinkage) between first grade and eighth grade.

According to the ECLS-K ability scale, the gaps grew little: by just one-fifth in reading and by just 6 percent in math. But according to the GRD ability scale, gaps grew more: by nearly half in math and by nearly three-quarters in reading. The disagreement between the ECLS-K and GRD ability scales is a little disquieting because both are IRT ability scales with similar claims to vertical interval scaling.

Standardization and reliability adjustment changed the results but did not bring the two ability scales into agreement. Before standardization, ability gaps grew little in the ECLS-K and more in the GRD. After standardization and reliability adjustment, ability gaps grew little in the GRD and more in the ECLS-K.

Although the disagreement between the two IRT ability scales is a little disconcerting, both suggested much less gap growth than did scores with measurement artifacts. According to the ECLS-K number-right score, which suffered from artifacts, the math gap more than doubled and the reading gap more than sextupled. According to the BSS Thurstone scale, which also suffered from artifacts, the math gap more than tripled.

The BSS results were very unusual in reading. According to the BSS, black children started first grade slightly (though not significantly) ahead of white children in reading; black children only fell behind after first grade began. As far as we know, this pattern is unique in the literature, which typically shows black children lagging white children well before the beginning of school. In math, the BSS gave a more typical result, with black children trailing white children from the start of first grade. In the GRD and ECLS-K, black children trailed white children from the start of kindergarten, in both reading and math.

