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
LBJ School of Public Affairs
University of Texas, Austin

Caitlin Hamrock
E3 Alliance
(formely University of Texas, Austin)

Draft of February 10, 2018

Abstract

A classic question in the social sciences is whether test-score gaps between advantaged and disadvantaged children originate inside or outside schools. One approach to this 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 and fall. These artifacts distorted many past estimates of gap growth.

Net of artifacts, the most replicable finding is that gaps form mainly in early childhood, before schooling begins. Gaps are already large when children enter school, and grow very little afterward; some gaps even shrink. Evidence is inconsistent regarding whether gaps grow faster during the school years or during summer vacations. We substantiate these conclusions in reanalyses of two datasets used in widely cited studies of gap growth—the Beginning School Study of Baltimore school children and the Early Childhood Longitudinal Study, Kindergarten cohort of 1998-99—as well as new analyses from the Growth Research Database of the Northwest Evaluation Association.

Correspondence: Paul T. von Hippel, paulvonhippel.utaustin@gmail.com

Publication history: This is SSRN working paper # 2745527, first created on March 9, 2016, most recently revised on February 10, 2018.

Acknowledgments: We thank the WT Grant Foundation and the University of Texas Institute for Urban Policy Research and Analysis for grants supporting this work.

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

Introduction

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 & Condron, 2016; Downey, von Hippel, & Broh, 2004; Jencks, 1972; Jennings, Deming, Jencks, Lopuch, & Schueler, 2015; Raudenbush & Eschmann, 2015). The question is fundamental to both research and policy. If inequality grows primarily inside schools, then identifying and eliminating the school-based mechanisms that exacerbate inequality should be a major priority. But if inequality grows primarily outside of schools—in family, care, and other non-school settings—then other priorities come to the fore. If inequality comes primarily from outside schools, then research and policy should focus more on identifying and reducing non-school sources of inequality, and on evaluating ways that schools and other institutions can compensate for non-school inequality.

One way to assess the relative importance of school and non-school sources of inequality is with a multi-year longitudinal study that tests children repeatedly, starting near the beginning of their school career. The idea is that test score gaps that are observed when children enter school can only be due to non-school influences, while later growth in the gaps is due to a mix of school and non-school factors (e.g., Phillips, Crouse, & Ralph, 1998).

The relative influence of school and non-school effects can be separated further if a longitudinal study uses a *seasonal* design that tests children twice a 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 non-school influences, while school-year learning is due to a mix of school and non-school factors. 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 children start attending school, the major sources of inequality lie elsewhere (e.g., Alexander, Entwisle, & Olson, 2001; Downey et al., 2004; Hayes & 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 & Magnuson, 2011; Heckman & 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 & Grether, 1969, 1983). And many studies attribute gaps and gap growth to schools (e.g., Condron, 2009; Hanushek & Rivkin, 2009; Heck, 2007).

Results have differed for different types of gap 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 & Alexander, 1992), while the gap between black and white children has been reported to grow fastest during school (Condron, 2009; Downey et al., 2004; Entwisle & 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 seasonal studies suffered from measurement artifacts that were much larger than any true growth in the skill gaps between children.

If we discard studies with serious measurement artifacts—or if we accept those studies and attempt 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, but in most analyses 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 up before kindergarten begins. Results that suggest otherwise 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 non-poor families, and the gaps between high-poverty and low-poverty schools.

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

Artifact 1: Changes of test form

Perhaps the most famous seasonal learning study is the Beginning School Study (BSS) of students who started first grade in Baltimore City Schools in 1982. The BSS found that the gap between higher- and lower-SES children tripled between first and ninth grade, with all of the gap growth occurring during summer vacations and no gap growth at all during the school years (Alexander, Entwisle, & Olson, 2007a).

This finding may be an artifact of measurement. More precisely, it could result from two artifacts.

The first potential artifact comes from *changes in test form*. The BSS used Form C of the California Achievement Test (CAT), which was a fixed-form paper test. In each grade, all children filled in a paper form with the same set of questions in fall and spring, then switched to a harder form when they started a new grade the next fall. For example, children used the CAT’s level 11 form in the fall and spring of first grade, and then switched to the level 12 form in the fall of second grade.

What this implies 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. So 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 be that, even with no change in abilities, the new form would produce more dispersed scores than the old. There is no way to be sure. The CAT's publisher (CTB/McGraw-Hill, 1979) tried to align forms so that scores could be compared across grades, but there is no guarantee that alignment was successful, and there is reason to be concerned that it was not, as we will discuss in the next section. In addition, the mere fact of switching to a new test often causes children's scores to drop (Koretz, 2009), and the drop may be more pronounced for some types of children than it is for others.

Before we discuss the next artifact, we should say that the BSS was not the only seasonal learning study to change forms over the summer. Essentially all seasonal studies did so from the 1960s into the 1990s. This was not the study authors' fault; it was simply what was done at the time. Yet it implies even summer learning pattern which a 1996 meta-analysis suggested were well-replicated (Cooper, Nye, Charlton, Lindsay, & Greathouse, 1996) may all suffer from change-of-form artifacts.

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 suggest that they have greater ability. So adaptive tests do not need to change abruptly at the start of a new school year, and summer learning can be estimated the same way as school year learning, with less concern about artifacts.

Artifact 2: Test score scaling

The second potential artifact is test score *scaling*. Scaling is the mathematical method by which a pattern of right and wrong answers gets translated into a test score. If a score is used to estimate gap growth, its scaling should ensure that it is an *interval* measure of students' ability which is *vertically aligned* so that students can be compared at different ages. 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 when and whether score gaps appear to grow. This has been known for a long time, but its implications for studies of school and summer gap growth have not been fully appreciated.

In fact, the very test that was used in the BSS—the CAT form C—was the focus of a controversy about test score scaling. The controversy erupted in 1985 (year 4 of the BSS), when the CAT switched from Forms C and its alternate Form D to Forms E and F.[1] The new forms radically changed researchers' impressions about whether achievement gaps grew with age. On the old Forms (C and D), the standard deviation (SD) of reading and math scores at least doubled between 1st and 8th grade, and continued to grow through 12th grade. But on the new Forms (E and F), the SD did not grow at all; instead, it shrank between 1st and 8th grade, then held steady through 12th grade (Yen, 1986).

A possible reason for this contrast was that the old CAT forms used Thurstone scaling (Thurstone, 1925, 1938), while the new forms used item response theory (IRT) scaling (Lord &

Novick, 1968). While Thurstone and IRT scaling sometimes yield similar results, in the CAT they clearly did not.

Observers who were used to the old CAT scores expressed dismay at the new scores. They opined that "something's awry in the state of test mark," and argued that the old Thurstone scales were better than the new IRT scales (Clemans, 1993). But defenders of the new CAT countered that IRT scales had better psychometric properties and could detect growth in the dispersion of students' skills if growth were actually present (Clemans, 1995; Yen, Burket, & Fitzpatrick, 1995a, 1995b).

Substantive arguments could not resolve the issue. Some observers offered "common-sense" arguments that gaps must grow with age. Surely the gaps between high school students in honors or remedial classes must be larger than the gaps between 1st graders taught in the same classroom. Yet defenders of IRT scores countered with sensible arguments that gaps might well be greater in the early grades (Yen, 1986 for a review). Could there be any gap greater than the gap between a 1st grader who can read and one who does not know the alphabet?

While no truce was ever called in the CAT dispute, in retrospect the defenders of Thurstone scaling were fighting a rear guard action. Since the CAT adopted IRT scaling in 1985, IRT has become by far the most popular method for scaling achievement tests, while Thurstone scaling has fallen into disuse. There are good reasons for the abandonment of Thurstone scaling; its assumptions are more restrictive and less realistic than the assumptions of IRT scaling, and this can lead to artifacts. We will define the differences between Thurstone and IRT scaling in the Data section.

The history of scaling in the CAT has implication for the BSS. The BSS started in 1982 with the Thurstone-scaled CAT Form C, and continued to use that form even after the IRT-scaled Forms E and F were released in 1985. So the BSS finding that SES gaps tripled between 1st and 6th grade may be an artifact of CAT Form C which could not have been replicated using Forms E and F. Likewise the BSS finding that "a large portion of the [8th grade] achievement gap originates over the summer" (Alexander, Entwisle, & Olson, 2007b) would be harder to support using an IRT test that showed little gap growth after children begin school. Again this is not the fault of anyone involved with the BSS; the study simply occurred during a transitional period in the history of scaling.

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 1998-99 (ECLS-K). These studies reported that the SES gap in reading and math grew fastest during summer vacation, but the black-white gap grew faster during the school year and displayed substantial growth from kindergarten through third grade (Condron, 2009; Downey et al., 2004; Fryer & 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 the ECLS-K was released with new ability scores that had a stronger claim to vertical interval scaling (Najarian, Pollack, & Sorongon, 2009). When the new scores were used, the timing and extent of gap growth was different than it was on the old scores, and results that had been obtained using the old score did not necessarily replicate with the new scores (Koretz, 2009; Quinn, 2015; Reardon, 2008). Later we will show that results for summer gap growth are also sensitive to the way that the ECLS-K tests are scaled.

Preview

In the rest of this article, we estimate the change in score gaps from the beginning of school through eighth grade, and we evaluate whether those changes occur mainly during the school years or during summer vacations. To do this, we draw evidence from three different data sets: the BSS, the ECLS-K, and an extract from the Growth Research Database (GRD) maintained by the Northwest Evaluation Association (NWEA). These datasets take different approaches to measurement; they also include children from different cohorts and different parts of the country.

We evaluate whether past conclusions about changes in score gaps have been affected by measurement artifacts. In the BSS, we assess artifacts by comparing results that use the study's original scores with results that attempt to statistically reduce potential artifacts in those scores. In the ECLS-K, we assess artifacts by comparing results obtained using the non-interval scores that were first released with the data to results using better scores that were released later. In the GRD, we are less worried about artifacts, but we still compare results on the published scale to results that have been statistically standardized and adjusted for reliability.

By comparing results across different datasets and measurements, we not only highlight which results are sensitive to artifacts. We also highlight which results are replicable. The results will show that many conclusions—about which gaps grow or shrink as children get older, and whether gap growth and shrinkage occurs during the school years or during summer vacations—are not particularly replicable across different datasets and scales. Once measurement artifacts are accounted for, the most replicable result is that substantial gaps are present when children begin school, and any changes in later school years and summers are relatively small.

Data

We compare three longitudinal datasets: the BSS, the ECLS-K, and the GRD. The datasets draw from different populations of children and give different tests on different testing schedules.

Tests

Features of the tests used in the three datasets are summarized in Table 1. The BSS used parts of Form C of the California Achievement Test (CAT). The GRD used the NWEA's 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 & Pollack, 2002). The tests differ in several ways.

Fixed form vs. adaptive testing

One difference between the tests is that the BSS used fixed-form testing, while the ECLS-K and GRD used adaptive testing. On a fixed-form test, all students in the same grade get a form with a fixed set of items. The form changes when students progress to a new grade. As discussed in the introduction, this raises the possibility of artifacts since school-year learning is estimated by comparing scores from the same test form, while summer learning is estimated by comparing spring and fall scores from different test forms.

The GRD and ECLS-K are less vulnerable to change-of-form artifacts because they used adaptive testing. In adaptive testing, different students get different items drawn from a large item pool. The difficulty of the items presented to a particular student is calibrated according to an estimate of the student's ability. The ECLS-K uses two-stage adaptive testing, in which a first-stage routing test provides an initial estimate of ability, and that estimate is used to assign students to a second-stage test of appropriate difficulty (Najarian et al., 2009). The GRD uses continuous adaptive testing, which re-estimates student ability after each response, and uses the current ability estimate to choose the difficulty of the next question (Northwest Evaluation Association, 2010). Either way, adaptive testing ensures that students get more difficult questions as they grow older and more capable, with no need for abrupt changes of test form each fall.

Test scaling

The tests also differ with respect to scaling. The BSS used CAT Form C which, along with its alternate Form D, was the last CAT form to use Thurstone scaling (CTB/McGraw-Hill, 1979). The BSS continued to use Form C even after Forms E and F, which used IRT scaling, were published in 1985 (Entwisle, Alexander, & Olson, 1997; Yen, 1986). The ECLS-K and GRD both used IRT scaling, but not exactly the same type. Specifically, the GRD used a 1-parameter logistic (1PL) IRT model, also known as a Rasch model, while the ECLS-K used a 3-parameter (3PL) 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 attempts 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, 1PL, and 3PL 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 from 0 up to 1 as soon as a student's ability exceeds the difficulty of the item (Thurstone, 1925):

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

Under the 1PL IRF, the probability of a correct response increases more gradually, following an inverse logit function which rises as a student's ability approaches and then exceeds the difficulty of an item (DeMars, 2010):

$$P(correct) = logit^{-1}(a_i(\theta_s - d_i)) \qquad \text{(1PL IRF)}$$

Under a 3PL IRF, items differ not just in difficulty but in *discrimination* d_i and *guessability* c_i (DeMars, 2010):

$$P(correct) = c_i + (1 - c_i)logit^{-1}(a_i(\theta_s - d_i)) \qquad \text{(3PL IRF)}$$

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 4 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).

Comparing the three models, we see that the 1PL IRF is like the 3PL IRF if there were no guessing (c_i=0) and all items discriminated equally (a_i=1). The Thurstone IRF is like the 3PL IRF if there were no guessing and all items discriminated perfectly ($a_i \rightarrow \infty$). When a 1PL model is used, test developers try to choose items that come close to satisfying the assumption of equal discrimination, but when a Thurstone model is used there is often no way to satisfy the assumption of perfect discrimination. That assumption is unrealistic.

Violations of an IRF's assumptions can bias the estimated distribution of student ability. For example, suppose that items given on a second-grade test have higher discrimination than the items given on a first-grade test. The 3PL IRF can model this, but the Thurstone and 1PL IRFs cannot since they assume all items discriminate equally. The Thurstone and 1PL IRFs will effectively mistake the increase in item discrimination for an increase in the dispersion of student ability. For example, if item discrimination doubles between first and second grade, a Thurstone or 1PL model will estimate that the SD of student ability has doubled instead.[2]

This is one possible explanation for the history of the CAT assessments discussed in the Introduction. When the CAT used Thurstone scaling, it appeared that the SD of ability was increasing with age, but when the CAT switched to a 3PL IRT model, it appeared that the SD of ability was constant or decreasing (Yen, 1986). The reason for the change in the SD was never made clear, but one possibility is that CAT items increased in discrimination as children grew older, and the Thurstone scale was mistaking the increase in discrimination for an increase in the dispersion of ability.

Another difference between Thurstone and IRT models is that the Thurstone model assumes that ability is normally distributed (Thurstone, 1925; Yen et al., 1995b). As students get older, the Thurstone model assumes that the shape of the ability distribution remains normal and only the mean and SD change. IRT models do not assume that ability is normally distributed, and in some settings it may not be. It is not hard to imagine settings where ability is positively skewed, with many students of low to moderate ability, and a few "gifted" students with extraordinarily high ability. Even if students start school with a normal ability distribution, the distribution may not remain normal as they get older. That said, this does not appear to be a major issue in the GRD and ECLS-K, where according to the IRT estimates the ability distribution does appear to be approximately normal.

In general, the modern IRT scales used by the GRD and ECLS-K are preferable to the old-fashioned Thurstone scale used by the BSS. The IRT assumptions are more flexible and realistic.

Ability scores vs. number-right scores

Recent releases of the ECLS-K provide 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 & Pollack, 2002)). The number-right score estimated how many correct answers the child would have given had they been given every item in the full item pool, instead of just the sample of questions drawn from the item pool for the first- and second-stage tests that students actually took.

Unfortunately, number-right scores are 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 children's ability grows, their number-right score grows slowly at first because the item pool has few easy questions. Then, as students reach the moderate ability levels where the bulk of questions are pitched, the number-right score reaches an inflection point and grows more quickly with student ability. Because it 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 will grow on the number-right scale 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 grows over this period, but on the ability score it appears that the gap shrinks.

The timing of apparent gap growth can also be an artifact of the S-shaped relationship. Advantaged students reach the inflection point in the S first, and when they reach it they pull away from disadvantaged students on the number-right score—but not necessarily on the ability score. So the timing of gap growth on the number-right score depends on when advantaged students reach that inflection point. If they reach it 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 after 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 these artifacts may affect different comparisons differently. The moment when advantaged students turn the corner 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 black-white gap opens during school, but the gap between the children of college and non-college-educated mothers opens during summer. Yet all these differences may be illusory, artifacts of the number-right score. 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 (Condron, 2009; Downey et al., 2004; Downey, von Hippel, & Hughes, 2008; Fryer & Levitt, 2006; Reardon, 2003; von Hippel, 2009). Its shortcomings were not yet appreciated, and in any case 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 ability scores are used instead of number-right scores.

The GRD MAP tests report scores using "Rasch units" (RIT) which are a linear transformation of ability: RIT=10θ+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 does not change any substantive conclusions about the growth or shrinkage of ability and ability gap. 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: it looks like other scores that parents and teachers are used to. In particular, the RIT scale takes positive three-digit values and increases by 5 to 20 points per year—whereas raw ability scores θ can take negative values and typically increase by less than a point per year.

Table 2 presents the means and SDs of the reading and math scores used in the BSS, the ECLS-K, and the GRD. Some scales spread with age, while others do not. Specifically, the SDs of the BSS Thurstone scores and the ECLS-K number-right scores more than double between 1st and 8th grade, while the SDs of the GRD ability scores grow by less than half, and the SDs of the ECLS-

K ability scores actually shrink. What this means is that, as they grow older, children's scores will spread relatively little on the GRD and ECLS-K ability scales, but will spread substantially on the ECLS-K number-right and BSS Thurstone scale.

Table 3 estimates the scores' *reliability*, classically defined as the proportion of variance that is due to true ability rather than measurement error (Lord & Novick, 1968). Most of the reliability estimates come from technical documentation published by the test developers (CTB/McGraw-Hill, 1979; Najarian et al., 2009; Northwest Evaluation Association, 2010).

The reliability of the BSS reading scale increases markedly with age, while the reliability of the other scales increases little if at all. This is important because an increase in score reliability will cause an artifactual increase in standardized score gaps. That is, if the reliability of a test increases with age, standardized score gaps will grow faster than true ability gaps. We will return to this issue in the Methods section when we discuss reliability adjustment.

