NBER WORKING PAPER SERIES

THE BLACK-WHITE EDUCATION-SCALED TEST-SCORE GAP IN GRADES K-7

Timothy N. Bond
Kevin Lang

Working Paper 19243
http://www.nber.org/papers/w19243

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
July 2013

We are grateful to participants in the Harvard Inequality Seminar, conferences at Linnaeus University, Northwestern and Oberlin, seminars at UC San Diego and UC Irvine and an informal brown bag lunch at Purdue, Ivan Fernandez-Val, Jon Guryan, Jesse Rothstein, and Justin Tobias for helpful comments and suggestions. The usual caveat applies. The views expressed herein are those of the authors and do not necessarily reflect the views of the National Bureau of Economic Research.

The Black-White Education-Scaled Test-Score Gap in Grades K-7
Timothy N. Bond and Kevin Lang
NBER Working Paper No. 19243
July 2013
JEL No. C18,I24,J15

ABSTRACT

We address the ordinality of test scores by rescaling them by the average eventual educational attainment of students with a given test score in a given grade. We show that measurement error in test scores causes this approach to underestimate the black-white test score gap and use an instrumental variables procedure to adjust the gap. While the unadjusted gap grows rapidly in the early school years, particularly in reading, after correction for measurement error, the education-scaled gap is large, exceeds the actual black-white education gap and is roughly constant. Strikingly, the gap in all grades is largely explained by a small number of measures of socioeconomic background. We discuss the interpretation of scales tied to adult outcomes.

Timothy N. Bond
Department of Economics
Krannert School of Management
Purdue University
403 W. State Street
West Lafayette, IN 47907
tnbond@purdue.edu

Kevin Lang
Department of Economics
Boston University
270 Bay State Road
Boston, MA 02215
and NBER
lang@bu.edu

1 Introduction

In Bond and Lang (forthcoming) we argue that test scales are inherently ordinal. We show that without restrictions beyond ordinality on the scales, the bounds on the evolution of the black-white test score gap from kindergarten through third grade are uninformative. They permit results ranging from "the gap is small in kindergarten and declines thereafter" to "the gap is initially nonexistent and grows to be significant." We further suggest that tying test scores to adult outcomes would solve the problem if we are willing to accept the adult outcome as an appropriate metric, an approach taken in a different context by Cunha and Heckman (2008) and Cunha, Heckman and Schennach (2010).

In this paper, we follow our suggestion and show that our optimism was partially justified. We use the relation between educational attainment and test scores to measure the gap and use an instrumental variables procedure to correct for the effect of measurement error in the tests on the scales derived from adult outcomes. We show that, measured in this way the gap in reading is roughly constant from kindergarten through seventh grade at around .7 years of predicted education while the gap in math is close to a full year. When we measure educational attainment not in years but in the associated average log earnings, the pattern is similar, but the gaps are notably larger at around 10 to 12 percent.

Junker, Schofield and Taylor (2012) have been very critical of economists' tendency to treat test scores as measured without error. They show that using the reliability estimates from an underlying item response theory (IRT) model to correct regression estimates can have large effects on their magnitude. Boyd et al (2012) show that measurement error is noticeably larger than suggested by reliability estimates. We show that the scale based on adult outcomes is inherently a shrinkage estimator. When applied to the outcomes of multiple tests, the extent of shrinkage is too large. We derive an instrumental variables estimator that allows us determine how much these "shrunk" estimates need to be "stretched" and confirm the importance of doing so. If we do not correct for excess shrinkage, we observe a pattern similar to that reported by Fryer and Levitt (2004, 2006); the black-white test score gap is initially small but rises rapidly during the early school years.

Murnane et al (2006) argue that the Fryer/Levitt result differs from that found in earlier work (e.g., Jencks and Phillips, 1998) because they use a different test, a point confirmed in our earlier work. Our results suggest that differences between the tests may lead to more measurement error in "achievement tests" in the early grades.

When we combine information from the math and two reading tests, the estimated "education-scaled" gap averages .9 years from kindergarten through seventh grade after adjusting for excess shrinkage. The gap on the PPVT, a test similar to those used in earlier studies, is also .9 years although this is admittedly not corrected for excess shrinkage.

Strikingly, the "education-scaled" test gap in the early years, particularly in math, is at least as large as and probably larger than the actual gap in educational attainment. At the same time, barring some surprising narrowing of the black-white earnings gap, the "earnings/education-scaled" test gap in kindergarten through seventh grades is likely to be less than the future earnings gap, suggesting either that test scores contain more information than just their effect on education, as argued in Neal and Johnson (1996), or continued labor market discrimination.