Gap between black and white children: Reading

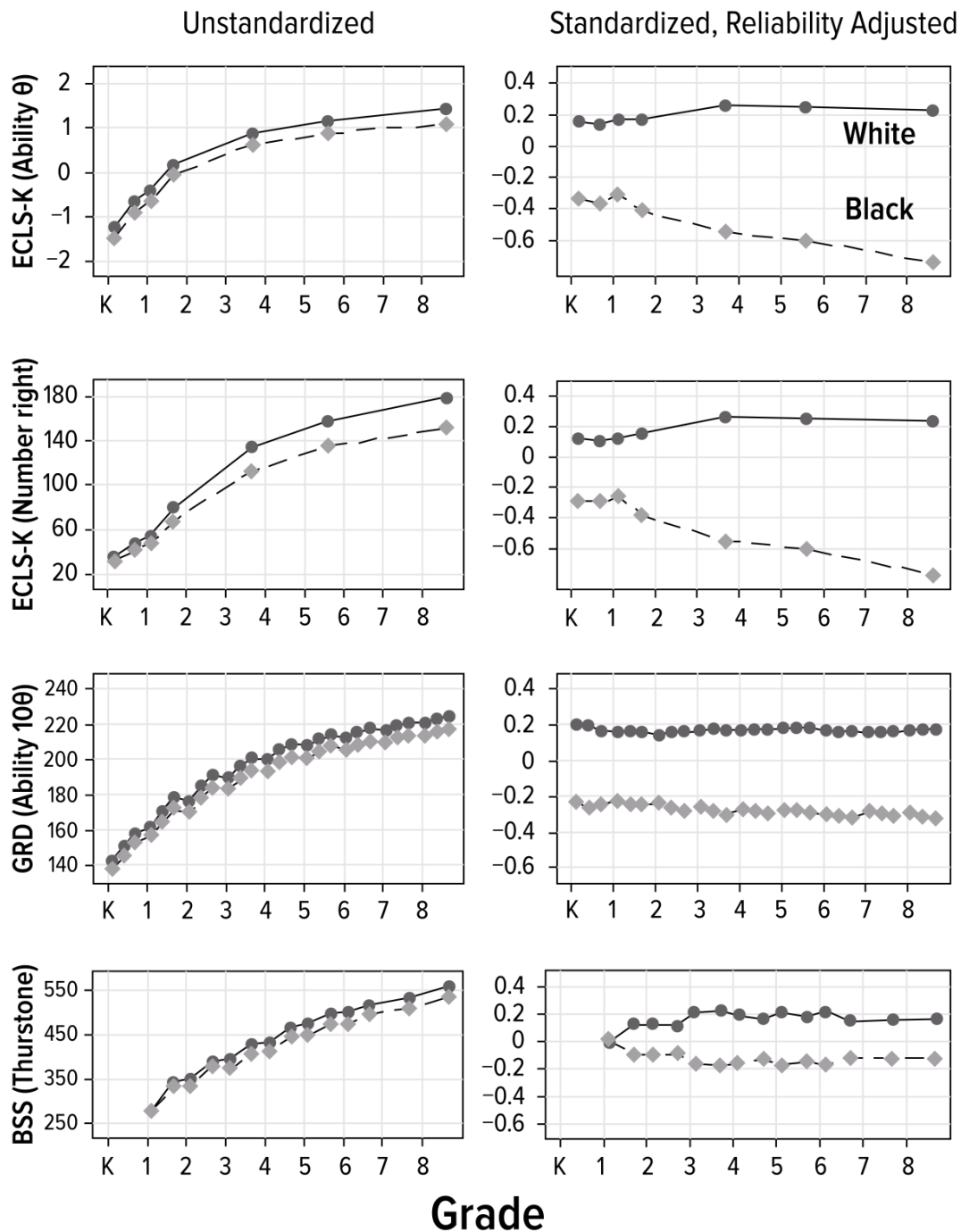


Figure 6: The reading gap between white and black children grew far less in the GRD than in the BSS and ECLS-K. In the BSS, the gap grew during the early years, but in the ECLS-K, the gap was stable in the early years and grew later.

Did gaps grow faster during school or during summer? Table 4d asks whether the white–black gaps grew faster (or shrank slower) during the school year or during summer vacation.

The answer to this question was inconsistent. According to the ECLS-K number-right score, gaps grew faster during school than during summer—a finding has inspired at least one article about how early school experiences disadvantage black children (Condron 2009). But this finding used the ECLS-K’s number-right scale, which suffered from measurement artifacts. When we switched to the ECLS-K ability scale, which had fewer artifacts, we found no significant difference between school and summer growth in the reading or math gaps between white and black children. The BSS Thurstone scale also showed no significant difference between school and summer growth in the reading and math gaps between white and black children—though the BSS, too, suffered from measurement artifacts.

If we used the GRD ability scale, we found more consistent evidence of seasonality. At first glance, the seasonal pattern appeared to be that white–black gaps grew faster during the school year. But closer inspection revealed something surprising: The pattern was not that gaps grew faster during school, it was that gaps grew during school *and shrank during summer*. It is hard to make sense of this result. If out-of-school disadvantages caused black children to start kindergarten behind white children, then why would black children catch up when school let out for summer vacation?

Gaps between White and Hispanic Students

Figure 7 shows the mean scores of Hispanic and non-Hispanic white children in reading. Hispanic and non-Hispanic white children were present in both the ECLS-K and the GRD, but not in the BSS.

How much did gaps grow between first grade and eighth grade? Table 3e summarizes growth (or shrinkage) in the white–Hispanic gaps between first grade and eighth grade.

According to the ECLS-K ability scale, the white–Hispanic gap shrank by one-fifth in reading and by two-fifths in math. But according to the GRD ability scale, the gap did not shrink; it grew by one-fifth in both subjects. The disagreement between the ECLS-K and GRD ability scales is a little disquieting because both are IRT ability scales with similar claims to vertical interval scaling. Yet, although the two ability scales did not agree on whether gaps grew or shrank, they did agree that any changes in the gaps were relatively small. Standardization and reliability adjustment reduced the disagreement between the two IRT ability scales. After standardization and reliability adjustment, the ECLS-K and GRD ability scales agreed that math gaps shrank by one-fifth to one-third, whereas reading gaps changed by less than 5 percent.

Scores with measurement artifacts exaggerated gap growth. According to the ECLS-K number-right score, which suffered from artifacts, the white–Hispanic gap grew by half in math and grew sixfold in reading. Standardization and reliability adjustment reduced the appearance of growth substantially. After standardization