For each scale, Table 3 presents two estimates of reliability. The first is an internal consistency estimate,[7] obtained by comparing a child's answers to different questions during the same administration of the same test form. The second is a test-retest estimate, obtained by correlating the scores on alternate forms of the same test taken a few days or weeks apart. Test-retest estimates are typically lower than internal consistency estimates of reliability because test-reliability estimates account for day-to-day as well as item-to-item variations in a child's performance (Shavelson & Webb, 1991). For some tests the difference between test-retest and internal-consistency estimates is large; for other tests is it is small. For example, in the CAT test used by the BSS, the fall 1st grade reading scores are .68 reliable according to the internal consistency estimate but only .50 reliable according to the test-retest estimate; however, the fall 1st grade math scores are .80-.81 reliable according to either estimate.

Strictly speaking, only the CAT offers a proper estimate of test-retest reliability, obtained by correlating scores from alternate test forms[8] (Forms C and D) taken 2-3 weeks apart (CTB/McGraw-Hill, 1979). For the GRD and ECLS-K, we do not have a proper estimate of test-reliability, but in the ECLS-K we do have correlations between fall and spring tests taken approximately 6 months apart, and in the GRD we have correlations between fall, winter, and spring tests taken approximately 3-4 months apart. These correlations (displayed in Table 3) are typically about .8-.9, and this figure may be treated as a lower bound on reliability. It is a lower bound because six months is long enough for the true distribution of ability to change; that is, even if the tests were perfectly reliable, the correlation between fall and spring scores would be less than 1.0.

Test content

The tests from the BSS, ECLS-K, and GRD also differ in content. The content tested in the BSS was limited. Although the CAT Form C had several sections, not all of them were used in the BSS. The BSS reading scores were limited to the Reading Comprehension section of the CAT and did not include other CAT reading sections: namely Reading Vocabulary and (in grades 1-3) Phonic Analysis and Structural Analysis. Likewise the BSS math scores were limited to the CAT Mathematics Concepts and Applications section and did not include the CAT section on Mathematics Computation (compare CTB/McGraw-Hill, 1979; to Entwisle 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 included ranged from "basic skills" and "vocabulary" to "reading comprehension skills" such as "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 is aligned with state curriculum standards, and many state standards are now aligned with the common core. MAP items are screened for validity and bias through a multi-stage 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 datasets differ in the number of school years and summers that they cover. The testing schedules can be visualized by noticing the patterns of filled and empty cells in Table 2 and Table 3, where a filled cell indicates an occasion when a test was administered and an empty cell indicates an occasion when it was not.

The BSS tested students twice a year, in the fall and spring of every school year from 1st grade through 6th grade, and then once a year, in the springs of 7th and 8th grade. This design lets us estimate learning rates for every school year and summer from 1st grade to 6th grade. After 6th grade, we can still estimate annual learning rates, but we cannot separate summer learning from school year learning.

The ECLS-K tested students in the fall and spring of kindergarten and 1st grade, and then every 2-3 years, in the springs of 3rd, 5th, and 8th grade. (The fall 1st grade test was limited to a random 30% subsample of schools.) This design lets us separate school-year from summer learning rates during kindergarten, 1st grade, and the summer between. After 1st grade, we cannot separate summer learning from school year learning, but we can still estimate average learning rates over periods of 2 or 3 years.

Our GRD extract followed an accelerated longitudinal design (Raudenbush & Chan, 1992). During two consecutive school years (2008-09 and 2009-10), we followed eight cohorts of student: we followed one from kindergarten through 1st grade, one from 1st through 2nd grade, and so on through the oldest cohort which we followed from 7th through 8th grade. The advantage of an accelerated design is that, with just two years of data, we can estimate learning over every school year and summer from kindergarten through 8th grade, with less attrition than if we followed a single cohort for nine years.

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

Ideally tests would be given on the first and last day of the school year, but none of the studies did that. Our models will compensate for the fact that exams were not taken on the first and last day. The GRD and ECLS-K record the exact date on which each student took each test. Most commonly the fall test was taken in October, and the spring test was taken in May. The BSS also gave tests in October and May (Entwisle et al., 1997), but the exact dates do not appear in the data. In analyses where we need exact test dates for the BSS, we plug 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 need to know the dates on which each school year started and ended. The ECLS-K makes these dates available as restricted data. The GRD does not include school dates, but we found them by looking up the calendars of participating schools on the internet. The BSS does not include school dates, either, but staff at the Baltimore City School District told us that, in 1982-90 when the BSS was running, Baltimore's public schools opened on the day after Labor Day. We assumed that that Baltimore schools closed

284 days later, since 284 is the average number of calendar days between the start and end of the school year in the ECLS-K. Making other plausible assumptions about the end of the BSS school year did not materially affect our results.

Children and schools

The three datasets differ in the characteristics of the children and schools that were tested. Table 4 compares the school, child, and family characteristics of the BSS, ECLS-K, GRD, highlighting some of the strengths and weaknesses of each dataset.

The BSS is the oldest dataset. It began in the fall of 1982 with 790 first graders sampled from 20 public schools in the Baltimore City School District (Alexander & 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 Schools, which means that they were, on average, somewhat poorer than the US population. The BSS is limited to black and white children, since there were very few Hispanic or Asian families living in Baltimore in 1982.

The BSS is historically important and it provides detailed information on schools and families, but its sample is 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 less than three-fifths had a mother who had graduated high school.

The ECLS-K is much more diverse. It began with a US 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 which oversampled private schools and areas with high enrollments of Asian-Americans. (Sampling weights compensate for the oversampling.) Our analytic ECLS-K sample excludes 27 year-round schools where children attend school much of the summer (von Hippel, 2016). After this exclusion, our analytic sample totals 17,779 children in 977 schools.

Our GRD extract is very large, containing 177,549 students in 389 schools, 25 school districts, and 14 US states (AK, CO, IN, KS, KY, MN, NM, NV, OH, SC, TX, WA, WI, WY). Although the sample is not nationally representative since it is limited to districts that use NWEA's MAP tests, those districts are quite diverse, including large urban districts as well as suburban, small-town, and rural districts. Across GRD districts, the percentage of students who are white ranges from 16% to 98%, the percentage of students who receive subsidized meals ranges from 12% to 73%, and the average math percentile score ranges from 22 to 70.

At the school level, we can exploit the diversity of the GRD, because the GRD is supplemented with ample information on school characteristics and demographics from the Common Core of Data. At the child level, unfortunately, the GRD is quite limited. It contains no child or family variables except for race/ethnicity and gender. We requested data on meal subsidy and maternal education, but NWEA staff told us that few districts collect information on maternal education and, although most districts collect meal subsidy status, that variable is not integrated into the GRD.

Methods

We use all three datasets to estimate the size and growth or shrinkage of various test score gaps through school years and summers from the start of kindergarten (1st grade in the BSS) until the end of 8th grade. Our overarching questions are these:

1. How large are gaps at the start of kindergarten (1st grade in the BSS)?
2. How much do gaps grow or shrink by the end of 8th grade?
3. To the degree that gaps grow, is the growth faster during the summers or during the school years?

We evaluate whether the answer to these question are consistent across different types of gap, different datasets, different populations, different tests and different test scales.

Gap types

We estimate gaps between advantaged and disadvantaged children, where advantage and disadvantage are defined in different ways. Because each dataset has somewhat different measures of advantage, not every dataset has every gap.

Past studies of the BSS and ECLS-K have measured SES using custom indices that combine measures of parental education, income, employment, and occupation (Alexander et al., 2007b; Downey et al., 2004). We cannot use these indices here because the measures that they require are survey-specific and cannot be replicated across different surveys. The ECLS-K uses an SES index which we cannot replicate in the BSS, and the BSS uses an SES index which we cannot replicate in the ECLS-K.

Instead using composite SES indices, we use three simple SES components: child poverty, school poverty, and maternal education.

- We estimate the *gap between non-poor and poor children*, where poor children are defined as children eligible for meal subsidy (free/reduced lunch). Note that this gap is not available in the GRD, which lacks data on individual students' meal subsidy status.
- We estimate the *gap between low-poverty and high-poverty schools*. Our original intention was to use Title I status, but that is not available in the BSS, so instead we define high-poverty schools as schools with at least 40 percent of students qualifying for meal subsidy. Other schools are defined as low-poverty.
- We estimate the *gap between the children of more and less educated mothers*. We define three levels of maternal education at the time of the first tests: mothers who had not completed high school, mothers who had completed at least a high school diploma or equivalent, and mothers who had completed at least a bachelor's degree. Maternal education is unavailable in the GRD. It is available in both the ECLS-K and the BSS, but in the BSS only mothers reported having a bachelor's degree, so we group them with other high school graduates.

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

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

- We estimate the *gap between white and black students*. This is available in all three datasets.
- We estimate the *gap between white and Hispanic students*. This is available in the ECLS-K and the GRD, but not in the BSS where there are no Hispanics.

We code each gap so that it is positive. For example, instead of the black-white gap we examine the white-black gap, defined as the average number of points by which white students lead black students. Likewise, we define the non-poor-poor gap as the average number of points by which students who do not receive meal subsidies lead those who do. When gaps are defined in this way, growth in a gap is always a positive number, and shrinkage in the gap is always negative. This simplifies interpretation.

Standardization and reliability adjustment (SRA)

We can get a lot of insight by simply graphing the average scores of advantaged and disadvantaged students against average test dates from the fall of kindergarten or 1st grade through the spring of 8th grade. On these graphs we can eyeball the size of achievement gaps and identify periods when gaps are growing or shrinking rapidly.

A limitation of these graphs is that they are vulnerable to measurement artifacts because they take scores at face value. Another limitation is that it is hard to compare results across different datasets, because each dataset scales test scores differently.

While it is important to see patterns on each test's native scale, it is also desirable to have a common, less scale-dependent metric on which the results from different datasets can be compared and certain artifacts can be reduced.

A first attempt to construct such a scale is to standardize each score with respect to the mean and SD of scores for the same test and occasion (e.g., for the MAP scores in the fall of 3rd grade). The result is a *standard score Z* which on each occasion has mean 0 and SD 1 regardless of the original scale of the test. Standard scores are familiar, and their use reduces but does not eliminate the differences between scales.

But standard scores are vulnerable to changes in *reliability*. Reliability is the fraction of variance in test scores that is due to true ability rather than measurement error. The BSS tests are less reliable than the ECLS-K and NWEA tests, and all three tests are less reliable in the early grades than they are later on (Table 3).[9] 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 the kindergarten test is 0% reliable, so that all of the variance in test scores is due to measurement error and none is due to true ability. Then students' scores will be completely random, and there will be no average score gaps between the standard kindergarten score of high- and low ability students. As students progress into higher grades and scores become more reliable, gaps in standard scores will grow, even if gaps in true ability do not change at all.

To correct for differences in reliability, we use a standardized and reliability-adjusted (SRA) score $Z_{SRA} = Z/\sqrt{\hat{\rho}}$, where $\hat{\rho}$ is an estimate of reliability (Ho, 2009; Reardon, 2011). We adjust for reliability using the internal-consistency reliability estimates from Table 3; the test-retest estimates led to very similar results.

Reliability adjustment spreads out scores from less reliable tests, so that two groups that differ by one SD in average true ability will differ by one unit in average Z_{SRA}, regardless of reliability. Reliability adjustment noticeably increases estimated gaps if reliability is low, but not if reliability is high. For example, if a test is 68% reliable—like the first reading test in the BSS—then a reliability-adjusted gap using Z_{SRA} is 21% larger than an unadjusted gap using *Z*. But if a test is 90% reliable—like the ECLS-K and GRD tests—then a gap measured with Z_{SRA} is only 5% larger than a gap measured with *Z*.

Reliability adjustment can also affect estimates of gap *growth* if the reliability of a test changes with age. For example, between 1st and 2nd grade the test-retest reliability of the BSS reading score increases from 68% to 89%. Under these circumstances, a gap that appears to grow by 14% on the *Z* scale will not grow at all on the Z_{SRA} scale.

Despite their advantages, Z_{SRA} scores do not necessarily offer an interval scale for estimating growth in score gaps. By construction, Z_{SRA} scores produce a scale where the SD of ability does not change with age. But 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 captured by Z_{SRA}.

Growth Modeling

We fit a multilevel growth model to estimate rates of test score growth during each school year and summer (Raudenbush & Bryk, 2001; Singer & Willett, 2002). We fit this model to test scores on their native scale, and also to Z_{SRA} scores.

We describe in detail the model that we fit to the GRD, which is the only dataset that covered every school year and summer. Similar models were fit to the BSS and ECLS-K, but some modification was needed because those datasets did not cover every school year and summer.[10]

Our basic growth model was

$$Y_{sct} = \beta_0 + \alpha_0 Kinder_{sct} + \beta_1 Summer1_{sct} + \alpha_1 First_{sct} + \beta_2 Summer2_{sct} + \cdots + u_s + e_{sct}$$

where Y_{sct}, the dependent variable, is the reading or math score for child *c* in school *s* on measurement occasion *t*. All models were fit to the unstandardized original scores, as well as to Z_{SRA} scores that had been standardized and adjusted for reliability.

The exposure variables $Kinder_{sct}$, $Summer1_{sct}$, etc. are the number of months that the child has been exposed to each school year and summer (kindergarten, summer 1, etc.) as of measurement occasion *t*. These exposure variables are coded to reflect the fact that tests were not given on the first and last day of each school year; for example, at the time of the first test, the kindergarten exposure variable is approximately $Kinder_{sct}$=1.5 months, instead of 0.

The coefficients $\alpha_1, \alpha_2, \ldots$ are average monthly learning rates during each school year (kindergarten, 1st grade, etc.), and the coefficients $\beta_1, \beta_2, \ldots$ are average monthly learning rates rates during each summer (summer 1, summer 2, etc.). β_0 is the average score at the start of kindergarten. A school random effect u_s accounts for the correlations among children from the same school, and the residual e_{sct} has an autocorrelated structure that accounts for the correlation among tests given to the same child. One version of the residual model also accounted for heteroscedasticity, though this made little difference to the estimates.[11] The model for the GRD data includes a fixed effect for each cohort, but these cohort fixed effects were small and made little difference to other estimates.

To estimate gaps in scores and score growth between advantaged and disadvantaged children, we added to the model a dummy variable X_{sc} which was 1 for advantaged children or schools and 0 for disadvantaged children or schools. In different model runs, X_{sc} was defined to indicate different measures of advantage—e.g., being white (vs. black), being non-poor (vs. poor), being in a low-poverty (vs. high-poverty) school.

$$\begin{aligned} Y_{sct} = \beta_0 + \alpha_0 Kinder_{sct} + \beta_1 Summer1_{sct} + \alpha_1 First_{sct} + \beta_2 Summer2_{sct} + \cdots + \gamma_0 X_{sc} \\ + \gamma_1 Kinder_{sct} X_{sc} + \delta_1 Summer1_{sct} X_{sc} + \gamma_2 First_{sct} X_{sc} + \delta_2 Summer2_{sct} X_{sc} \\ + \cdots + u_s + e_{sct} \end{aligned}$$

The coefficient γ_0 of X_{sc} represents the average score gap between advantaged and disadvantaged children at the start of kindergarten. We let X_{sc} interact with variables representing exposure to kindergarten, summer 1, etc. so that the coefficients $\gamma_1, \gamma_2, \ldots$ and $\delta_1, \delta_2, \ldots$ of the interactions represent the average growth (or shrinkage) of the gap during each school year and summer.

Linear combinations of coefficients

Our research questions are addressed not by the coefficients themselves, but by linear combinations of the coefficients.

We use linear combinations to estimate total growth in the gaps across all school years and summers from the beginning of 1st grade to the end of 8th—a period covered by all 3 datasets. The school years average 9.37 months, and the summers average 2.63 months, so the total growth in the score gap is $\gamma_{1-8} = 9.37\sum_{i=1}^{8}\gamma_i$ over the 1st-8th grade school years and $\delta_{2-8} = 2.63\sum_{i=2}^{8}\delta_i$ over the summers between. Then the total growth in the gap from the start of 1st grade to the end of 8th grade is $\gamma_{1-8} + \delta_{2-8}$. The gap at the start of 1st grade is $\gamma_{start1} = \gamma_0 + 9.37\gamma_1 + 2.63\delta_2$, and the gap at the end of 8th grade is $\gamma_{start1} + \gamma_{1-8} + \delta_{2-8}$.

We also use linear combinations to compare the average monthly rate of gap growth during the school years and during summer vacations. In the GRD, the average monthly growth in the gap is $\bar{\gamma} = 1/9\sum_{i=1}^{9}\gamma_i$ across the 9 school years and $\bar{\delta} = 1/8\sum_{i=1}^{8}\delta_i$ across the 8 summers. So the average monthly difference between summer gap growth and school year gap growth is the contrast $\bar{\delta} - \bar{\gamma}$. Similar contrasts can be defined for shorter periods such as grades K-1 in the ECLS-K or grades 1-6 in the BSS.

Missing Data

All three datasets had missing test scores and covariates which we handled using multiple imputation. We imputed each missing value 20 times. We used a fully conditional specification in which each variable was imputed by regressing it on all the others. We used linear regression to impute continuous variables, and binomial, ordinal, and multinomial logit regression to impute categorical variables (Raghunathan, Lepkowski, Van Hoewyk, & Solenberger, 2001).

To account for correlations among tests given to the same student, before imputing we reshaped each dataset into wide format, so that there was one row per student and one column for each test occasion. This forced the imputation model to estimate correlations between different tests given to the same student (Allison, 2002). To account for correlations among students in the same school, we first imputed a school-level dataset containing school variables and school-level averages of child variables; then we joined the imputed school-level dataset to the child-level dataset and imputed the child-level variables.

Imputed test scores were used in graphs of mean test scores. For our multilevel growth models, we used imputed covariates but deleted imputed test scores. This approach is known as *multiple imputation with deletion*; it produces slightly more efficient estimates of growth parameters and, more importantly, it reduces the sensitivity of the results to debatable choices made in the imputation model (von Hippel, 2007).