Finally and strikingly, much, and in some cases all, of the education-scaled gaps can be explained by a small number of controls representing the child's early environment. Results that condition on sociodemographics should be treated with great caution due to the sociological fallacy (Jensen, 1969). However this suggests that our previous inability to explain the test gap by environmental factors may have reflected scaling decisions. The achievement gap may be due to racial differences in socioeconomics rather than a specific racial component in human capital acquisition or the environment more generally.

It is important to understand what our results do and do not mean. It would be easy to interpret our results as saying that the entire black-white education gap is due to pre-school factors. This is true only in the sense that the entire gap can be predicted on the basis of kindergarten test scores. But it is not true if it is interpreted to mean that subsequent events do not affect the gap. To the extent that low kindergarten scores predict future attendance at lower quality schools, less future parental support, etc., the gap is "explained" by factors known in first grade. But this does not mean that differences in these factors no longer matter. Instead, our results tell us that blacks do no better or worse, on average, than would be predicted by their early test scores.

2 Theoretical Framework

2.1 Intuition: Meaning of Rescaling

Suppose we had a very reliable and valid test in the sense that it captured everything about a student's current academic achievement that predicts some adult outcome, say educational attainment, and does so perfectly. Moreover, we have good longitudinal data that provide us with such test scores as students progress through school and on their eventual educational attainment. It would then be possible to rescale the scores on each test by the average eventual educational attainment of students with that score in that grade. If the sample size for each score/grade combination were sufficiently large, we could ignore sampling error in the calculation of average education and simply say that a score of 11 on the test predicts 10.6 years of education, 12 predicts 10.8, and so on.

Of course, no student with a rescaled score of 10.6 would get exactly 10.6 years of education. Some would encounter better teachers in future years. Others would have unanticipated personal problems. But, by construction, they would get an average of 10.6 years of education, and, more generally, by construction, the difference between students' actual and predicted education would be orthogonal to their predicted education.

Because we are temporarily assuming that measurement is perfect, if we observe that blacks and whites have average scores of 12.3 and 13.1 in kindergarten, we can say that the predicted education gap in kindergarten is .8 years. By measuring the gap as students progress through school, we can determine whether it grows or shrinks.

It is important to recognize that even if the test score does not change over time for an individual or group, this *does not* mean that intervening events have been unimportant. Indeed, by construction, the mean population test score does not change over time. It means that subsequent events have (on average) been those predicted by the initial test score. The low test score in kindergarten that may have reflected low parental investment in the child's human capital, low quality pre-school and/or cognitive deficits also predicts continued low investment, low quality elementary and secondary schools and continued cognitive deficits. If the gap between two groups does not change, it means that group membership does not provide additional information about the future beyond the information contained in the test score. We emphasize that this is a fundamental limit of tying test scores to an adult outcome.

2.2 Intuition: Handling Measurement Error

So far we have assumed that the test measurement is perfect. In practice our test scores will be imperfect because of randomness due to guessing, factors affecting a student's performance on a particular day and test content. Students who perform well on a test will, on average, have made lucky guesses, been feeling particularly good and be relatively strong in the content of the exam. If we use a reading test, we will underpredict the educational attainment of students who are particularly strong in math.

For those schooled in the use of tests as performance measures and Bayesian statistics, it is natural to think about shrinkage estimators. An extensive discussion of shrinkage estimators is beyond both the scope of this paper and our expertise. Shrinkage estimators can be thought of in two complementary ways. They are biased estimators that have lower mean-squared error than unbiased estimators. They are estimators that use additional information to improve unbiased estimates. We are quite used to doing this in daily life.

Early in the season, young baseball fans are often excited that player X is batting over .400 (hits divided by official at bats) and might be the first player to do so since Ted Williams in 1941. Older fans would certainly not accept an even bet that the player will finish the season batting above .400 even though the maximum likelihood estimate based on performance to date is above .400. The mere fact that there have been tens of thousands of player/seasons since 1941 and no player has batted above .400, makes it unlikely that such an outstanding performance early in the season will persist. One does not have to be a formally-trained Bayesian to understand this. Depending on how far into the season we are, we may "shrink" our estimate of X's eventual batting average to, say, .350. Similarly, if we compare the test scores of the highest and lowest performing students on a test, and we are asked to predict the gap between them on a similar test, we should predict that the gap will be smaller than the one observed on the initial test.