Gap between Hispanic and white children: Reading

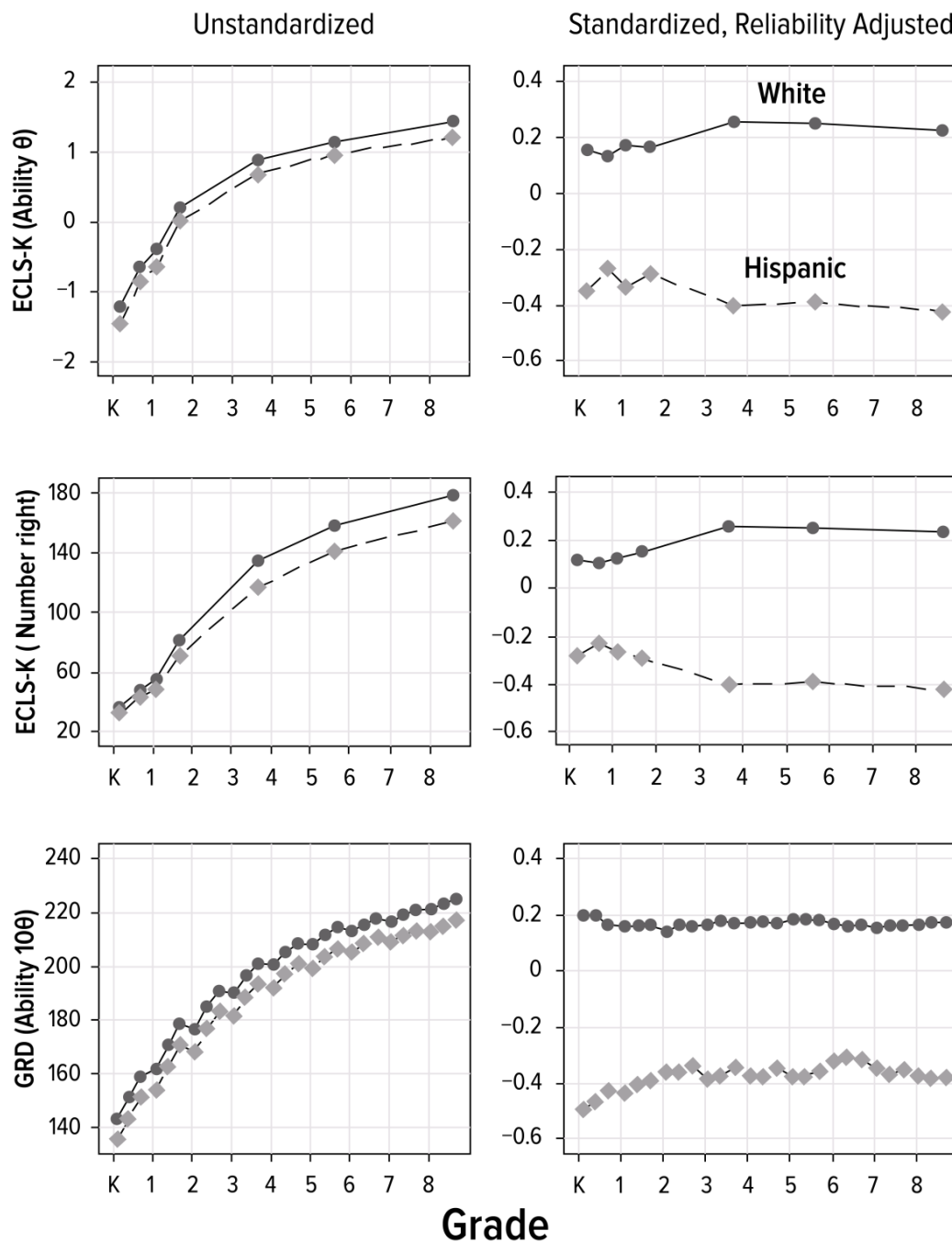


Figure 7: Before standardization, the reading gap between Hispanic children and non-Hispanic white children shrank on the ECLS-K ability scale but grew on the GRD ability scale. After standardization and reliability adjustment, the gap grew on the ECLS-K ability scale but shrank on the GRD ability scale. The ECLS-K number-right scale displayed substantial gap growth, but this was due to measurement artifacts, which standardization and reliability adjustment reduced.

and reliability adjustment, the gap on the ECLS-K number-right score less than doubled in reading and actually shrank in math.

Did gaps grow faster during school or during summer? Table 4e asks whether the white–Hispanic gaps grew faster (or shrank slower) during the school year or during summer vacation. The answer to this question was inconsistent. According to the ECLS-K ability score, gaps grew faster during summer, but according to the GRD ability score, most gaps grew faster during the school year. After standardization and reliability adjustment, the ECLS-K and GRD ability scores agreed that gaps grew faster during summer in math, but they continued to disagree about gap growth in reading.

Closer inspection showed that the GRD math gaps actually *shrank* during summer. It is hard to make sense of this result. If out-of-school disadvantages caused Hispanic children to start kindergarten behind non-Hispanic white children, then why would Hispanic children catch up when school let out for summer vacation?

Scores with measurement artifacts could distort the results. According to the ECLS-K number-right score, which suffered from artifacts, the reading gap grew faster during the school year, and the math gap grew at about the same rate during school and summer. By contrast, the ECLS-K ability score, which had fewer artifacts, suggested that both reading and math gaps grew faster during summer.