Alternative methods

In addition to the methods described above, we explored several other ways to estimate test score gaps across datasets. One way is to treat scores as ordinal rather than interval measures of ability. On an ordinal scale, the score gap between groups A and B can be summarized by the probability *P*=*P*(A>B) that a randomly chosen member of group A will outscore a randomly chosen member of group B. The *P* gap can be converted to a *Z* gap which represents the mean standardized difference that would exist between the scores of groups A and B if each group's scores were normal and equally variable (Ho, 2009). We experimented with this approach and found that the resulting Z gaps were very similar to those that we got by simply standardizing the scores directly. This occurred because the test scores in our datasets, especially in the GRD and ECLS-K, are quite close to normal and equally variable.

An older suggestion is to convert scores to an age or grade equivalent, so that we can speak of one group of students as being, say, one year ahead of another group. This approach is intuitive, but it has a problem. Since test score growth usually slows with age (see Figure 3-Figure 12), then a gap which 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. On a years-of-learning scale, gaps will appear to grow with age, simply because 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 & Lang, 2013). As a practical matter, we could not use this approach in our study since only one of our three datasets, the BSS, includes data on adult outcomes. The GRD participants are not adults yet, while the ECLS-K participants have reached young adulthood but have not been surveyed as adults. These are common limitations of test score data. Even when adult outcomes are available, it is not clear to us how longitudinal changes in score gaps can be compared using adult outcomes that are measured only once.

Results

Figure 3-Figure 12 plot mean reading and math score gaps in the BSS, ECLS-K, and GRD from the fall of kindergarten or 1st grade through the spring of 8th grade. The Figures present several gaps—the gap between non-poor and poor children, the gap between low-poverty and high-poverty schools, the gaps between the children of more- and less-educated mothers, and the gaps between white, black, and Hispanic children. The Figures present these gaps on four different scales—the BSS Thurstone scale, the ECLS-K number-right scale, the ECLS-K ability scale θ, and the GRD's 10θ scale. Gaps are presented before and after standardization and reliability adjustment (SRA).

Visual impressions about the timing and extent of gap growth change with the dataset and scale that are used. On the IRT ability or θ scales provided by the ECLS-K and GRD, gaps display little visible growth between the start of kindergarten and the end of 8th grade. Gaps grow substantially on the BSS Thurstone scale and the ECLS-K number right scale, but that growth is an artifact of flawed scaling. The discrepancies between scales are substantially reduced, but not eliminated, when scales have been standardized and adjusted for reliability.

In the rest of the Results, we make these impressions more concrete by focusing on two questions:

1. How much do reading and math gaps grow between 1st and 8th grade?
2. Do gaps grow faster during the summers or during the school year?

We demonstrate the sensitivity of the answers to the dataset and scale that are used, and we ask whether, despite that sensitivity, the questions can be broadly answered.

How much do gaps grow between 1st and 8th grade?

Table 5 and Table 6 summarize the growth of different reading and math gaps between the beginning of 1st grade and the end of 8th grade—a range of grades covered by all three datasets.

Gaps grow very little on the most trustworthy scales. On the ability scales provided by the GRD and ECLS-K, gaps shrink about as often as they grow. On average , across all gap types (white-black, poor-non-poor, etc.), the ability scales show just 4 percent growth in math gaps and 3 percent growth in reading gaps. By contrast, on the flawed scales—the BSS Thurstone scale and the ECLS-K number-right scale—most math gaps at least double (increase by over 100 percent) and most reading gaps at least triple (increase by over 200 percent).

This contrast is disquieting, but its origins are clear. In Table 2 we saw that the SD of test scores grew with age on the BSS Thurstone scale and the ECLS-K number-right scale, but not on the ability scales provided by the GRD and ECLS-K. If the SD of a scale grows, then the gaps between advantaged and disadvantaged students will grow as well. However, the spread of the Thurstone and number-right scales is a scaling artifact which does not necessarily indicate a spread in students' ability.

Standardization and reliability adjustment reduce but do not eliminate the discrepancies among scales. After standardization and reliability adjustment, the average gap on the ECLS-K and GRD ability scales, averaged across all gap types, grows by just 17 percent in reading and actually shrinks by 4 percent in math. On less trustworthy scales—the BSS Thurstone scale or the ECLS-K number-right scale—the average gap grows by 37 percent in reading and by 7 percent in math.

The exact amount of gap growth depends on the type of gap. Patterns of gap growth are somewhat different for the white-black gap than for the white-Hispanic gap, and somewhat different for the gap between nonpoor and poor children than for the gap between low-and high-poverty schools. Some patterns are different for reading than for math. Yet the differences between subjects and gap types are dwarfed by the differences between an IRT ability score, derived from an adaptive test, and a flawed score with serious measurement artifacts.

Socioeconomic gaps

Poor vs. non-poor children

Table 5a and Table 6a describe the gap between non-poor children and poor children. According to the ECLS-K ability scale, both reading and math gaps shrink by a quarter between 1st and 8th grade. But on the flawed ECLS-K number-right scale, the math gap nearly doubles and the reading gap more than triples. On the flawed Thurstone scale used in the BSS, apparent gap growth is even larger; the gap grows almost fivefold in math and more than eightfold in reading. So the true ability gap between non-poor and poor children shrinks after school begins, but it can appear to grow if scores are badly scaled.

Standardization and reliability adjustment reduce but do not eliminate these discrepancies. After SRA, the reading gap grows by just 6 percent on the ECLS-K ability scale but grows by about half on the flawed ECLS-K number-right scale and the flawed Thurstone scale used in the BSS.

Note that we cannot compare non-poor and poor children in the GRD, which lacks data on individual children's meal subsidy status.

Low- vs. high-poverty schools

Table 5b and Table 6b describe growth in the reading gap between low-poverty and high-poverty schools. Both the ECLS-K and the GRD measure this gap using ability scales, but they do not agree. According to the ECLS-K, both the reading and math ability gaps shrink by about 15 percent, but according to the GRD the reading gap grows by a third and the math gap doubles. Since the ECLS-K and GRD scale ability similarly, we are not sure what to make of these discrepancies. We do note that the ECLS-K is nationally representative, while the GRD, though diverse, is not.

The flawed Thurstone scale used in the BSS gives much larger estimates of growth in the gap between low-poverty and high-poverty schools. It suggests that reading and math gaps more than triple, but this result is exaggerated by scaling artifacts. In the ECLS-K, the flawed number-right scale suggest, as usual, more gap growth than the ability scale. In particular, the number-right scale suggests that the gap grows by half in math and more than doubles in reading, but the ability scale shows that both gaps shrink by about 15 percent.

Again, standardization and reliability adjustment reduce but do not eliminate these discrepancies.

More- vs. less educated mothers

Table 5c and Table 6c summarize growth in the gaps between children whose mothers did or did not complete a high school diploma or equivalent. According to the ECLS-K ability scale, both the reading and math gaps shrink by a third between 1st and 8th grade. But on the flawed ECLS-K number-right scale, the math gap doubles and the reading gap quadruples. On the flawed Thurstone scale used in the BSS, apparent gap growth is even larger; the gap grows more than fourfold in math and almost sixfold in reading. So the true ability gap between children whose mothers did and did not finish high school shrinks after school begins, but can appear to grow if scores are badly scaled. Standardization and reliability adjustment reduce but do not eliminate these discrepancies.

Table 5d and Table 6d summarize growth in the gaps between children whose mothers did or did not complete a bachelors degree. The results are similar to those for a high school diploma. According to the ECLS-K ability scale, both the reading and math gaps shrink by a quarter between 1st and 8th grade. But on the flawed ECLS-K number-right scale, the math gap almost doubles and the reading gap more than triples. So again, the true ability gap between children whose mothers did and did not finish a bachelors degree shrinks after school begins, but can appear to grow if scores are badly scaled. Standardization and reliability adjustment eliminates the disparity between scales in math but not in reading.

Note that we cannot estimate maternal education gaps in the GRD, which lacks data on maternal education. And we cannot estimate the gap between mothers with and without a bachelors degree in the BSS, where only three mothers had a bachelors degree.

Racial and ethnic gaps

White and black students

Table 5e and Table 6e summarize growth in the gaps between white and black children. Both the ECLS-K and the GRD measure this gap using ability scales, but they do not agree. According to the ECLS-K, the reading gap grows by a fifth and the math gap grows by an amount that is statistically indistinguishable from zero. Yet according to the GRD the math gap grows by half and the reading gap grows by three-quarters. Since the ECLS-K and GRD scale ability similarly, we are not sure what to make of their disagreement. We do note that the ECLS-K is nationally representative, while the GRD, though diverse, is not.

The flawed ECLS-K number-right scale gives much larger estimates for growth in the white-black gap, suggesting that it more than doubles in math and sextuples in reading. Clearly this is a scaling artifact. The true growth in the ability gap in small to negligible, but the badly scaled number-right score makes it look enormous.

The flawed Thurstone scale used in the BSS also gives very large estimates for growth in the white-gap gap, suggesting that it nearly triples in math. In reading, the BSS suggests that black children actually start 1st grade slightly, though not significantly, ahead of white children. We do not know what to make of this initial lead for black children. As far as we know, it is unique in the literature on the white-black test score gap.

Standardization and reliability adjustment reduce but do not eliminate the differences between estimates of gap growth obtained with different scores.

White and Hispanic students

Table 5f and Table 6f summarize growth in the gaps between white and Hispanic children. Both the ECLS-K and the GRD measure this gap using ability scales, but they do not agree. According to the GRD, both gaps grow by a fifth. But according to the ECLS-K, both gaps shrink—the reading gap shrinks by a fifth and the math gap shrinks by two-fifths. Since the ECLS-K and GRD scale ability similarly, we are not sure what to make of this discrepancy. We do note that the ECLS-K is nationally representative, while the GRD, though diverse, is not.

The flawed ECLS-K number-right scale gives much larger estimates for growth in the white-Hispanic gap, suggesting that grows by half in math and sextuples in reading. Clearly this is a scaling artifact. The true growth in the ability gap in small to negligible, but the badly scaled number-right score makes it look enormous.

Standardization and reliability adjustment reduce but do not eliminate the discrepancies between different scales. Note that we cannot estimate the white-Hispanic gap in the BSS, which has no Hispanics.

Do gaps grow faster during school or during summer?

We now ask whether test score gaps grow fastest during school or during summer. Given the previous section's findings that most ability gaps grow little (or even shrink) between 1st and 8th grade, the question of whether they grow faster during school or summer seems less important and may be harder to answer. We ask the question anyway. It is traditional, and it may shed light on how school and non-school factors affect inequality after the age of 5.

Because different datasets take seasonal measurements at different times (see Table 2), we compare school and summer growth over three different grade ranges: grades K-1 (which are available in the ECLS-K and GRD), grades 1-6 (available in the BSS and GRD), and grades K-8 (available in the GRD only).

Socioeconomic gaps

Perhaps the most famous result in the summer learning literature is that the gaps between socioeconomically advantaged and disadvantaged students grow faster during summer than they do during the school year. We now show that this result is sensitive to the dataset that is used, the scale on which test scores are measured, and the variable that is used to define socioeconomic status.

Poor vs. non-poor children

Table 7a and Table 8a summarize seasonal patterns in the reading and math gaps between non-poor children and poor children. The gap grows faster during summer than during school. This is true in both math and reading, in both the ECLS-K and the BSS, and both before and after standardization and reliability adjustment.

The ECLS-K's flawed number-right score distorts this pattern of summer gap grwoth. The number-right score suggests that the reading gap grows faster during school and the math gap grows at about the same rate during school and summer. Standardization and reliability adjustment do not affect conclusions from the ability score, but change the number-right patterns. The inconsistency of results is confusing, but we need not worry about them since we know the number-right score is flawed.

Note that we cannot contrast non-poor and poor children in the GRD, which lacks data on individual children's meal subsidy status.

Low- vs high-poverty schools

The previous section suggested that the gap between non-poor and poor children grows faster during summer than during school. But that evidence came from the ECLS-K, which has only one summer, and the BSS, whose test scores are flawed.

Table 7b and Table 8b summarize seasonal patterns in the reading and math gaps between low- and high-poverty schools. Again the ECLS-K ability scores and the BSS Thurstone scores suggest that both reading and math gaps grow faster during summer. But now we also have evidence from the GRD—and the GRD does not agree. In grades K-1, the GRD suggests that math gaps grow faster in summer, but reading gaps grow faster during school. In grades 1-6, the pattern reverses and the GRD suggests that reading gaps grow faster during school, but math gaps grow faster during summer. Over all grades, from K through 8, the GRD finds that reading gaps grow at about the same rate during school and summer, while math gaps grow faster during school and actually shrink during summer.

The results are highly inconsistent across datasets, scores, and subjects. Standardization and reliability adjustment do not resolve the discrepancies.

More- vs. less-educated mothers

Table 7c and Table 8c summarize seasonal patterns in the reading and math gaps between children whose mothers did or did not graduate high school. In the ECLS-K, the ability scores show that gaps grow faster during summer, but the flawed number-right scores show no significant difference between school and summer gap growth. In the BSS, the flawed Thurstone scores also show no significant difference between school and summer gap growth.

Table 7d and Table 8d summarize seasonal patterns in the reading and math gaps between children whose mothers did or did not complete a bachelors degree. Again, in the ECLS-K the ability scores show that gaps grow faster during summer, but the flawed number-right scores show no significant difference between school and summer gap growth.

Note that we cannot estimate maternal education gaps in the GRD, which lacks data on maternal education. And we cannot estimate the gap between mothers with and without a bachelors degree in the BSS, where only three mothers had a bachelors degree.

<u>Racial and ethnic gaps</u>

The white-black gap

Almost as famous as the finding that SES gaps grow fastest in the summer is the finding that the gaps between white and black students grow fastest during the school year. We now show that this finding, too, is sensitive to measurement.

Table 7e and Table 8e summarize seasonal patterns of growth in the reading and math gaps between white and black and children. The ECLS-K ability score shows no significant difference between rates of gap growth during school and during summer. The ECLS-K number-right score suggests that the gap grows faster during school—but that score is flawed. The flawed BSS Thurstone score also shows no significant difference between school and summer gap growth.

These null findings are striking, because both the BSS and the ECLS-K have been used to support the claim that the white-black gap grows faster during school (Condron, 2009; Downey et al., 2004; Entwisle & Alexander, 1994).

The GRD generally suggests that gaps grow faster during school. This is true in both subjects, in both grades K-8 and grades 1-6. In reading, it is also true in grades K-1.

After standardization and reliability adjustment, nearly every white-black gap grows fastest during school. But closer inspection reveals a surprising pattern. The pattern is not so much that gaps grow faster during school; it is that gaps grow during school *and shrink during summer*. This is very surprising. If out-of-school disadvantages cause black children to start kindergarten behind white children, then why would black children catch up when school lets out for summer vacation? We have no answer to this question.

The white-Hispanic gap

Table 7f and Table 8f summarize seasonal patterns of growth in the reading and math gaps between white and Hispanic children. The ECLS-K ability score suggests that the gaps grows faster during summer, but the GRD ability score suggests that it grows faster during school. Again, closer inspection shows that GRD math gaps actually shrink during summer—a pattern that is hard to explain since Hispanic children presumably suffer from out-of-school disadvantages.

After standardization and reliability adjustment, every math gap grows faster during summer, and so does the ECLS-K ability gap in reading.

Conclusion

Measurement artifacts

Do the test score gaps between advantaged and disadvantaged students grow before, during, or between the school years? Different studies have given discrepant answers to this fundamental question. In part, these discrepancies may result from the fact that different studies test different samples of students in different times and places (Table 4). But we have shown that the discrepancies result in large part from test measurement artifacts. Even when the students are held constant, changes in scaling can lead to very different conclusions about how much score gaps grow. And studies that change test forms at the end of the summer risk confounding test-form changes with summer learning.

The potential for change-of-form artifacts is reduced by adaptive testing, and the potential for scaling artifacts is reduced by using IRT ability scores. So when estimating gap growth, we should favor IRT ability scores that are calculated from adaptive tests.

Do such scales offer vertical interval measures of student ability? They certainly have a stronger claim to interval scaling than do number-right scales or Thurstone scales. But different IRT ability scales can still yield very different conclusions about the growth of gaps. For example, the IRT ability scale used in the ECLS-K suggests that the white-black reading gap doubles between 1st and 8th grade—but the GRD's IRT ability scale suggests that the same gap grows by only a third. This difference may occur partly because the ECLS-K and GRD refer to different populations, but measurement artifacts can also play a role. Test scores are affected by test content, and 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 use somewhat different IRT models; the GRD's is a 1PL model, while the ECLS-K is a 3PL model.

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 is a linear function of the *log odds* that the student will correctly answer an item of given difficulty, discrimination, and guessability. If we don't like the log odds interpretation, then we can plausibly transform θ_s into a different parameter that is interval-scaled in a different technical sense. For example, the transformation $\omega = exp(\theta)$ 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 correct answer an item of given difficulty etc. So even within the IRT framework, the problem of measuring ability on an interval scale remains slippery (Lord & Novick, 1968). Some modern studies treat test scores as ordinal (Ho, 2009).

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 suggest that score gaps double or triple after children start school, these results are invariably obtained using Thurstone or number-right scores which are not interval measures of student ability. If we focus on IRT ability scores, then despite limitations and disagreements we find that no gap so much as doubles between kindergarten and 8th grade, and some gaps even shrink. All this suggests that reading and math gaps grow substantially more in the first five years of life, before school begins, than they do in the nine years after.

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

The growing policy and research 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 just as vital to understand what distinguishes effective programs from programs that 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 have managed to cut gaps in half over a period of 3 school years (e.g., Dobbie & Fryer, 2009; 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 is that it is not clear how much summers really contribute to test score gaps. There are some well-known findings suggesting that substantial test score gaps accumulate over summer vacations, but these findings were obtained using test scales that spread with age and/or fixed-form tests that change at the end of the summer. Similar findings may be hard to replicate using modern IRT ability scales and adaptive tests. It is somewhat intuitive to think that children are more unequal during summer than during the school year, and while the balance of evidence perhaps tips in that direction, several of our findings point in the opposite direction, and in general the size of summer learning gaps is difficult to discern through the fog of potential measurement artifacts. In light of the finding that there is little net growth in test score gaps after the age of five, it is understandably difficult to slice that growth finely enough to assign different components to school or summer. Perhaps it is safest to say that neither schools nor summers contribute a great deal 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 paper (Figure 3-Figure 12). Even when these figures do not suggest that score gaps grow in summer, they invariably show that summer learning is slow for practically all children, including children from advantaged groups. This means that the summer offers disadvantaged children a window of opportunity to catch up. It is important to make the most of that opportunity, either through summer learning programs or through extended school years for disadvantaged children.