There are variety of approaches using Bayesian methods to shrink the estimates. Clearly, the best approach would be to observe, in a single season, a large number of batters hitting, say .410 after 100 at bats, and calculating their mean batting average at the end of the season. We could then say that a .410 after 100 at bats corresponds to an expected batting average of perhaps .335 at the end of the season.[1] In general, we need sophisticated statistical techniques to make up for the absence of such data.

Thus if we have the true mean eventual education associated with a test score, there is no need to shrink the estimates. In a sense, we have done so already. If the test score is pure noise, the predicted educational attainment for each test score will just be the overall mean education. As measurement error shrinks, the variance of the rescaled test scores grows.

In fact, if we have multiple tests, we have the opposite problem. Let us continue our baseball example. Suppose we are asked to predict a player's batting average based on ten official at bats. Suppose, we find that the proper shrinkage estimate is .252 + .001*(number of hits) for up to five hits, the most observed in our fictional example. Thus a player batting .200 (two hits) has an estimated true batting average of .254 while one with a batting average of .300 (three hits) has an estimated true batting average of .255. We expect that in reality there should be even more shrinkage, but this degree of shrinkage is adequate to make our point.

Now suppose that we are comparing two teams of 25 players each of whom has had 10 at bats. We take the average of the shrunk scores for each team. Team A's batting average using the shrunk scores is .254. Its players have an average of two hits each. The team has

[1] Baseball fans will understand that we will not have these data. We leave it to those with a greater interest in baseball statistics to figure out how to use *ex ante* information most efficiently in this case.

a total of 50 hits in 250 at bats. If the true mean batting average for the team is .254, the odds of having only 50 hits in 250 at bats is only about 2.5 percent.[2] Suppose that team B's batting average using the shrunk scores is .255. The team has a total of 75 hits in 250 at bats. If its true team batting average is .255, the odds of it getting 75 or more hits in 250 at bats are only about 5 percent. The likelihood that the true team batting averages differ by only .001 is very small. Using the average of the shrunk estimates understates the gap.

The same logic applies when we have many students taking tests. If there is a genuine gap in the average performance of two exogenously selected groups, then the average of the individually shrunk gaps will be smaller than the true average gap. Our application is in many ways more extreme than the baseball example we provide. Our shrinkage estimator is based on a single test, but we will have many students of each race. We will have to undo the natural shrinkage from scaling by average education. We do this by multiplying the estimated gap by the inverse of the shrinkage parameter. We discuss how we can estimate this in the next subsection.

There is one last complication we must discuss. So far, we have assumed that the number of observations with each test score/grade combination is large. In practice, this will not be the case. Therefore, in addition to the other sources of measurement error discussed in the literature, our transformed scales will be subject to sampling error and thus will be the shrunk estimates plus sampling error. However as the number of observations gets large, this sampling error will go to 0.

2.3 A More Formal Presentation

Suppose we are interested in the difference in average "achievement" between blacks and whites in a given grade. This poses two immediate problems. The first is that we cannot observe achievement directly. The second is that achievement has no natural scale.

To solve the latter problem, suppose we are interested in achievement because it predicts future levels of education. We can then normalize achievement at a given time to be in units of expected completed schooling, S, so that for each individual i in grade g,

$$S_i = A_{ig} + \varepsilon_{ig} \tag{1}$$

where, A_{ig} is units of normalized achievement and ε_{ig} is a mean zero error term that reflects determinants of educational attainment that arise after the measurement of achievement. We assume that $E\left(A_{ig}\varepsilon_{ig}\right) = 0$ and $E\left(\varepsilon_{ig}\varepsilon_{jg}\right) = 0$ for $i \neq j$. The interpretation of ε is

[2] The probability that a randomly chosen player would have five hits out of ten is also about .02, but the probability that at least one of fifty players would have five hits is about .66.

important. We will therefore discuss it in greater detail later in this section.

We assume that we have access to a series of test scores. In each grade the test score, τ_{ig}, is a function of A_{ig} and some noise component ν_{ig},

$$\tau_{ig} = \tau_g(A_{ig}) + \nu_{ig}. \tag{2}$$

The function τ_g reflects the fact that each test is scored in some arbitrary fashion so that it is not necessarily linear in (normalized) units of achievement while ν_{ig}, reflects the measurement error associated with any test.