Discussion: Measurement Artifacts

Do the test score gaps between advantaged and disadvantaged students grow before, during, or between the school years? Studies have disagreed about the answer to this fundamental question. We have shown that many of these disagreements stem from test measurement artifacts. Even when measuring the same students or schools, changes in scaling could lead researchers to very different and sometimes incorrect conclusions about how much ability gaps grew. An example occurred in the 1980s, when the California Achievement Test switched from Thurstone scaling to IRT ability scaling. Another example occurred in the 2000s, when the ECLS-K, which at first had released number-right scores, began releasing IRT ability scores as well. In general, gap growth looks much smaller on IRT ability scales than it did on Thurstone or number-right scales. On IRT ability scales, most gaps grow rather little, and some gaps even shrink.

The question of whether gaps grow faster during the school year or during summer is also sensitive to measurement artifacts. Scaling artifacts remain a concern, but changes in test form are an even bigger problem. Studies that change test forms at the end of the summer risk confounding summer learning with changes in test form. The potential for change-of-form artifacts is reduced by adaptive testing, which does not use fixed test forms and does not require changing forms at the end of the summer.

Clearly, adaptive tests scored with IRT ability scales are preferable to older alternatives. Yet different adaptive IRT ability scales can still yield different conclusions about gap growth. For example, between first grade and eighth grade, the ECLS-K's IRT ability scale suggested that the white–black reading gap doubled, but the GRD's IRT ability scale suggested that the same gap grew by only one-third. This

difference may stem partly from the fact that the ECLS-K and GRD drew samples from different populations, but measurement could also play a role. The skills tested by the GRD and ECLS-K tests may be somewhat different even if both fall under the broad heading of “reading” or “math.” In addition, the GRD and ECLS-K used somewhat different IRT models. The GRD model had one parameter, whereas the ECLS-K model had three.

Even when results from different IRT ability scales agree, IRT ability scales are only interval measures in the narrow technical sense that the student ability parameter θ_s is a linear function of the log odds that the student will correctly answer an item of a given difficulty. But the log odds are not an intuitive scale, and we can plausibly transform θ_s into a different parameter that is interval scaled in a different technical sense. For example, the transformation $\omega_s = \exp(\theta_s)$ gives a new parameter ω_s , which is interval scaled in the sense that it is a linear function of the simple odds that a student will correctly answer an item of a given difficulty. So even within the IRT framework, the problem of measuring ability on an interval scale remains slippery (Lord and Novick 1968).

Some modern studies sidestep the problem of interval scaling by treating test scores as if they were merely ordinal. On an ordinal scale, the score gap between groups A and B can be summarized by the probability that a randomly chosen member of group A will outscore a randomly chosen member of group B. This probability gap can be converted to a Z gap, which represents the standardized difference that would exist between groups A and B if each group’s scores were normal and had equal variances (Ho 2009). We experimented with this approach, but we found that the resulting Z gaps were very similar to those that we got by simply standardizing the scores directly. This happened because the test scores in our data sets, especially in the GRD and ECLS-K, had something close to a normal distribution with equal within-group variances.

An older and more problematic approach is to convert scores to an age or grade equivalent, so we can speak of one group of students as being, say, one year or grade level ahead of another group. This sounds intuitive, but it has a problem. Because test score growth decelerates with age (this deceleration is evident in Figures 3–7), a gap that is equivalent to one year’s learning in the early grades will be equivalent to more than one year’s learning in the later grades. So on a years-of-learning scale, gaps will appear to grow with age, but this gap growth is an artifact of the fact that the ruler used to measure gaps is shrinking.

A newer idea is to use test scores to predict an adult outcome, such as educational attainment or income (Bond and Lang 2018). Yet it is not clear to us how longitudinal changes in score gaps can be evaluated by using adult outcomes that are measured only once. If the adult outcome method is valid, it is not often applicable because researchers must wait many years for adult outcomes to be collected—if they are collected at all. Only one of our three data sets, the BSS, includes adult outcomes. The ECLS-K participants, though grown, have not been surveyed as adults. The GRD participants are still children, and there are many other children who are taking standardized tests today. We would like to know whether the achievement gaps between these children are growing or shrinking with age, and we would like to have the answer before they are grown.

Conclusion: What We Can Know in Spite of Artifacts

Despite measurement artifacts, our results justify some broad conclusions about when test score gaps grow.

Early Childhood

The first conclusion is that gaps grow fastest in early childhood. Although some results suggested that score gaps grew twofold to sixfold after children start school, those results were invariably obtained using Thurstone or number-right scores, which are not vertical interval measures of student ability. If we focus on results obtained from IRT ability scores, which have a plausible claim to vertical interval scaling, we find that no gap so much as doubled between first grade and eighth grade. Some gaps even shrank. Across all the comparisons considered in this article (black vs. white, poor vs. nonpoor, etc.), the average gap growth between first grade and eighth grade was just 7 percent on the IRT ability scales. All this suggests that reading and math gaps grow substantially more in the first five years of life than they do in the nine years after school begins.