Seasonal research design

A final implication of our findings is that seasonal research designs may not be able to authoritatively answer the question of whether inequality originates primarily at school or at home. The question of whether test score gaps grow faster during summer or during the school year is important, but that question turns out to be difficult to answer because of measurement artifacts. The question of where inequality comes from remains important, but it is not a question that can be answered by one research design alone. While seasonal longitudinal studies can contribute to the answer, other sources of evidence are needed as well.

Endnotes

[1] Forms C and D were alternate forms which contained different items but were meant to have interchangeable scores. The BSS only used Form C. Forms E and F were also alternate forms.

[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 $logit(A_1(\theta_s-d_i))$ in first grade and $logit(A_2(\theta_s-d_i))$ in second grade. The same responses can be modeled using a 1PL IRF $logit(\theta_s-d_i)$, 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 3PL 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 since it pertains only to a small number of questions which 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] There are several ways to estimate reliability from internal consistency. Documentation for the CAT Form C uses the KR-20 formula (CTB/McGraw-Hill, 1979; Kuder & Richardson, 1937), while documentation from the tests used in the ECLS-K and GRD uses IRT formulas (Najarian, Pollack, & Sorongon, 2009; Northwest Evaluation Association, 2010).

[8] Alternate test forms were only used in grades 3-8; in grades 1-2 there were no alternate forms.

[9] The reliability of the ECLS-K tests also declines in 8th grade.

[10] In the GRD, tests were given in the fall and spring of every grade, so we could estimate the model as described with a separate growth rate for every school year and summer. In the ECLS-K, however, we could only estimate school and summer growth rates through the end of first grade, after which we could only estimate average growth rates over periods of 2-3 years. Similarly, in the BSS, we could only estimate school and summer growth rates from 1st through 5th grade; after 5th grade, we could only estimate 12-month growth rates, and before 1st grade we couldn't estimate anything because no tests were given.

[11] We modeled the serial correlation in different ways, first using a heteroscedastic AR1 model, and then using a spatial power model in which the residual correlation decreased as the time between tests increased (Littell, Milliken, Stroup, Wolfinger, & Schabenberger, 2006). The results were materially unchanged. An alternative way to account for correlations among tests on the same child is to include a child random intercept and possibly random slopes as well. But child random effects greatly increase runtime and memory use, to the point that SAS would not run models with child random effects on the GRD.

References

Alexander, K. L., & Entwisle, D. R. (2003). *The Beginning School Study, 1982-2002* (No. hdl:1902.1/01293 UNF:3:hlT7u4CbArgxawKtYr87fg==). Cambridge, MA: Murray Research Archive. Retrieved from http://dvn.iq.harvard.edu/dvn/dv/mra/faces/study/StudyPage.xhtml?studyId=467

Alexander, K. L., Entwisle, D. R., & Olson, L. S. (2001). Schools, Achievement, and Inequality: A Seasonal Perspective. *Educational Evaluation and Policy Analysis*, *23*(2), 171–191.

Alexander, K. L., Entwisle, D. R., & Olson, L. S. (2007a). Lasting Consequences of the Summer Learning Gap. *American Sociological Review*, *72*(2), 167–180. https://doi.org/10.1177/000312240707200202

Alexander, K. L., Entwisle, D. R., & Olson, L. S. (2007b). Summer learning and its implications: Insights from the Beginning School Study. *New Directions for Youth Development*, *2007*(114), 11–32. https://doi.org/10.1002/yd.210

Allison, P. D. (2002). *Missing Data*. Thousand Oaks, Calif.: Sage Publications.

Apted, M. (1964). *Seven Up!* Granada Television.

Bond, T. N., & Lang, K. (2013). *The Black-White Education-Scaled Test-Score Gap in Grades K-7* (Working Paper No. 19243). National Bureau of Economic Research. Retrieved from http://www.nber.org/papers/w19243

Clemans, W. V. (1993). Item Response Theory, Vertical Scaling, and Something's Awry in the State of Test Mark. *Educational Assessment*, *1*, 329–347. https://doi.org/10.1207/s15326977ea0104_3

Clemans, W. V. (1995). Reply to Yen, Burket, and Fitzpatrick. *Educational Assessment*, *3*(2), 191.

Coleman, J. S., Campbell, E. Q., Hobson, C. F., Mood, A. M., Weinfeld, F. D., & York, R. L. (1966). *Equality of Educational Opportunity*. Washington, DC: Department of Health, Education and Welfare.

Condron, D. J. (2009). Social Class, School and Non-School Environments, and Black/White Inequalities in Children's Learning. *American Sociological Review*, *74*(5), 685–708. https://doi.org/10.1177/000312240907400501

Cooper, H. M., Nye, B., Charlton, K., Lindsay, J., & Greathouse, S. (1996). The Effects of Summer Vacation on Achievement Test Scores: A Narrative and Meta-Analytic Review. *Review of Educational Research*, *66*(3), 227–268. 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, C. (2010). *Item Response Theory*. Oxford University Press, USA.

Dobbie, W., & Fryer, R. G. (2009). *Are High Quality Schools Enough to Close the Achievement Gap? Evidence from a Social Experiment in Harlem* (NBER Working Papers No. 15473). Cambridge, MA: National Bureau of Economic Research.

Downey, D. B., & Condron, D. J. (2016). Fifty Years since the Coleman Report: Rethinking the Relationship between Schools and Inequality. *Sociology of Education*, *89*(3), 207–220. https://doi.org/10.1177/0038040716651676

Downey, D. B., von Hippel, P. T., & Broh, B. A. (2004). Are Schools the Great Equalizer? Cognitive Inequality during the Summer Months and the School Year. *American Sociological Review*, *69*(5), 613.

Downey, D. B., von Hippel, P. T., & Hughes, M. (2008). Are “Failing” School Really Failing? Using Seasonal Comparisons to Evaluate School Effectiveness. *Sociology of Education*, *81*(3), 242–270.

Duncan, G. J., & Magnuson, K. (2011). The nature and impact of early achievement skills, attention skills, and behavior problems. In G. J. Duncan & R. J. Murnane, *Whither Opportunity?: Rising Inequality, Schools, and Children’s Life Chances* (pp. 47–69). New York City: Russell Sage Foundation.

Entwisle, D. R., & Alexander, K. L. (1992). Summer Setback: Race, Poverty, School Composition, and Mathematics Achievement in the First Two Years of School. *American Sociological Review*, *57*(1), 72–84.

Entwisle, D. R., & Alexander, K. L. (1994). Winter Setback: The Racial Composition of Schools and Learning to Read. *American Sociological Review*, *59*(3), 446–460.

Entwisle, D. R., Alexander, K. L., & Olson, L. S. (1997). *Children, Schools & Inequality*. Boulder, Colorado: Westview Press.

Fryer, R. G., & Levitt, S. D. (2006). The Black-White Test Score Gap Through Third Grade. *Am Law Econ Rev*, *8*(2), 249–281. https://doi.org/10.1093/aler/ahl003

Gershon, R. C. (2005). Computer Adaptive Testing. *Journal of Applied Measurement*, *6*(1), 109–127.

Hanushek, E. A., & Rivkin, S. G. (2009). Harming the best: How schools affect the black-white achievement gap. *Journal of Policy Analysis and Management*, *28*(3), 366–393. https://doi.org/10.1002/pam.20437

Hayes, D. P., & Grether, J. (1969). The School Year and Vacations: When Do Students Learn? In *Eastern Sociological Association Convention*. New York City.

Hayes, D. P., & Grether, J. (1983). The School Year and Vacations: When Do Students Learn? *Cornell Journal of Social Relations*, *17*, 56–71.

Heck, R. H. (2007). Examining the Relationship Between Teacher Quality as an Organizational Property of Schools and Students’ Achievement and Growth Rates. *Educational Administration Quarterly*, *43*(4), 399–432. https://doi.org/10.1177/0013161X07306452

Heckman, J. J., & Masterov, D. V. (2007). The Productivity Argument for Investing in Young Children. *Applied Economic Perspectives and Policy*, *29*(3), 446–493. https://doi.org/10.1111/j.1467-9353.2007.00359.x

Heyns, B. (1978). *Summer Learning and the Effects of Schooling*. New York: Academic Press.

Ho, A. D. (2009). A Nonparametric Framework for Comparing Trends and Gaps Across Tests. *Journal of Educational and Behavioral Statistics*, *34*(2), 201–228. https://doi.org/10.3102/1076998609332755

Jencks, C. S. (1972). *Inequality: A reassessment of the effect of family and schooling in America*. New York: Basic Books.

Jennings, J. L., Deming, D., Jencks, C., Lopuch, M., & Schueler, B. E. (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*(1), 56–82. https://doi.org/10.1177/0038040714562006

Johnson, S. B., Riis, J. L., & Noble, K. G. (2016). State of the Art Review: Poverty and the Developing Brain. *Pediatrics*, peds.2015-3075. https://doi.org/10.1542/peds.2015-3075

Koretz, D. M. (2009). *Measuring Up: What Educational Testing Really Tells Us*. Cambridge, Mass.: Harvard University Press.

Kuder, G. F., & Richardson, M. W. (1937). The theory of the estimation of test reliability. *Psychometrika*, *2*(3), 151–160. https://doi.org/10.1007/BF02288391

Littell, R. C., Milliken, G. A., Stroup, W. W., Wolfinger, R. D., & Schabenberger, O. (2006). *SAS for Mixed Models* (2nd ed.). SAS Institute.

Lord, F. M., & Novick, M. R. (1968). *Statistical Theories of Mental Test Scores*. Addison-Wesley Publishing Company, Inc.

Murnane, R. J. (1975). *The impact of school resources on the learning of inner city children*. Cambridge, Mass.: Ballinger Pub. Co.

Najarian, M., Pollack, J. M., & Sorongon, A. G. (2009). *Early Childhood Longitudinal Study, Kindergarten Class of 1998 - 99 (ECLS-K), Psychometric Report for the Eighth Grade*. Retrieved from http://nces.ed.gov/pubsearch/pubsinfo.asp?pubid=2009002

National Center for Education Statistics. (2009, July 17). 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, from 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, M., Crouse, J., & Ralph, J. (1998). Does the Black-White Test Score Gap Widen after Children Enter School? In C. S. Jencks & M. Phillips, *The black-white test score gap*. Brookings Institution Press.

Quinn, D. M. (2015). Black–White Summer Learning Gaps Interpreting the Variability of Estimates Across Representations. *Educational Evaluation and Policy Analysis*, *37*(1), 50–69. https://doi.org/10.3102/0162373714534522

Raghunathan, T. E., Lepkowski, J. M., Van Hoewyk, J., & Solenberger, P. W. (2001). A Multivariate Technique for Multiply Imputing Missing Values Using a Sequence of Regression Models. *Survey Methodology*, *27*(1), 85–95.

Raudenbush, S. W., & Bryk, A. S. (2001). *Hierarchical Linear Models: Applications and Data Analysis Methods* (2nd ed.). Sage Publications, Inc.

Raudenbush, S. W., & Chan, W.-S. (1992). Growth Curve Analysis in Accelerated Longitudinal Designs. *Journal of Research in Crime and Delinquency*, *29*(4), 387–411. https://doi.org/10.1177/0022427892029004001

Raudenbush, S. W., & Eschmann, R. D. (2015). Does Schooling Increase or Reduce Social Inequality? *Annual Review of Sociology*, *41*(1), 443–470. https://doi.org/10.1146/annurev-soc-071913-043406

Reardon, S. F. (2003). *Sources of Educational Inequality: The Growth of Racial/Ethnic and Socioeconomic Test Score Gaps in Kindergarten and First Grade* (Working Paper). State College, PA: Pennsylvania State University, Population Research Center.

Reardon, S. F. (2008). *Thirteen ways of looking at the black-white test score gap* (Working paper No. 2008-08). Stanford University: Insitute for Research on Education Policy and Practice.

Reardon, S. F. (2011). The widening socioeconomic status achievement gap: new evidence and possible explanations. In R. Murnane & G. Duncan, *Social Inequality and Educational Disadvantage*. Washington, DC: Brookings Institution.

Rock, D. A., & Pollack, J. M. (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.

Shavelson, R. J., & Webb, N. M. (1991). *Generalizability Theory: A Primer: A Primer*. SAGE Publications.

Singer, J. D., & Willett, J. B. (2002). *Applied longitudinal data analysis: modeling change and event occurrence*. Oxford; New York: Oxford University Press.

Thurstone, L. L. (1925). A method of scaling psychological and educational tests. *Journal of Educational Psychology*, *16*(7), 433–451. https://doi.org/10.1037/h0073357

Thurstone, L. L. (1938). Primary mental abilities. *Psychometric Monographs*, *1*, ix + 121.

Tuttle, C. C., Gill, B., Gleason, P., Knechtel, V., Nichols-Barrer, I., & Resch, A. (2013). *KIPP Middle Schools: Impacts on Achievement and Other Outcomes. Final Report.* Mathematica Policy Research, Inc. Retrieved from http://eric.ed.gov/?id=ED540912

von Hippel, P. T. (2007). Regression With Missing Ys: An Improved Strategy For Analyzing Multiply Imputed Data. *Sociological Methodology*, *37*, 83–117.

von Hippel, P. 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*(2), 187–213.

von Hippel, P. T. (2016). Year-round school calendars: Effects on summer learning, achievement, maternal employment, and property values. In K. L. Alexander, S. Pitcock, & M. Boulay, *The Summer Slide: What We Know and Can Do About Summer Learning Loss*. New York: Teachers College Press.

Yen, W. M. (1986). The Choice of Scale for Educational Measurement: An IRT Perspective. *Journal of Educational Measurement*, *23*(4), 299–325.

Yen, W. M., Burket, G. R., & Fitzpatrick, A. R. (1995a). Rejoinder to Clemans. *Educational Assessment*, *3*(2), 203.

Yen, W. M., Burket, G. R., & Fitzpatrick, A. R. (1995b). Response to Clemans. *Educational Assessment*, *3*(2), 181.

Tables

Table 1. Test characteristics

	BSS	ECLS-K	GRD
Test	CAT, Form C	Custom	MAP
Domains	Reading comprehension	Reading	Reading
	Math concepts	Math	Math
Scaling	Thurstone	IRT (3PL logistic)	IRT (1PL logistic)
Scales	Number-right score	Ability score θ	RIT (Rasch unit) = $10\theta + 200$
		Number-right (number-right) score	
Medium	Paper	Paper in 8th grade, computer before	Computer
Adaptive?	No: fixed form	2-stage adaptive	Continuously adaptive

Table 2. Means and SDs of test scales

		CAT Thurstone (BSS)				ECLS-K ability θ				ECLS-K number-right				GRD ability 10θ			
		Reading		Math		Reading		Math		Reading		Math		Reading		Math	
Grade	Season	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD
K	Fall					-1.31	.52	-1.16	.47	35	10	26	9	141	12	142	13
	Winter													149	13	149	14
	Spring					-.74	.51	-.68	.46	46	14	36	12	157	13	158	15
1	Fall	279	41	292	32	-.52	.51	-.45	.46	52	18	43	14	159	14	161	15
	Winter													168	15	171	15
	Spring	339	46	340	37	.10	.46	.04	.42	77	24	61	18	176	15	179	15
2	Fall	340	53	338	38									174	17	177	14
	Winter													182	16	185	14
	Spring	384	52	379	41									188	16	192	15
3	Fall	384	58	378	44									188	16	190	14
	Winter													194	16	198	14
	Spring	416	61	412	51	.78	.31	.71	.39	126	28	98	25	198	15	204	14
4	Fall	421	67	413	52									198	16	202	14
	Winter													203	15	208	14
	Spring	455	73	446	57									206	15	213	15
5	Fall	459	72	451	57									205	16	211	15
	Winter													209	15	217	16
	Spring	484	77	477	64	1.04	.30	1.09	.41	149	26	122	25	212	15	222	17
6	Fall	484	72	479	66									210	16	217	16
	Winter													213	15	222	16
	Spring	504	80	499	70									215	15	225	17
7	Fall													214	16	224	17
	Winter													217	16	227	18
	Spring	520	83	519	75									219	15	230	18
8	Fall													218	16	229	18
	Winter													220	16	233	19
	Spring	546	92	538	82	1.30	.39	1.42	.44	169	28	140	22	222	15	235	19

Table 3. Reliabilities of test scales

		BSS (CAT Thurstone scale)				ECLS-K (IRT ability scale θ)				GRD (IRT ability scale 10θ)			
		Reading		Math		Reading		Math		Reading		Math	
Grade	Season	Internal consistency	Test-retest	Internal consistency	Test-retest	Internal consistency	Corr. with next test	Internal consistency	Corr. with next test	Internal consistency	Corr. with next test	Internal consistency	Corr. with next test
K	Fall					.92	.80	.91	.83		.81		.82
	Winter									.90	.84	.91	.83
	Spring					.95	.89	.93	.90		.86		.85
1	Fall	.68	.50	.83	.80	.96	.83	.94	.82		.86		.85
	Winter									.93	.88	.94	.88
	Spring	.84		.87		.96	.76	.94	.73		.83		.82
2	Fall	.89	.73	.87	.80						.87		.87
	Winter									.96	.88	.96	.88
	Spring	.91		.90							.84		.84
3	Fall	.91	.78	.92	.83						.85		.85
	Winter									.95	.86	.95	.87
	Spring	.91		.93		.94	.86	.95	.85		.85		.86
4	Fall	.91	.80	.91	.83						.85		.88
	Winter									.94	.86	.95	.89
	Spring	.93		.93							.85		.88
5	Fall	.92	.81	.92	.83						.86		.89
	Winter									.94	.86	.96	.90
	Spring	.93		.94		.93	.79	.95	.78		.84		.88
6	Fall	.91	.75	.89	.77						.85		.89
	Winter									.94	.85	.96	.90
	Spring	.92		.91							.84		.90
7	Fall	.91	.78	.91	.76						.84		.90
	Winter									.94	.85	.97	.91
	Spring	.92		.92	.77						.83		.91
8	Fall	.88	.77	.91	.78						.84		.91
	Winter										.84	.97	.92
	Spring	.90		.92	.79	.87	.79	.92	.78				

Table 4. Characteristics of the three samples

a. Periods, sample sizes

	BSS	ECLS-K	GRD
Years of testing	1982-1990	1998-2007	2008-10
Students	825	17,825	177,549
Schools	20	977	419
Districts	1	438	25
States	1	41	14

b. Child characteristics

	BSS	ECLS-K	GRD
Race/Ethnicity			
White	45%	58%	53%
Black	55%	16%	18%
Hispanic	0%	18%	19%
Asian, Native Hawaiian, Pacific Islander	0%	4%	6%
American Indian	0%	2%	2%
Multi-ethnic	0%	2%	3%
Female	50%	49%	49%

c. Family characteristics

	BSS	ECLS-K	GRD
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/reduced price lunch)	65%	45%	---

d. School characteristics

	BSS	ECLS-K	GRD
% Free/Reduced price lunch	51%	37%	52%
High poverty schools (>40% poor 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 samples do not have weights.