We further assume that measurement error is uncorrelated over time, so that $E[\nu_{ir}\nu_{is}] = 0, \ \forall r \neq s$. As Boyd, Lankford, Loeb and Wyckoff (2012) discuss in detail there are a number of factors that contribute to measurement error other than those captured by test publishers' estimates of reliability. Some of these such as luck or how the student was feeling on a particular day are very likely to be uncorrelated over time. This assumption is less obvious in the case of the items or domains included in the exam. We will present evidence that suggests that such serial correlation is unlikely to be a significant problem in our data.

We derive an "education-normalized scale" in the following fashion. Suppose that τ is discrete, as it is in our data. We define the scale by the population mean of S_i at τ_{ig} :

$$s_g(q) = \frac{\sum_{\tau_{ig}=q} S_i}{N_q} \tag{3}$$

where N_q is the number of individuals in that grade with a score of $\tau = q$. Thus $s_g\left(q\right)$ is the average education ultimately attained by individuals with a score of q on the test in grade g.

We note that in practice this approach will add sampling error to the other sources of measurement error because we have only a finite number of observations at each score in each grade. In part to address this issue, we also provide one set of estimates based on a kernel estimator that combines test scores from three tests.

Henceforth we drop the subscript g when doing so will not cause confusion.

It is tempting to define the education-normalized test score gap by the difference in the mean of s for the two groups. However, this will be incorrect. Suppose

$$\tau_i = A_i + \nu_i \tag{4}$$

and that A and ν are independent and normally distributed with variances σ_A^2 and σ_v^2, respectively. Note that we are fortunate in this example because the test scores have already been scaled to equal the achievement scale.

A standard result from statistical theory gives

$$E\left(A|\tau=a\right)=\beta_1 a+\left(1-\beta_1\right)\overline{A} \tag{5}$$

where

$$\beta_1=\frac{\sigma_A^2}{\sigma_A^2+\sigma_\nu^2}. \tag{6}$$

Now because $s\left(a\right)$ is just the average of A given that τ equals a, by the law of large numbers

$$plim_{N(a)\rightarrow\infty}s(a)=E\left(A|\tau=a\right)=\beta_1 a+\left(1-\beta_1\right)\overline{A}. \tag{7}$$

While the conclusion that s_i is the shrinkage estimate of $\left(A|\tau_i\right)$ does not rely on the assumption of normality, we will maintain this assumption for the rest of the example. Suppose we have a large number of test scores from group c. Then

$$\begin{aligned} s_c &\equiv \frac{\Sigma_{i\in c}s_i}{N_c}=\beta_1\frac{\Sigma_{i\in c}\tau_i}{N_c}+\left(1-\beta_1\right)\overline{A} && (8)\\ &= \beta_1\frac{\Sigma_{i\in c}\left(A_i+\nu_i\right)}{N_c}+\left(1-\beta_1\right)\overline{A} && (9)\end{aligned}$$

where N_c is the number of members of group c. But

$$plim_{N_c\rightarrow\infty}\left(\beta_1\frac{\Sigma_{i\in c}\left(A_i+\nu_i\right)}{N_c}+\left(1-\beta_1\right)\overline{A}\right)=\beta_1 A_c+\left(1-\beta_1\right)\overline{A} \tag{10}$$

where A_c is the mean achievement of group c.

By the same logic

$$plim_{N_c\rightarrow\infty}s_c-plim_{N_b\rightarrow\infty}s_b=\beta_1\left(A_c-A_b\right). \tag{11}$$

To find a consistent estimate of the differences in achievement between the two groups, we need to augment the difference between their mean education-scaled test scores by a factor of β^{-1}.

To address this overcorrection, we approximate the relation between s and A by a linear function

$$s_i=\beta_0+\beta_1 A_i+\mu_i. \tag{12}$$

The linear relation is exact under the normality assumption but need not be otherwise. In principle, we could allow for a more general relation.

If we observed A_i, we could estimate β_1 by regressing s_i on A_i. We do not observe A, but we do observe S which, from (1), is a noisy measure of A. We estimate

$$s_i = \beta_0 + \beta_1 S_i + \varepsilon_i \tag{13}$$

Because $S_i = A_i + \varepsilon_i$ with $E\left(A\varepsilon\right) = 0$, the measurement error is classical. Therefore, we can estimate β_1 consistently if we can find a suitable instrument for S. A natural instrument for S is s_{g-1}, the (renormed) test score from a prior test. However, the renorming includes S_i and therefore is correlated with ε_i. Therefore, we construct a "leave-one-out" instrument which is the average eventual educational attainment of all *other* individuals with the same test score on the prior test

$$s^*_{ig-1} = \frac{\sum_{g_{jg-1}=q, j\neq i} S_{ig-1}}{N_q - 1}. \tag{14}$$

s^*_{ig-1} is correlated with A_{ig} since achievement is persistent.