This finding is consistent with neurocognitive research showing that the brain is more plastic at younger ages (Johnson, Riis, and Noble 2016). It is also consistent with economic research suggesting that investments made early in childhood have a greater potential return than do investments later on (Heckman and Masterov 2007). There is some truth in the old Jesuit maxim (Apted 1964), “Give me a child until the age of seven” (or even five), “and I will give you the man” (or woman).

The growing interest in early-childhood programs—such as preschool, home visits, and new-parent training—is justified. It is vital to invest in early-childhood programs, and it is just as vital to understand why some early-childhood programs succeed in shrinking gaps, whereas others fail to realize their potential.

The finding that gaps emerge in early childhood does not mean that nothing can be accomplished later on. Although many school-based programs have had disappointing effects on achievement gaps, a few, such as the Knowledge Is Power Program (KIPP) and the Harlem Children’s Zone middle schools, have managed to cut gaps in half over a period of three school years (e.g., Dobbie and Fryer 2011; Tuttle et al. 2013). Such programs are impressive, but we should recognize that they are remediating gaps that already exist by the time children start school. They are not preventing gaps from opening in the first place.

Summer Learning

Another implication of our results is that it is unclear how much summers contribute to test score gaps. There are well-known findings suggesting that substantial test score gaps accumulate over summer vacation, but those findings were obtained using test scales that spread with age and fixed-form tests that change at the end of the summer. Patterns of summer gap growth do not necessarily replicate when using modern adaptive tests that are scored on IRT ability scales. If summer learning gaps are present, most of them are small and hard to discern through the fog of potential measurement artifacts.

Because there is little net growth in test score gaps after the age of five, it is understandably difficult to slice that growth finely and assign different slices to the school years and to summer vacations. Perhaps it is safest to say that neither schools nor summer vacations contribute much to test score gaps.

This does not mean that summer learning programs have little potential. The potential of summer learning programs is clear from nearly every figure in this article. Although the figures do not consistently show that score gaps grow in summer, they do consistently show that summer learning is slow for nearly all children, including children from advantaged groups. Summer slowdown among advantaged children offers disadvantaged children a window of opportunity to catch up. It is important to make the most of that opportunity, through summer learning programs, through extended school years for disadvantaged children, or through policies that help poor parents and improve the home environments of disadvantaged children.

Notes

- 1 Forms E and F were “alternate forms” meant to produce similar scores with a different set of questions. Form C also had an alternate, Form D, but Form D was not used in the BSS.
- 2 To see this in a simplified setting, suppose that ability has an SD of σ in first grade and the same SD in second grade. Suppose that guessing is impossible and that every first-grade item has a discrimination of A_1 and every second-grade item has a discrimination of A_2 . Then, the true IRF is $\text{logit}^{-1}(A_1(\theta_s - d_i))$ in first grade and $\text{logit}^{-1}(A_2(\theta_s - d_i))$ in second grade. The same responses can be modeled by using a one-parameter logistic IRF $\text{logit}^{-1}(\theta_s - d_i)$, but the resulting estimates will suggest that the SD of ability is $A_1\sigma$ in first grade and $A_2\sigma$ in second. If $A_1 < A_2$, it will appear that the SD of student ability has increased, although the true SD has not. The same thing can happen with a Thurstone IRF, which is just the three-parameter logistic IRF with no guessing and infinite discrimination.
- 3 The ECLS-K codebook uses the term “number right” to indicate the number of questions that a student answered correctly on their first- and second-stage tests. That number-right score is not particularly useful because it pertains only to a small number of questions that are different for different students. To our knowledge, what the ECLS-K calls a number-right score has never been used in a published analysis, so without risk of confusion, we use the term number right to describe what the ECLS-K calls a scale score.
- 4 It is also, to a lesser degree, a function of the discrimination and guessability of those items.
- 5 If the distribution of difficulty across the item pool were uniform, then Figure 2 would be a straight line and the number-right score would be an interval measure of ability. The fact that Figure 2 has an S shape indicates that the cumulative distribution of item difficulty is not uniform but closer to normal (cf. Reardon 2008).
- 6 In fact, the scaling of the ability score θ is arbitrary up to a linear transformation.
- 7 The reliability of the ECLS-K tests also declined in eighth grade.