Table 5. Reading gaps: Growth from 1st through 8th grade

a. Reading gap between non-poor children and poor children

	Unstandardized				Standardized, reliability-adjusted			
Dataset (scale)	Start 1st	End 8th	Growth	%Growth	Start 1st	End 8th	Growth	%Growth
BSS (Thurstone)	8.253*	66.988***	58.735***	712%	0.394**	0.662***	0.268†	68%
	(3.785)	(8.010)	(8.418)		(0.126)	(0.101)	(0.153)	
ECLS-K (# right)	4.442***	14.396***	9.955***	224%	0.355***	0.525***	0.170***	48%
	(0.489)	(0.521)	(0.370)		(0.021)	(0.019)	(0.015)	
ECLS-K (ability θ)	0.255***	0.187***	-0.067***	-26%	0.466***	0.492***	0.026	6%
	(0.010)	(0.008)	(0.008)		(0.022)	(0.021)	(0.015)	
GRD (ability 10θ)	NA	NA						

b. Reading gap between low-poverty and high-poverty schools

	Unstandardized				Standardized, reliability-adjusted			
Dataset (scale)	Start 1st	End 8th	Growth	%Growth	Start 1st	End 8th	Growth	%Growth
BSS (Thurstone)	16.330*	56.104***	39.774***	244%	0.586***	0.660***	0.074	13%
	(6.666)	(9.603)	(8.111)		(0.157)	(0.143)	(0.145)	
ECLS-K (# right)	7.122***	16.700***	9.578***	134%	0.451***	0.646***	0.195***	43%
	(0.677)	(0.592)	(0.519)		(0.032)	(0.029)	(0.021)	
ECLS-K (ability θ)	0.295***	0.256***	-0.039***	-13%	0.560***	0.630***	0.070**	13%
	(0.014)	(0.013)	(0.009)		(0.034)	(0.029)	(0.024)	
GRD (ability 10θ)	6.044***	7.930***	1.886***	31%	0.460***	0.506***	0.046	10%
	(0.389)	(0.448)	(0.418)		(0.026)	(0.031)	(0.028)	

c. Reading gap between children whose mothers did or did not have a high school diploma

	Unstandardized				Standardized, reliability-adjusted			
Dataset (scale)	Start 1st	End 8th	Growth	%Growth	Start 1st	End 8th	Growth	%Growth
BSS (Thurstone)	8.229*	48.363***	40.135***	488%	0.401**	0.489***	0.089	22%
	(3.401)	(7.341)	(7.776)		(0.121)	(0.090)	(0.145)	
ECLS-K (# right)	3.621***	14.966***	11.345***	313%	0.348***	0.546***	0.198***	57%
	(0.793)	(0.633)	(0.640)		(0.030)	(0.022)	(0.026)	
ECLS-K (ability θ)	0.291***	0.197***	-0.094***	-32%	0.515***	0.512***	-0.003	-1%
	(0.011)	(0.010)	(0.009)		(0.031)	(0.022)	(0.029)	
GRD (ability 10θ)	NA	NA						

d. Reading gap between children whose mothers did or did not have a bachelor’s degree

	Unstandardized				Standardized, reliability-adjusted			
Dataset (scale)	Start 1st	End 8th	Growth	%Growth	Start 1st	End 8th	Growth	%Growth
BSS (Thurstone)	NA	NA						
ECLS-K (# right)	2.844***	10.156***	7.312***	257%	0.275***	0.373***	0.099***	36%
	(0.681)	(0.633)	(0.445)		(0.022)	(0.018)	(0.018)	
ECLS-K (ability θ)	0.187***	0.138***	-0.049***	-26%	0.310***	0.351***	0.041*	13%
	(0.008)	(0.008)	(0.005)		(0.023)	(0.018)	(0.019)	
GRD (ability 10θ)	NA	NA						

e. Reading gap between white and black children

	Unstandardized				Standardized, reliability-adjusted			
Dataset (scale)	Start 1st	End 8th	Growth	%Growth	Start 1st	End 8th	Growth	%Growth
BSS (Thurstone)	-3.719	16.942*	20.661*	-556%	-0.188	0.117	0.306*	-162%
	(3.984)	(8.147)	(8.148)		(0.124)	(0.106)	(0.144)	
ECLS-K (# right)	2.537**	16.134***	13.597***	536%	0.226***	0.617***	0.391***	173%
	(0.796)	(0.645)	(0.516)		(0.029)	(0.026)	(0.018)	
ECLS-K (ability θ)	0.192***	0.234***	0.042***	22%	0.319***	0.577***	0.258***	81%
	(0.013)	(0.011)	(0.009)		(0.031)	(0.025)	(0.025)	
GRD (ability 10θ)	3.948***	6.758***	2.810***	71%	0.326***	0.441***	0.115***	35%
	(0.277)	(0.369)	(0.456)		(0.019)	(0.025)	(0.031)	

f. Reading gap between white and Hispanic children

	Unstandardized				Standardized, reliability-adjusted			
Dataset (scale)	Start 1st	End 8th	Growth	%Growth	Start 1st	End 8th	Growth	%Growth
BSS (Thurstone)	NA	NA						
ECLS-K (# right)	1.768†	10.849***	9.081***	514%	0.234***	0.398***	0.164***	70%
	(0.952)	(0.858)	(0.588)		(0.031)	(0.025)	(0.025)	
ECLS-K (ability θ)	0.193***	0.153***	-0.040***	-21%	0.342***	0.350***	0.008	2%
	(0.012)	(0.010)	(0.009)		(0.032)	(0.023)	(0.028)	
GRD (ability 10θ)	6.582***	7.982***	1.400**	21%	0.547***	0.528***	-0.020	-4%
	(0.273)	(0.362)	(0.451)		(0.019)	(0.025)	(0.031)	

*p<.05, **p<.01, ***p<.001. Growth and % Growth indicate how much a gap grows between 1st and 8th grade. NA indicates that a variable is not available in a particular dataset; for example, individual meal subsidy status is not available in the GRD.

Table 6. Math gaps: Growth from 1st through 8th grade

a. Math gap between non-poor children and poor children

	Unstandardized				Standardized, reliability-adjusted			
Dataset (scale)	Start 1st	End 8th	Growth	%Growth	Start 1st	End 8th	Growth	%Growth
BSS (Thurstone)	12.023***	56.436***	44.412***	369%	.534***	.663***	.129	24%
	(2.833)	(6.881)	(6.955)		(.097)	(.100)	(.127)	
ECLS-K (# right)	5.564***	1.184***	4.619***	83%	.473***	.451***	-.021	-4%
	(.458)	(.443)	(.311)		(.021)	(.018)	(.017)	
ECLS-K (ability θ)	.240***	.175***	-.066***	-27%	.515***	.436***	-.079***	-15%
	(.010)	(.008)	(.007)		(.023)	(.019)	(.016)	
GRD (ability 10θ)	NA	NA						

b. Math gap between low-poverty and high-poverty schools

	Unstandardized				Standardized, reliability-adjusted			
Dataset (scale)	Start 1st	End 8th	Growth	%Growth	Start 1st	End 8th	Growth	%Growth
BSS (Thurstone)	14.889**	47.320***	32.432***	218%	0.582***	0.632***	0.050	9%
	(5.622)	(8.287)	(6.699)		(0.140)	(0.144)	(0.122)	
ECLS-K (# right)	7.923***	12.145***	4.223***	53%	0.593***	0.584***	-0.009	-2%
	(0.572)	(0.491)	(0.411)		(0.032)	(0.029)	(0.018)	
ECLS-K (ability θ)	0.296***	0.242***	-0.054***	-18%	0.629***	0.572***	-0.058**	-9%
	(0.014)	(0.012)	(0.009)		(0.034)	(0.028)	(0.022)	
GRD (ability 10θ)	5.918***	11.630***	5.712***	97%	0.397***	0.575***	0.178***	45%
	(0.445)	(0.496)	(0.412)		(0.029)	(0.033)	(0.027)	

c. Math gap between children whose mothers did or did not have a high school diploma (or equivalent)

	Unstandardized				Standardized, reliability-adjusted			
Dataset (scale)	Start 1st	End 8th	Growth	%Growth	Start 1st	End 8th	Growth	%Growth
BSS (Thurstone)	8.552***	39.183***	30.632***	358%	0.431***	0.469***	0.038	9%
	(2.354)	(6.615)	(6.626)		(0.089)	(0.093)	(0.119)	
ECLS-K (# right)	4.785***	9.690***	4.905***	103%	0.444***	0.425***	-0.019	-4%
	(0.629)	(0.468)	(0.547)		(0.027)	(0.022)	(0.022)	
ECLS-K (ability θ)	0.245***	0.161***	-0.084***	-34%	0.523***	0.404***	-0.119***	-23%
	(0.012)	(0.009)	(0.011)		(0.030)	(0.021)	(0.027)	
GRD (ability 10θ)	NA	NA						

d. Math gap between children whose mothers did or did not have a bachelors degree

	Unstandardized				Standardized, reliability-adjusted			
Dataset (scale)	Start 1st	End 8th	Growth	%Growth	Start 1st	End 8th	Growth	%Growth
BSS (Thurstone)	NA	NA						
ECLS-K (# right)	4.180***	7.733***	3.552***	85%	0.396***	0.372***	-0.024*	-6%
	(0.466)	(0.426)	(0.346)		(0.017)	(0.018)	(0.011)	
ECLS-K (ability θ)	0.182***	0.141***	-0.041***	-23%	0.374***	0.346***	-0.028†	-8%
	(0.007)	(0.007)	(0.005)		(0.019)	(0.015)	(0.016)	
GRD (ability 10θ)	NA	NA						

e. Math gap between white and black children

	Unstandardized				Standardized, reliability-adjusted			
Dataset (scale)	Start 1st	End 8th	Growth	%Growth	Start 1st	End 8th	Growth	%Growth
BSS (Thurstone)	9.125**	24.798***	15.673*	172%	0.156	0.211*	0.056	36%
	(2.998)	(6.879)	(6.651)		(0.098)	(0.105)	(0.120)	
ECLS-K (# right)	6.277***	14.894***	8.617***	137%	0.528***	0.714***	0.186***	35%
	(0.635)	(0.448)	(0.520)		(0.029)	(0.026)	(0.020)	
ECLS-K (ability θ)	0.267***	0.283***	0.016†	6%	0.552***	0.679***	0.127***	23%
	(0.014)	(0.011)	(0.010)		(0.031)	(0.024)	(0.027)	
GRD (ability 10θ)	6.569***	9.672***	3.103***	47%	0.462***	0.490***	0.028	6%
	(0.272)	(0.362)	(0.447)		(0.018)	(0.024)	(0.030)	

f. Math gap between white and Hispanic children

	Unstandardized				Standardized, reliability-adjusted			
Dataset (scale)	Start 1st	End 8th	Growth	%Growth	Start 1st	End 8th	Growth	%Growth
BSS (Thurstone)	NA	NA						
ECLS-K (# right)	5.587***	8.230***	2.643***	47%	0.507***	0.381***	-0.126***	-25%
	(0.571)	(0.420)	(0.476)		(0.026)	(0.023)	(0.018)	
ECLS-K (ability θ)	0.241***	0.150***	-0.091***	-38%	0.533***	0.350***	-0.183***	-34%
	(0.011)	(0.010)	(0.008)		(0.028)	(0.021)	(0.024)	
GRD (ability 10θ)	7.196***	8.753***	1.558***	22%	0.531***	0.433***	-0.098***	-18%
	(0.262)	(0.349)	(0.434)		(0.017)	(0.023)	(0.029)	

*p<.05, **p<.01, ***p<.001. NS means p>.05. Growth and % Growth indicate how much a gap grows between 1st and 8th grade. NA means that the variable or grade range is not available in a particular dataset. For example, individual meal subsidy status is not available in the GRD.

Table 7. Reading gap growth, per month, during school and summer

a. Reading gap growth between non-poor children and poor children

		Unstandardized			Gap grows faster in	Standardized, reliability-adjusted			Gap grows faster in
Grades	Dataset (scale)	School	Summer	Difference		School	Summer	Difference	
K-1	ECLS-K (# right)	0.530***	0.229*	-0.301**	School	0.000	-0.002	-0.002	NS
		(0.024)	(0.097)	(0.111)		(0.001)	(0.004)	(0.005)	
	ECLS-K (ability θ)	-0.007***	0.012***	0.019***	Summer	-0.009***	0.015**	0.024***	Summer
		(0.001)	(0.002)	(0.003)		(0.001)	(0.004)	(0.005)	
	GRD (ability 10θ)	NA	NA						
1-6	BSS (Thurstone)	0.195	2.111**	1.916*	Summer	-0.005	0.033*	0.038*	Summer
		(0.195)	(0.700)	(0.850)		(0.004)	(0.014)	(0.017)	
	GRD (ability 10θ)	NA	NA						
K-8	GRD (ability 10θ)	NA	NA						

b. Reading gap growth between low-poverty and high-poverty schools

		Unstandardized			Gap grows faster in	Standardized, reliability-adjusted			Gap grows faster in
Grades	Dataset (scale)	School	Summer	Difference		School	Summer	Difference	
K-1	ECLS-K (# right)	0.470***	0.300*	-0.169	NS	-0.002	0.007	0.009	NS
		(0.032)	(0.152)	(0.177)		(0.001)	(0.006)	(0.007)	
	ECLS-K (ability θ)	-0.007***	0.017***	0.024***	Summer	-0.010***	0.026***	0.036***	Summer
		(0.000)	(0.002)	(0.003)		(0.002)	(0.007)	(0.008)	
	GRD (ability 10θ)	0.168***	-0.009	-0.177*	School	0.008***	-0.019***	-0.027***	School
		(0.021)	(0.079)	(0.090)		(0.001)	(0.005)	(0.006)	
1-6	BSS (Thurstone)	-0.366†	3.428***	3.794***	Summer	-0.014***	0.057***	0.071***	Summer
		(0.189)	(0.668)	(0.813)		(0.004)	(0.014)	(0.017)	
	GRD (ability 10θ)	-0.008	0.149***	0.157***	Summer	0.002***	-0.007***	-0.009***	School
		(0.008)	(0.029)	(0.035)		(0.001)	(0.002)	(0.002)	
K-8	GRD (ability 10θ)	0.040***	0.023	-0.017	NS	0.005***	-0.014***	-0.019***	School
		(0.008)	(0.024)	(0.029)		(0.001)	(0.002)	(0.002)	

c. Reading gap growth between children whose mothers did or did not complete a high school diploma (or equivalent)

		Unstandardized			Gap grows faster in	Standardized, reliability-adjusted			Gap grows faster in
Grades	Dataset (scale)	School	Summer	Difference		School	Summer	Difference	
K-1	ECLS-K (# right)	0.613***	0.343	-0.270	NS	0.003	0.005	0.002	NS
		(0.040)	(0.198)	(0.228)		(0.002)	(0.008)	(0.009)	
	ECLS-K (ability θ)	-0.007***	0.020***	0.026***	Summer	-0.007***	0.029***	0.036***	Summer
		(0.000)	(0.002)	(0.003)		(0.002)	(0.009)	(0.010)	
	GRD (ability 10θ)	NA	NA						
1-6	BSS (Thurstone)	0.246	0.981	0.734	NS	-0.003	0.013	0.016	NS
		(0.196)	(0.702)	(0.861)		(0.004)	(0.015)	(0.018)	
	GRD (ability 10θ)	NA	NA						
K-8	GRD (ability 10θ)	NA	NA						

d. Reading gap growth between children whose mothers did or did not complete a bachelor's degree

		Unstandardized			Gap grows faster in	Standardized, reliability-adjusted			Gap grows faster in
Grades	Dataset (scale)	School	Summer	Difference		School	Summer	Difference	
K-1	ECLS-K (# right)	0.444***	0.132	-0.313	NS	-0.001	-0.004	-0.003	NS
		(0.029)	(0.139)	(0.160)		(0.001)	(0.005)	(0.006)	
	ECLS-K (ability θ)	-0.006***	0.008***	0.015***	Summer	-0.008***	0.007	0.015*	Summer
		(0.000)	(0.001)	(0.002)		(0.001)	(0.006)	(0.007)	
	GRD (ability 10θ)	NA	NA						
1-6	BSS (Thurstone)	NA	NA						
	GRD (ability 10θ)	NA	NA						
K-8	GRD (ability 10θ)	NA	NA						

e. Reading gap growth between white and black children

		Unstandardized			Gap grows faster in	Standardized, reliability-adjusted			Gap grows faster in
Grades	Dataset (scale)	School	Summer	Difference		School	Summer	Difference	
K-1	ECLS-K (# right)	0.578****	-0.111	-0.689****	School	0.012****	-0.022****	-0.034****	School
		(0.031)	(0.149)	(0.171)		(0.001)	(0.004)	(0.005)	
	ECLS-K (ability θ)	0.001**	0.001	0.000	NS	0.008****	-0.011	-0.019*	School
		(0.000)	(0.003)	(0.003)		(0.001)	(0.007)	(0.008)	
	GRD (ability 10θ)	0.193***	-0.389***	-0.581***	School	0.005*	-0.026***	-0.030***	School
		(0.028)	(0.105)	(0.119)		(0.002)	(0.007)	(0.008)	
1-6	BSS (Thurstone)	0.044	1.307*	1.262	NS	0.001	0.019	0.017	NS
		(0.182)	(0.643)	(0.783)		(0.004)	(0.013)	(0.016)	
	GRD (10θ)	0.070***	-0.079*	-0.149***	School	0.005***	-0.014***	-0.019***	School
		(0.010)	(0.036)	(0.044)		(0.001)	(0.002)	(0.003)	
K-8	GRD (ability 10θ)	0.077***	-0.134***	-0.211***	School	0.005***	-0.016***	-0.020***	School
		(0.010)	(0.030)	(0.038)		(0.001)	(0.002)	(0.003)	

f. Reading gap growth between white and Hispanic children

		Unstandardized			Gap grows faster in	Standardized, reliability-adjusted			Gap grows faster in
Grades	Dataset (scale)	School	Summer	Difference		School	Summer	Difference	
K-1	ECLS-K (# right)	0.447****	-0.001	-0.448*	School	-0.003	0.003	0.005	NS
		(0.038)	(0.189)	(0.218)		(0.002)	(0.007)	(0.009)	
	ECLS-K (ability θ)	-0.008****	0.010****	0.018****	Summer	-0.014****	0.025**	0.039****	Summer
		(0.001)	(0.002)	(0.003)		(0.002)	(0.009)	(0.010)	
	GRD (ability 10θ)	0.130***	-0.123	-0.254*	School	-0.005**	0.004	0.010	NS
		(0.028)	(0.100)	(0.115)		(0.002)	(0.007)	(0.008)	
1-6	BSS (Thurstone)	NA	NA						
	GRD (ability 10θ)	0.014	-0.047	-0.060	NS	-0.002*	-0.002	0.000	NS
		(0.010)	(0.036)	(0.044)		(0.001)	(0.002)	(0.003)	
K-8	GRD (ability 10θ)	0.041***	-0.051+	-0.092*	School	-0.001*	0.001	0.003	NS
		(0.009)	(0.029)	(0.037)		(0.001)	(0.002)	(0.003)	

*p<.05, **p<.01, ***p<.001. NS means p>.05. NA means that the variable or grade range is not available in a particular dataset.