References

- Alexander, Karl L., and Doris R. Entwisle. 2003. *The Beginning School Study, 1982–2002*. Murray Research Archive. <https://dataverse.harvard.edu/dataset.xhtml?persistentId=hdl:1902.1/01293>.
- Alexander, Karl L., Doris R. Entwisle, and Linda S. Olson. 2001. "Schools, Achievement, and Inequality: A Seasonal Perspective." *Educational Evaluation and Policy Analysis* 23:171–91. <https://doi.org/10.3102/01623737023002171>.
- Alexander, Karl L., Doris R. Entwisle, and Linda S. Olson. 2007a. "Lasting Consequences of the Summer Learning Gap." *American Sociological Review* 72:167–80. <https://doi.org/10.1177/000312240707200202>.
- Alexander, Karl L., Doris R. Entwisle, and Linda S. Olson. 2007b. "Summer Learning and Its Implications: Insights from the Beginning School Study." *New Directions for Youth Development* 2007:11–32. <https://doi.org/10.1002/yn.210>.
- Apted, Michael. 1964. *Seven up!* DVD. Dover, United Kingdom: Granada Television.
- Bond, Timothy N., and Kevin Lang. 2018. "The Black–White Education-Scaled Test-Score Gap in Grades K–7." *Journal of Human Resources* 53:891–917. <https://doi.org/10.3368/jhr.53.4.0916.8242R>.
- Clemans, William V. 1993. "Item Response Theory, Vertical Scaling, and Something's Awry in the State of Test Mark." *Educational Assessment* 1:329–47. https://doi.org/10.1207/s15326977ea0104_3.
- Clemans, William V. 1995. "Reply to Yen, Burket, and Fitzpatrick." *Educational Assessment* 3:191. https://doi.org/10.1207/s15326977ea0302_5.
- Coleman, James S., Ernest Q. Campbell, Carol J. Hobson, Alexander M. Mood, Frederic D. Weinfeld, and Robert L. York. 1966. *Equality of Educational Opportunity*. Washington, DC: Department of Health, Education, and Welfare.
- Condron, Dennis J. 2009. "Social Class, School and Non-School Environments, and Black/White Inequalities in Children's Learning." *American Sociological Review* 74:685–708. <https://doi.org/10.1177/000312240907400501>.
- Cooper, Harris M., Barbara Nye, Kelly Charlton, James Lindsay, and Scott Greathouse. 1996. "The Effects of Summer Vacation on Achievement Test Scores: A Narrative and Meta-analytic Review." *Review of Educational Research* 66:227–68. <https://doi.org/10.3102/00346543066003227>.
- CTB/McGraw-Hill. 1979. *California Achievement Tests, Forms C & D, Technical Bulletin 1*. Monterey, CA: CTB/McGraw-Hill.
- DeMars, Christine. 2010. *Item Response Theory*. Cary, NC: Oxford University Press. <https://doi.org/10.1093/acprof:oso/9780195377033.001.0001>.
- Dobbie, Will, and Roland G. Fryer. 2011. "Are High-Quality Schools Enough to Increase Achievement among the Poor? Evidence from the Harlem Children's Zone." *American Economic Journal: Applied Economics* 3:158–87. <https://doi.org/10.1257/app.3.3.158>.
- Downey, Douglas B., and Dennis J. Condron. 2016. "Fifty Years since the Coleman Report: Rethinking the Relationship between Schools and Inequality." *Sociology of Education* 89:207–20. <https://doi.org/10.1177/0038040716651676>.
- Downey, Douglas B., Paul T. von Hippel, and Beckett A. Broh. 2004. "Are Schools the Great Equalizer? Cognitive Inequality during the Summer Months and the School Year." *American Sociological Review* 69:613. <https://doi.org/10.1177/000312240406900501>.

- Downey, Douglas B., Paul T. von Hippel, and Melanie Hughes. 2008. "Are 'Failing' Schools Really Failing? Using Seasonal Comparisons to Evaluate School Effectiveness." *Sociology of Education* 81:242–70. <https://doi.org/10.1177/003804070808100302>.
- Duncan, Greg J., and Katherine Magnuson. 2011. "The Nature and Impact of Early Achievement Skills, Attention Skills, and Behavior Problems." Pp. 47–69 in *Whither Opportunity?: Rising Inequality, Schools, and Children's Life Chances*, edited by G. J. Duncan and R. J. Murnane. New York, NY: Russell Sage Foundation.
- Entwisle, Doris R., and Karl L. Alexander. 1992. "Summer Setback: Race, Poverty, School Composition, and Mathematics Achievement in the First Two Years of School." *American Sociological Review* 57:72–84. <https://doi.org/10.2307/2096145>.
- Entwisle, Doris R., and Karl L. Alexander. 1994. "Winter Setback: The Racial Composition of Schools and Learning to Read." *American Sociological Review* 59:446–60. <https://doi.org/10.2307/2095943>.
- Entwisle, Doris R., Karl L. Alexander, and Linda S. Olson. 1997. *Children, Schools & Inequality*. Boulder, CO: Westview Press.
- Fryer, Roland G., and Steven D. Levitt. 2006. "The Black–White Test Score Gap Through Third Grade." *American Law Economics Review* 8:249–81. <https://doi.org/10.1093/aler/ahl003>.
- Gershon, Richard C. 2005. "Computer Adaptive Testing." *Journal of Applied Measurement* 6:109–27.
- Greene, William H. 2011. *Econometric Analysis*. 7th ed. Upper Saddle River, NJ: Prentice Hall.
- Gulliksen, Harold. 2013. *Theory of Mental Tests*. Abingdon, United Kingdom: Routledge. <https://doi.org/10.4324/9780203052150>.
- Hanushek, Eric A., and Steven G. Rivkin. 2009. "Harming the Best: How Schools Affect the Black–White Achievement Gap." *Journal of Policy Analysis and Management* 28:366–93. <https://doi.org/10.1002/pam.20437>.
- Hayes, Donald P., and Judith Grether. 1969. "The School Year and Vacations: When Do Students Learn?" Presented at the Eastern Sociological Association Convention, April 19, New York, NY.
- Hayes, Donald P., and Judith Grether. 1983. "The School Year and Vacations: When Do Students Learn?" *Cornell Journal of Social Relations* 17:56–71.
- Heck, Ronald H. 2007. "Examining the Relationship between Teacher Quality As an Organizational Property of Schools and Students' Achievement and Growth Rates." *Educational Administration Quarterly* 43:399–432. <https://doi.org/10.1177/0013161X07306452>.
- Heckman, James J., and Dimitriy V. Masterov. 2007. "The Productivity Argument for Investing in Young Children." *Applied Economic Perspectives and Policy* 29:446–93. <https://doi.org/10.1111/j.1467-9353.2007.00359.x>.
- Heyns, Barbara. 1978. *Summer Learning and the Effects of Schooling*. New York, NY: Academic Press.
- Ho, Andrew D. 2009. "A Nonparametric Framework for Comparing Trends and Gaps Across Tests." *Journal of Educational and Behavioral Statistics* 34:201–28. <https://doi.org/10.3102/1076998609332755>.
- Jencks, Christopher S. 1972. *Inequality: A Reassessment of the Effect of Family and Schooling in America*. New York, NY: Basic Books.
- Jennings, Jennifer L., David Deming, Christopher Jencks, Maya Lopuch, and Beth E. Schueler. 2015. "Do Differences in School Quality Matter More Than We Thought? New Evidence