Table 8. Math gaps: Monthly gap growth rates during school and summer

a. Math gap growth between non-poor children and poor children

		Unstandardized			Gap grows faster in	Standardized, reliability-adjusted			Gap grows faster in
Grades	Dataset (scale)	School	Summer	Difference		School	Summer	Difference	
K-1	ECLS-K (# right)	0.273***	0.327***	0.054	NS	-0.008***	0.010*	0.018**	Summer
		(0.018)	(0.086)	(0.100)		(0.001)	(0.005)	(0.006)	
	ECLS-K (ability θ)	-0.008***	0.009***	0.016***	Summer	-0.012***	0.014**	0.026***	Summer
		(0.000)	(0.002)	(0.003)		(0.001)	(0.005)	(0.006)	
	GRD (ability 10θ)	NA	NA						
1-6	BSS (Thurstone)	0.188	1.420**	1.232*	Summer	-0.010**	0.047***	0.057***	Summer
		(0.139)	(0.475)	(0.571)		(0.003)	(0.012)	(0.014)	
	GRD (ability 10θ)	NA	NA						
K-8	GRD (ability 10θ)	NA	NA						

b. Math gap growth between low-poverty and high-poverty schools

		Unstandardized			Gap grows faster in	Standardized, reliability-adjusted			Gap grows faster in
Grades	Dataset (scale)	School	Summer	Difference		School	Summer	Difference	
K-1	ECLS-K (# right)	0.239***	0.338**	0.099	NS	-0.009***	0.011*	0.020***	Summer
		(0.026)	(0.123)	(0.142)		(0.001)	(0.005)	(0.006)	
	ECLS-K (ability θ)	-0.007***	0.009***	0.016***	Summer	-0.012***	0.014*	0.026***	Summer
		(0.001)	(0.003)	(0.003)		(0.001)	(0.007)	(0.008)	
	GRD (ability 10θ)	0.088***	0.241**	0.153+	Summer	0.006***	-0.005	-0.011+	School
		(0.020)	(0.076)	(0.086)		(0.001)	(0.005)	(0.006)	
1-6	BSS (Thurstone)	0.223†	0.728	0.505	NS	-0.008*	0.034**	0.042**	Summer
		(0.133)	(0.453)	(0.545)		(0.003)	(0.012)	(0.014)	
	GRD (ability 10θ)	0.113***	-0.213***	-0.326***	School	0.005***	-0.011***	-0.016***	School
		(0.008)	(0.028)	(0.034)		(0.001)	(0.002)	(0.002)	
K-8	GRD (ability 10θ)	0.130***	-0.155***	-0.285***	School	0.005***	-0.011***	-0.017***	School
		(0.007)	(0.023)	(0.028)		(0.000)	(0.002)	(0.002)	

c. Math gap growth between children whose mothers only completed high school and children whose mothers dropped out

		Unstandardized			Gap grows faster in	Standardized, reliability-adjusted			Gap grows faster in
Grades	Dataset (scale)	School	Summer	Difference		School	Summer	Difference	
K-1	ECLS-K (# right)	0.288***	0.314†	0.026	NS	-0.007***	0.006	0.013†	NS
		(0.035)	(0.171)	(0.197)		(0.001)	(0.006)	(0.007)	
	ECLS-K (ability θ)	-0.008***	0.008**	0.016***	Summer	-0.013***	0.014†	0.027**	Summer
		(0.001)	(0.003)	(0.004)		(0.002)	(0.008)	(0.009)	
	GRD (ability 10θ)								
1-6	BSS (Thurstone)	0.268*	0.419	0.150	NS	-0.005	0.019	0.023	NS
		(0.134)	(0.474)	(0.571)		(0.003)	(0.012)	(0.014)	
	GRD (ability 10θ)	NA	NA						
K-8	GRD (ability 10θ)	NA	NA						

d. Math gap growth between children whose mothers went beyond high school and children whose mothers completed high school only

		Unstandardized			Gap grows faster in	Standardized, reliability-adjusted			Gap grows faster in
Grades	Dataset (scale)	School	Summer	Difference		School	Summer	Difference	
K-1	ECLS-K (# right)	0.213***	0.311**	0.098	NS	-0.008***	0.013***	0.021***	Summer
		(0.021)	(0.100)	(0.115)		(0.001)	(0.002)	(0.003)	
	ECLS-K (ability θ)	-0.005***	0.006***	0.011***	Summer	-0.007***	0.007	0.015**	Summer
		(0.000)	(0.001)	(0.002)		(0.001)	(0.005)	(0.006)	
	GRD (ability 10θ)	NA	NA						
1-6	BSS (Thurstone)	NA	NA						
	GRD (ability 10θ)								
K-8	GRD (ability 10θ)								

e. Math gap growth between white and black children

		Unstandardized			Gap grows faster in	Standardized, reliability-adjusted			Gap grows faster in
Grades	Dataset (scale)	School	Summer	Difference		School	Summer	Difference	
K-1	ECLS-K (# right)	0.475*** (0.032)	0.121 (0.154)	-0.353* (0.177)	School	0.008*** (0.001)	-0.010* (0.005)	-0.018** (0.006)	School
	ECLS-K (ability θ)	0.000 (0.001)	0.000 (0.003)	-0.001 (0.003)	NS	0.006*** (0.002)	-0.012 (0.008)	-0.018* (0.009)	School
	GRD (ability 10θ)	0.114*** (0.027)	0.055 (0.100)	-0.060 (0.113)	NS	0.001 (0.002)	0.001 (0.007)	0.000 (0.007)	NS
1-6	BSS (Thurstone)	0.327** (0.126)	-0.211 (0.422)	-0.537 (0.509)	NS	0.001 (0.003)	0.002 (0.011)	0.000 (0.014)	NS
	GRD (ability 10θ)	0.127*** (0.010)	-0.358*** (0.035)	-0.485*** (0.043)	School	0.004*** (0.001)	-0.013*** (0.002)	-0.017*** (0.003)	School
K-8	GRD (ability 10θ)	0.108*** (0.009)	-0.235*** (0.029)	-0.343*** (0.036)	School	0.001+ (0.001)	-0.007*** (0.002)	-0.008** (0.002)	School

f. Math gap growth between white and Hispanic children

		Unstandardized			Gap grows faster in	Standardized, reliability-adjusted			Gap grows faster in
Grades	Dataset (scale)	School	Summer	Difference		School	Summer	Difference	
K-1	ECLS-K (# right)	0.241*** (0.030)	0.170 (0.146)	-0.071 (0.168)	NS	-0.011*** (0.001)	0.009* (0.005)	0.020*** (0.005)	Summer
	ECLS-K (ability θ)	-0.008*** (0.000)	0.002 (0.002)	0.011*** (0.003)	Summer	-0.017*** (0.001)	0.011 (0.007)	0.028*** (0.008)	Summer
	GRD (ability 10θ)	0.009 (0.025)	0.262** (0.093)	0.254* (0.106)	Summer	-0.010*** (0.002)	0.027*** (0.006)	0.038*** (0.007)	Summer
1-6	BSS (Thurstone)								
	GRD (ability 10θ)	0.075*** (0.009)	-0.312*** (0.035)	-0.388*** (0.042)	School	-0.003*** (0.001)	0.003 (0.002)	0.006* (0.003)	Summer
K-8	GRD (ability 10θ)	0.074*** (0.009)	-0.193*** (0.028)	-0.267*** (0.035)	School	-0.005*** (0.001)	0.011*** (0.002)	0.016*** (0.002)	Summer

*p<.05, **p<.01, ***p<.001. NS means p>.05. NA means that the variable or grade range is not available in a particular dataset.

Figures

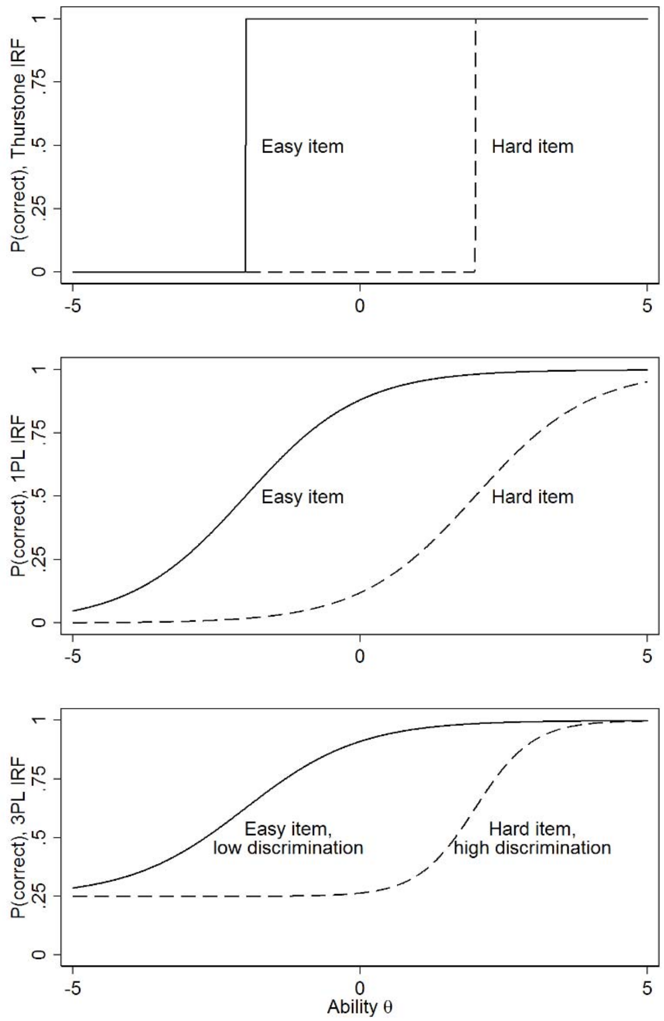


Figure 1. Item response functions (IRF) for three scaling methods: Thurstone scaling (used by the BSS), 1PL IRT scaling (used by the GRD), and 3PL IRT scaling (used by the ECLS-K). For each IRF, the figure illustrates the probability of a correct answer to an easy item and a hard item (d_i=-2 and +2) for students with abilities ranging from low to high (θ_s=-5 to +5). In the 3PL, the figure assumes that that the hard item is also twice as discriminating as the easy one (a_i=2 vs. 1), and that a student who doesn't know the correct answer has a 1 in 4 chance of guessing correctly (c_i=1/4).

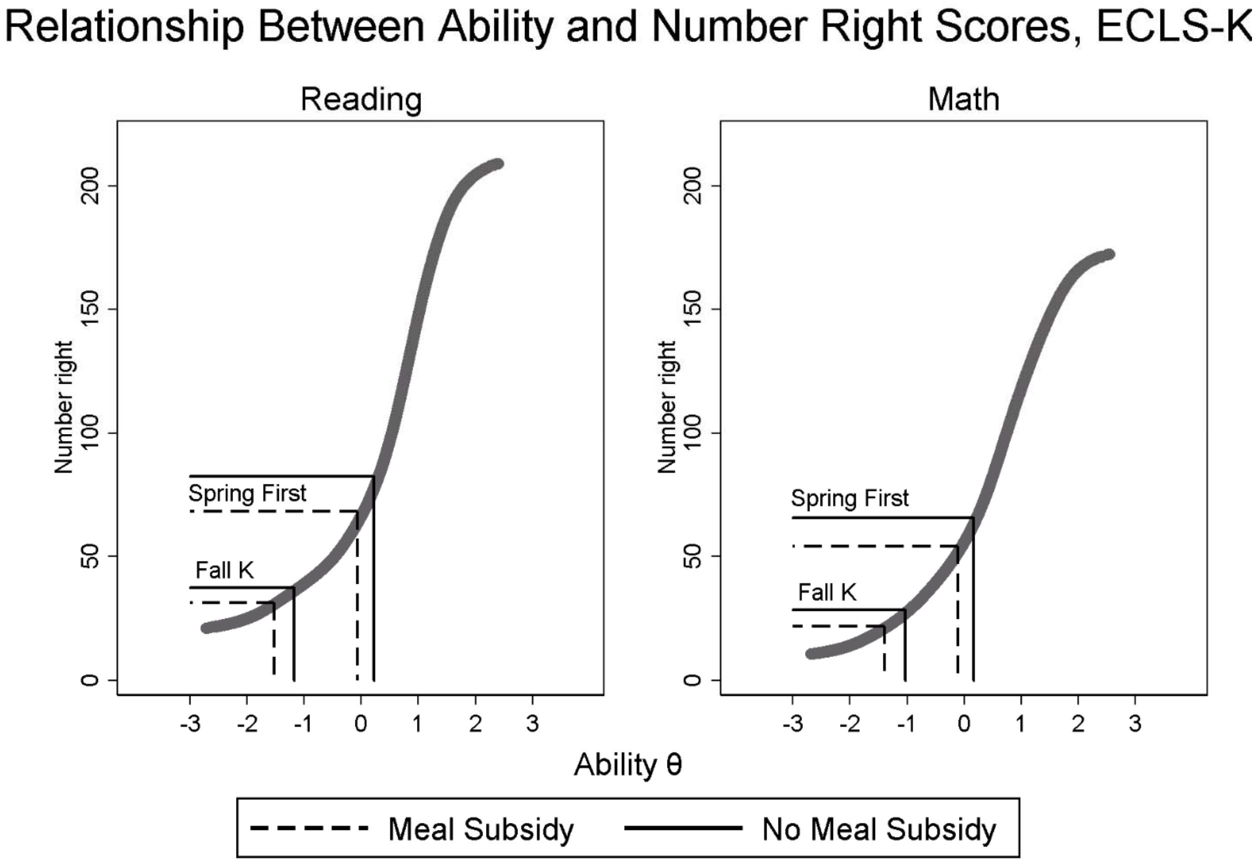


Figure 2. Functional relationship between ability and number-right scores in the ECLS-K. Solid and dotted lines point to the mean score for children with and without meal subsidies at various times in the survey.

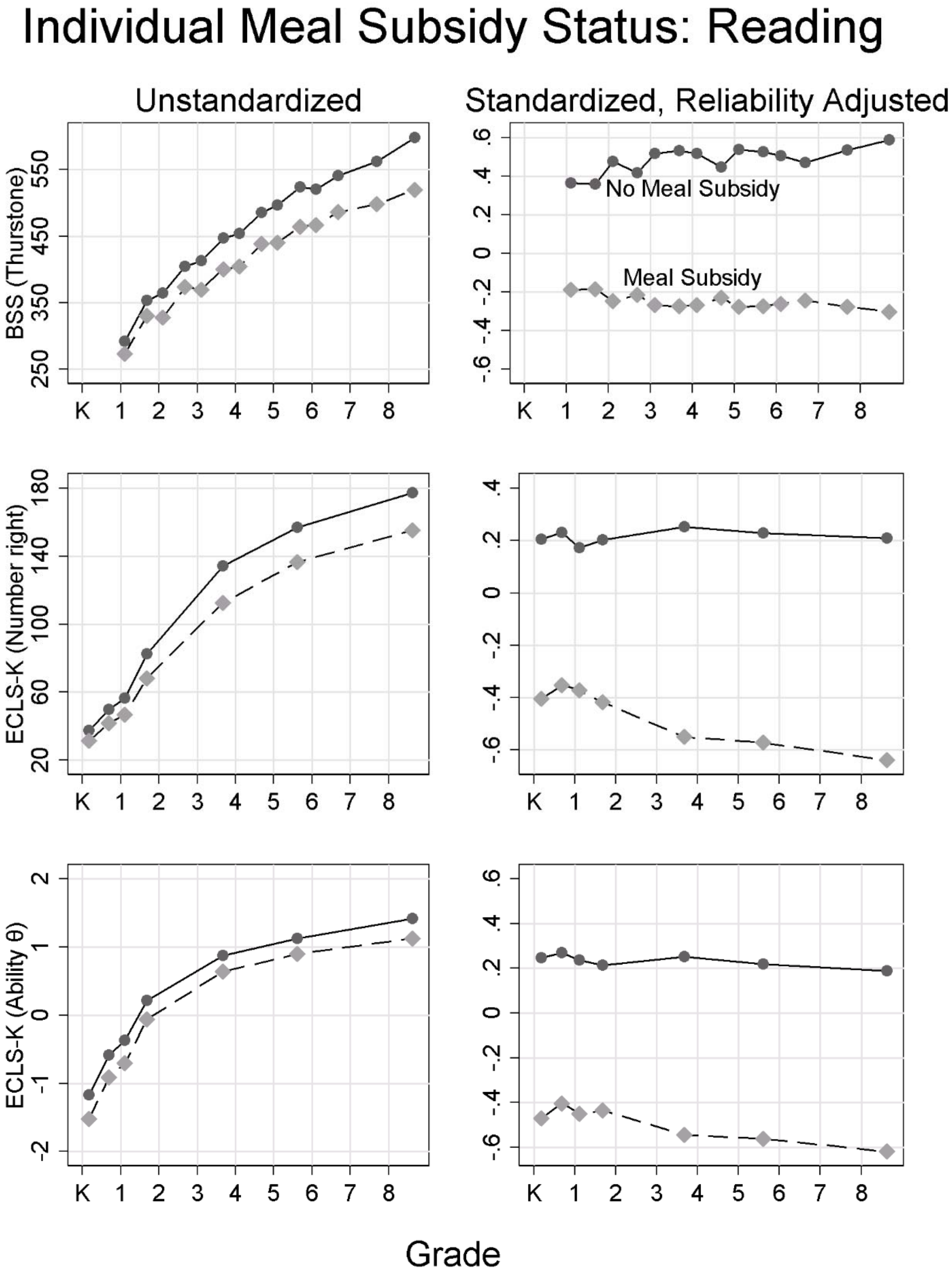


Figure 3. Reading gap between non-poor and poor children, on different scales in the BSS and the ECLS-K.

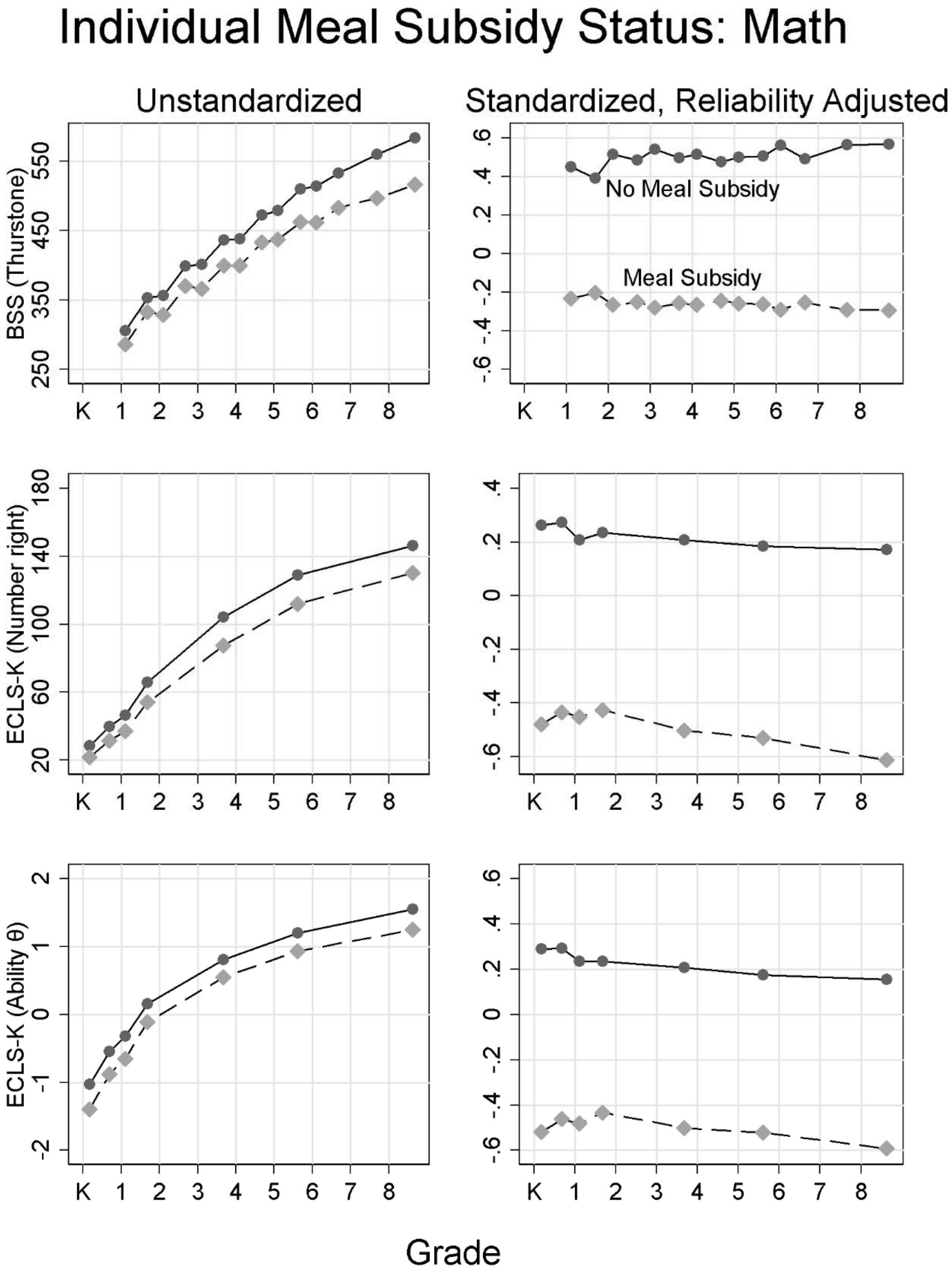


Figure 4. Math gap between non-poor children and poor children.

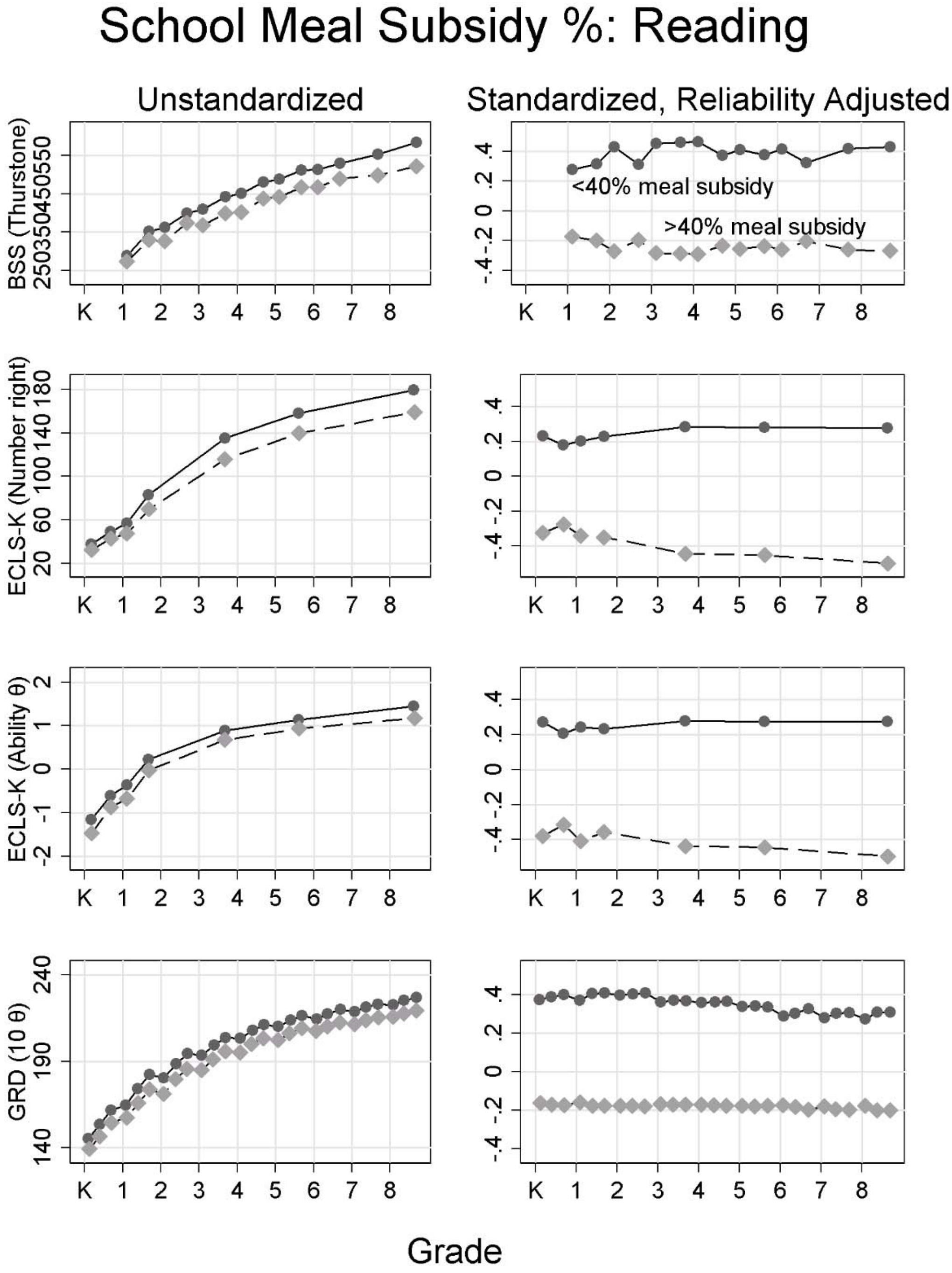


Figure 5. Reading gaps between low-poverty and high-poverty schools.

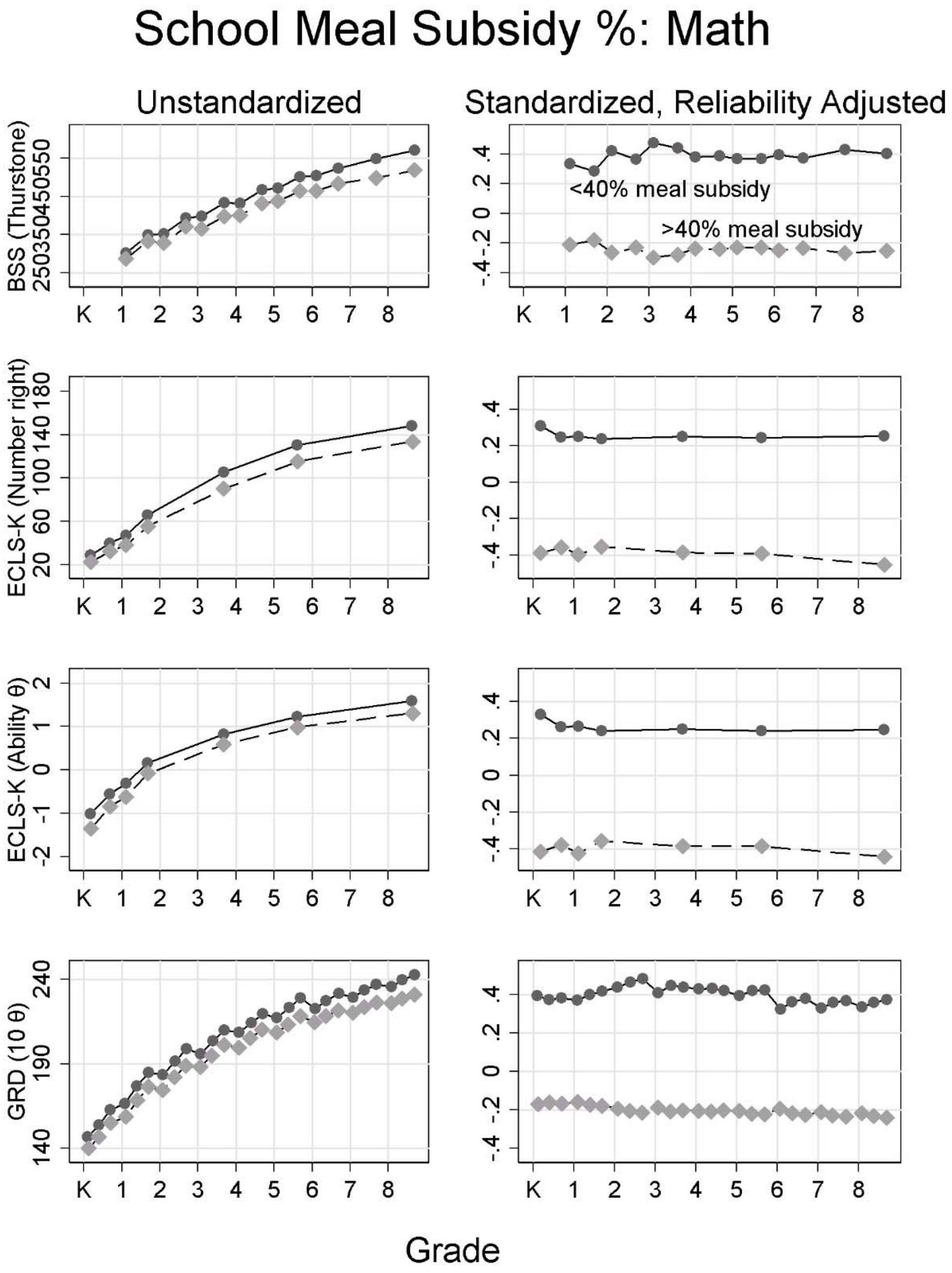


Figure 6. Math gaps between low-poverty and high-poverty schools.

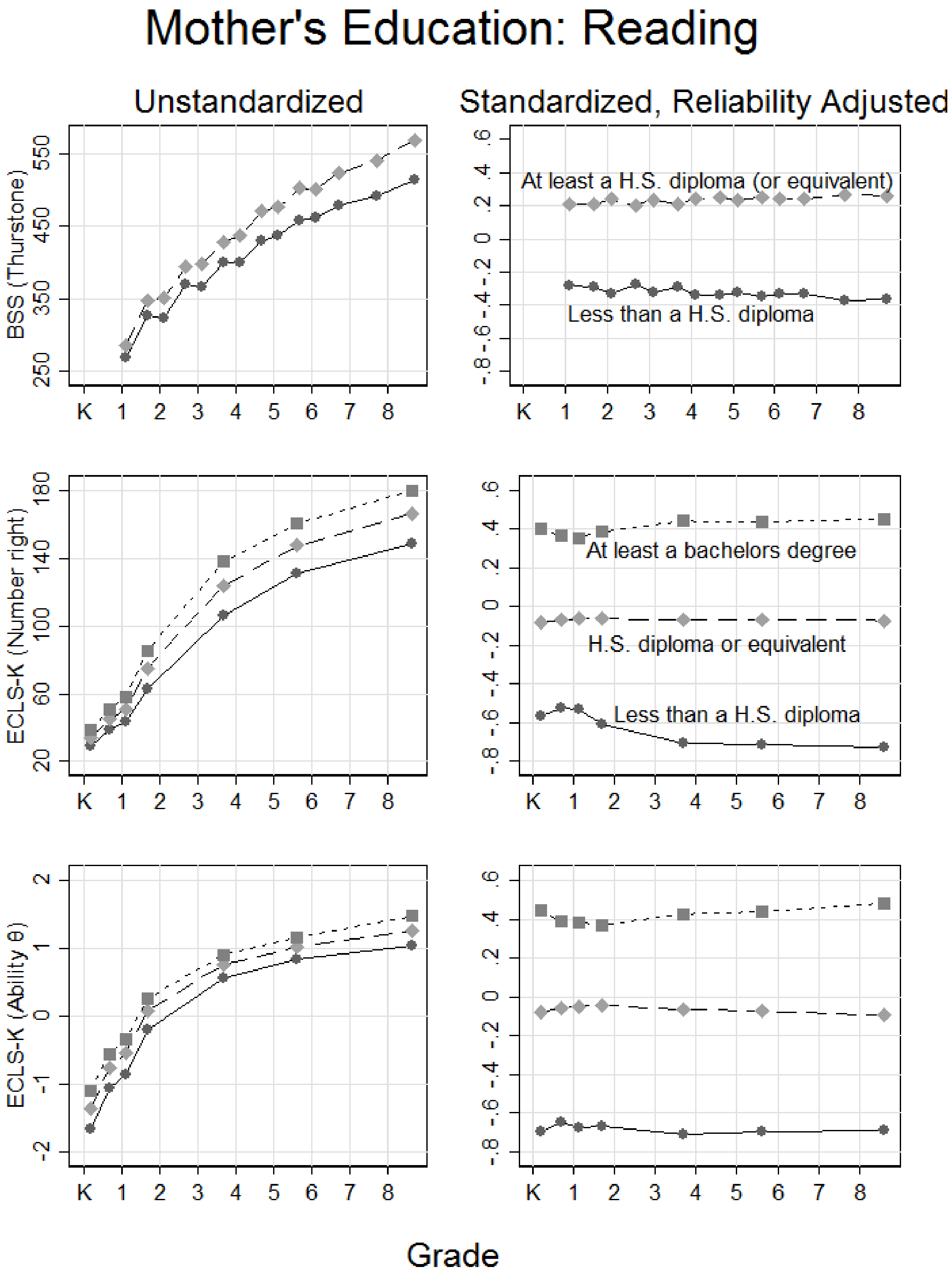
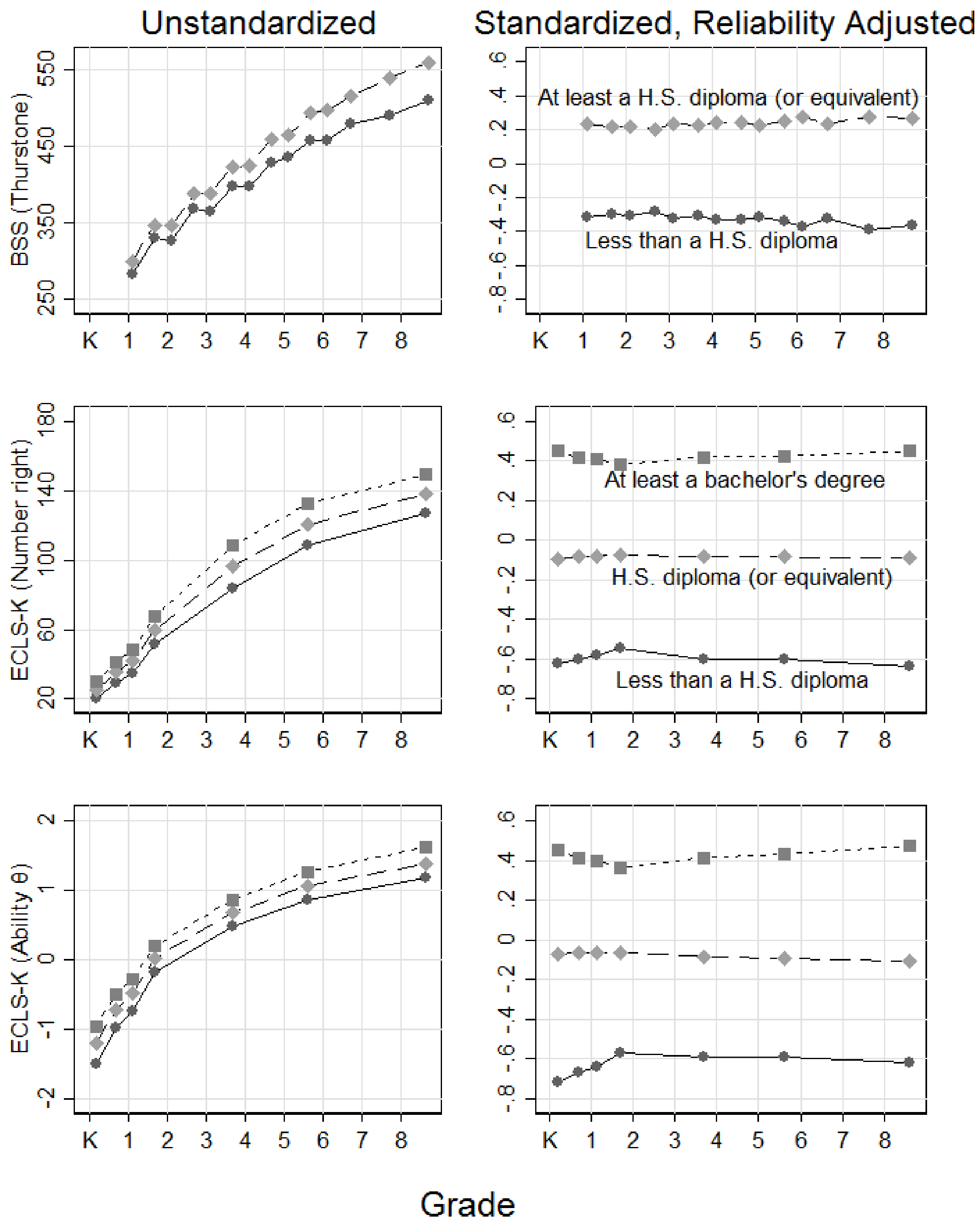


Figure 7. Reading gaps between the children of more and less educated mothers.