- on Educational Opportunity in the Twenty-First Century." *Sociology of Education* 88:56–82. <https://doi.org/10.1177/0038040714562006>.
- Johnson, Sara B., Jenna L. Riis, and Kimberly G. Noble. 2016. "State of the Art Review: Poverty and the Developing Brain." *Pediatrics* 137:e20153075. <https://doi.org/10.1542/peds.2015-3075>.
- Koretz, Daniel M. 2009. *Measuring up: What Educational Testing Really Tells Us*. Cambridge, MA: Harvard University Press.
- Lord, Frederic M., and Melvin R. Novick. 1968. *Statistical Theories of Mental Test Scores*. Boston, MA: Addison-Wesley Publishing Company, Inc.
- Murnane, Richard J. 1975. *The Impact of School Resources on the Learning of Inner City Children*. Cambridge, MA: Ballinger Publishing Company.
- Najarian, Michelle, Judith M. Pollack, and Alberto G. Sorongon. 2009. "Early Childhood Longitudinal Study, Kindergarten Class of 1998–99 (ECLS-K), Psychometric Report for the Eighth Grade." <http://nces.ed.gov/pubsearch/pubsinfo.asp?pubid=2009002>.
- National Center for Education Statistics. 2009. "Early Childhood Longitudinal Study, Kindergarten Class of 1998–99 (ECLS-K): Kindergarten Through Eighth Grade Full Sample Public-Use Data and Documentation." Retrieved January 20, 2012 (<http://nces.ed.gov/pubsearch/pubsinfo.asp?pubid=2009005>).
- Northwest Evaluation Association. 2010. *Technical Manual for Measures of Academic Progress and Measures of Academic Progress for Primary Grades*. Lake Oswego, OR: Northwest Evaluation Association.
- Phillips, Meredith, James Crouse, and John Ralph. 1998. "Does the Black–White Test Score Gap Widen after Children Enter School?" Pp. 229–72 in *The Black–White Test Score Gap*, edited by C. S. Jencks and M. Phillips. Washington, DC: Brookings Institution Press.
- Raudenbush, Stephen W., and Anthony S. Bryk. 2001. *Hierarchical Linear Models: Applications and Data Analysis Methods*. 2nd ed. Thousand Oaks, CA: Sage.
- Raudenbush, Stephen W., and Robert D. Eschmann. 2015. "Does Schooling Increase or Reduce Social Inequality?" *Annual Review of Sociology* 41:443–70. <https://doi.org/10.1146/annurev-soc-071913-043406>.
- Raudenbush, Stephen W., and Wing-Shing Chan. 1992. "Growth Curve Analysis in Accelerated Longitudinal Designs." *Journal of Research in Crime and Delinquency* 29:387–411. <https://doi.org/10.1177/0022427892029004001>.
- Reardon, Sean F. 2003. "Sources of Educational Inequality: The Growth of Racial/Ethnic and Socioeconomic Test Score Gaps in Kindergarten and First Grade." Working Paper, Pennsylvania State University Population Research Center, State College, PA.
- Reardon, Sean F. 2008. "Thirteen Ways of Looking at the Black–White Test Score Gap." Working Paper No. 2008–08, Stanford University Institute for Research on Education Policy and Practice.
- Reardon, Sean F. 2011. "The Widening Socioeconomic Status Achievement Gap: New Evidence and Possible Explanations." Pp. 91–116 in *Social Inequality and Educational Disadvantage*, edited by R. Murnane and G. Duncan. Washington, DC: Brookings Institution.
- Rock, Donald A., and Judith M. Pollack. 2002. *Early Childhood Longitudinal Study-Kindergarten Class of 1998–99 (ECLS-K): Psychometric Report for Kindergarten through First Grade*. No. NCES 2010009. Washington, DC: National Center for Education Statistics.
- Singer, Judith D., and John B. Willett. 2002. *Applied Longitudinal Data Analysis: Modeling Change and Event Occurrence*. New York, NY: Oxford University Press.

- Thurstone, Louis Leon. 1925. "A Method of Scaling Psychological and Educational Tests." *Journal of Educational Psychology* 16:433–51. <https://doi.org/10.1037/h0073357>.
- Thurstone, Louis Leon. 1938. "Primary Mental Abilities." *Psychometric Monographs* 1:ix,121.
- Tuttle, Christina Clark, Brian Gill, Philip Gleason, Virginia Knechtel, Ira Nichols-Barrer, and Alexandra Resch. 2013. "KIPP Middle Schools: Impacts on Achievement and Other Outcomes. Final Report." Mathematica Policy Research, Inc. <http://eric.ed.gov/?id=ED540912>.
- von Hippel, Paul T. 2009. "Achievement, Learning, and Seasonal Impact As Measures of School Effectiveness: It's Better to Be Valid than Reliable." *School Effectiveness and School Improvement* 20:187–213. <https://doi.org/10.1080/09243450902883888>.
- von Hippel, Paul T. 2016. "Year-Round School Calendars: Effects on Summer Learning, Achievement, Maternal Employment, and Property Values." Pp. 208–30 in *The Summer Slide: What We Know and Can Do about Summer Learning Loss*, edited by K. L. Alexander, S. Pitcock, and M. Boulay. New York, NY: Teachers College Press.
- Wooldridge, Jeffrey M. 2001. *Econometric Analysis of Cross Section and Panel Data*. 1st ed. Cambridge, MA: MIT Press.
- Yen, Wendy M. 1986. "The Choice of Scale for Educational Measurement: An IRT Perspective." *Journal of Educational Measurement* 23:299–325. <https://doi.org/10.1111/j.1745-3984.1986.tb00252.x>.
- Yen, Wendy M., George R. Burket, and Anne R. Fitzpatrick. 1995a. "Rejoinder to Clemans." *Educational Assessment* 3:203. https://doi.org/10.1207/s15326977ea0302_6.
- Yen, Wendy M., George R. Burket, and Anne R. Fitzpatrick. 1995b. "Response to Clemans." *Educational Assessment* 3:181. https://doi.org/10.1207/s15326977ea0302_4.

Acknowledgments: We thank Mina Kumar for research assistance. We thank the William T. Grant Foundation and the Institute for Urban Policy Research and Analysis for grants supporting this work.

Paul T. von Hippel: LBJ School of Public Affairs, University of Texas at Austin.
E-mail: paulvonhippel.utaustin@gmail.com.

Caitlin Hamrock: E3 Alliance. E-mail: chamrock@e3alliance.org.