Mother's Education: Math



Unstandardized

Standardized, Reliability Adjusted

Figure 8. Math gaps between the children of more and less educated mothers.

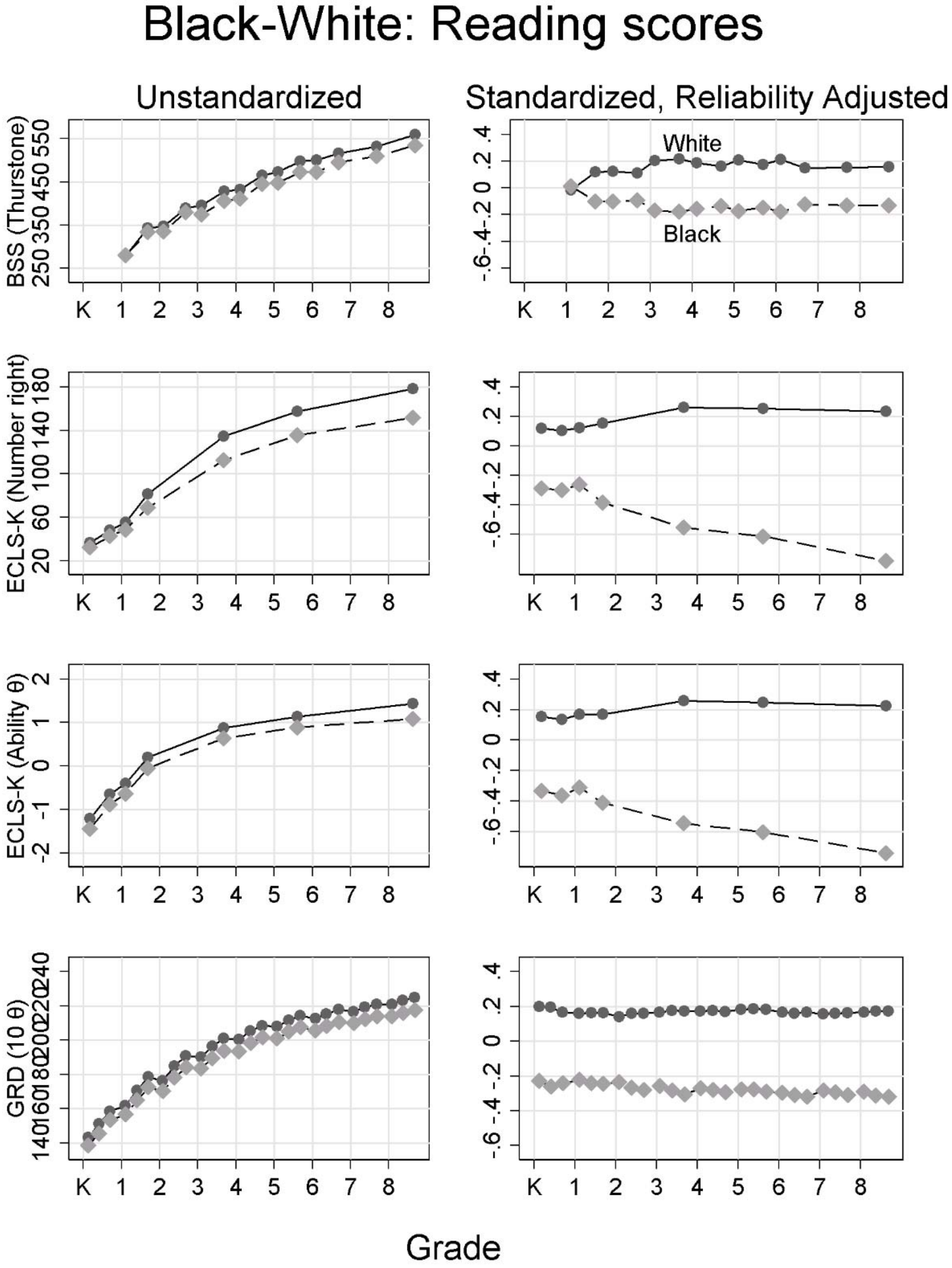


Figure 9. Reading gap between white and black students.

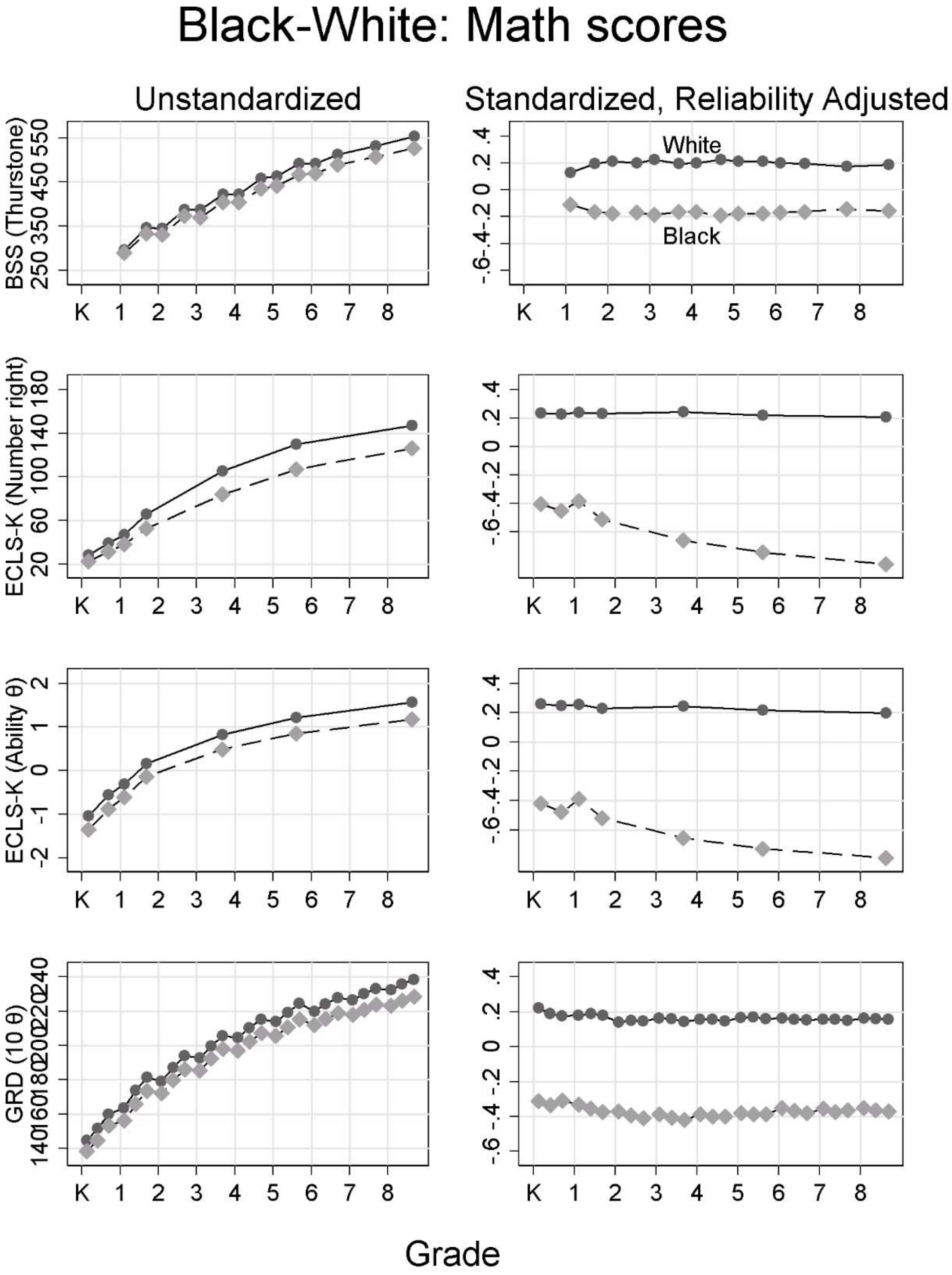


Figure 10. Math gap between white and black students.

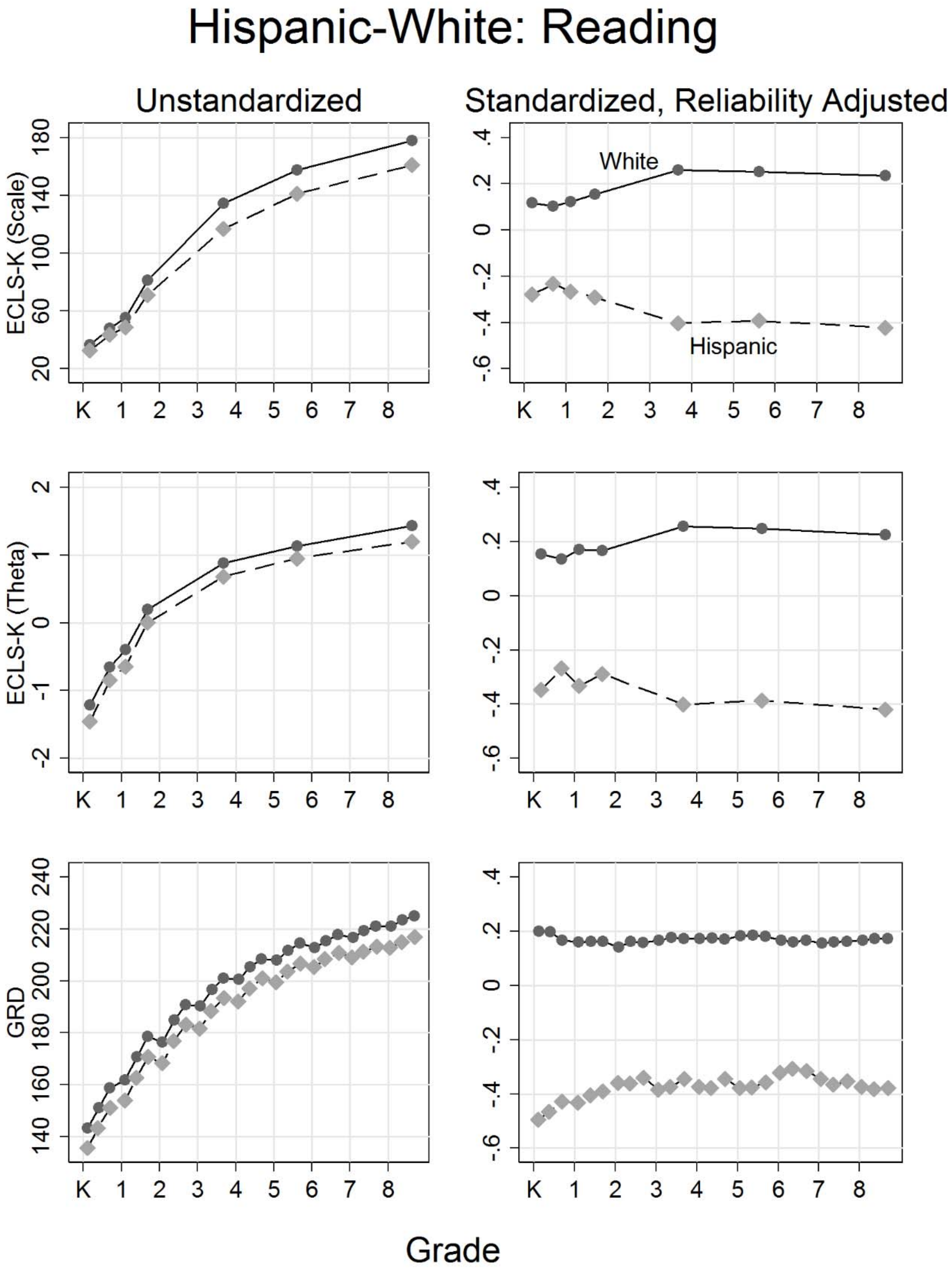


Figure 11. Reading gap between Hispanic and white stduents.

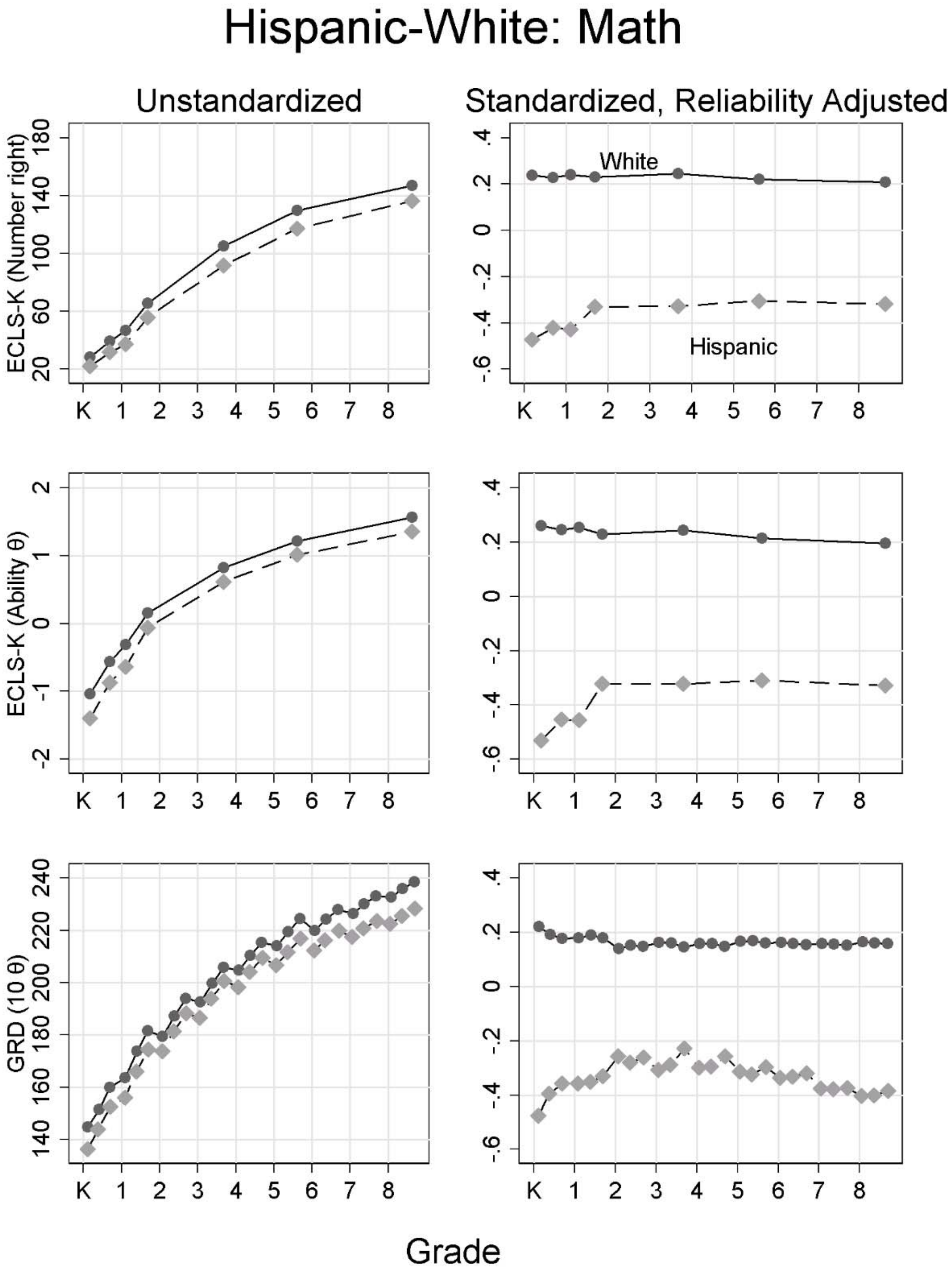


Figure 12. Math gap between Hispanic and white students.