Therefore we estimate the black-white achievement gap by

$$\widehat{\beta_1^{-1}}\left(s_w - s_b\right) = \frac{\Sigma S_i s^*_{ig-1}}{\Sigma s_{ig} s^*_{ig-1}}\left(s_w - s_b\right). \tag{15}$$

Note that if measurement error is positively correlated over time, we will underestimate β_1^{-1} and therefore the magnitude of the test-score gap. In finite samples, this will be a problem because some individuals will earn the same test score as each other in both grade g and grade $g-1$. Their completed schooling enters the calculation of both s_{ig} and s^*_{ig-1} creating correlation in the error measurement. Asymptotically this correlation goes to zero as both the overall measurement error and correlated measurement error go to zero. We return to a discussion of the importance of small sample bias later.

2.4 Interpretation and the Martingale Property

It is important to remember exactly what our estimates mean and how they should be interpreted. A_{ig} is predicted educational attainment based on achievement in grade g. If, for example, students are tracked in subsequent years based on achievement in grade g and if that tracking affects their subsequent educational attainment, we will attribute the effect of the subsequent tracking to their achievement in grade g. This attribution might be justified since there is a clear sense in which current achievement causes the future environment. However, in other cases current achievement may merely predict future environment without causing it. To take an extreme example, if only school quality determines educational attainment and school quality is perfectly correlated over time, achievement at the end of kindergarten will perfectly predict educational attainment. It would be easy to conclude that "nothing after kindergarten affects the achievement gap." Yet, as both examples make clear, this is not the case.

Finally we note that this interpretation has important implications for what we should find in the data. By definition

$$A_{ig+1} = A_{ig} + \omega_{ig+1}$$

where ω_{ig+1} is the innovation in achievement between grades and

$$E\left(A_{ig}\Sigma_{t=1}^{T}\omega_{ig+t}\right) = 0.$$

Thus, A is a martingale and s is a martingale augmented with measurement error. As discussed in Farber and Gibbons (1996), this means that the covariance of the test scores

$$\begin{aligned}\sigma_{g,g+t} &= E\left(A_o - \overline{A}_0 + \Sigma_{j=1}^{g+t}\omega_j + \mu_{g+t}\right)\left(A_o - \overline{A}_0 + \Sigma_{j=1}^{g}\omega_j + \mu_g\right) \\ &= \sigma_{A_0}^2 + \Sigma_{j=1}^{g}\sigma_{\omega_j}^2 + \sigma_{\mu_{g,g+t}}.\end{aligned}$$

The first two terms are independent of t. Under the assumption that measurement error is uncorrelated over time, the last term is 0 except when t equals 0. Note that in contrast, the covariance is increasing in g. Therefore, the model implies that the lower triangle of the covariance matrix is constant for all terms in a column below the diagonal and increasing from left to right. We can therefore cast light on the importance of serial correlation of the measurement error by examining the covariance matrix of the test scores.

3 Data

The Children of the National Longitudinal Survey of Youth (CNLSY) is a biennial survey of children born to women surveyed in the National Longitudinal Survey of Youth 1979 cohort (NLSY79). The NLSY79 is a longitudinal survey that has followed a sample of 12,686 youths who were age 14 through 21 in December 1978. The survey includes a nationally representative sample, as well as oversamples of blacks, Hispanics, military personnel, and poor whites. The military and poor white oversamples were dropped from later surveys.

Since 1986, the children of women from the NLSY79 have been surveyed and assessed every other year. Separate questions are asked for children and young adults. Children are eligible to enter the childhood sample at birth and advance to the young adult sample at age 15. As of 2010, a total of 11,506 children born to 4,931 unique mothers had been surveyed.

Our focus is on the Peabody Individual Achievement Tests (PIAT). Children were given three PIAT assessments in each survey in which they were age five through fourteen. The

PIAT Mathematics (PIAT-M) measures mathematics skill as typically taught in school. It is comprised of 84 multiple choice questions on a wide range of topics from number recognition to trigonometry. The PIAT Reading Recognition (PIAT-RR) is an oral reading test which assesses children's ability to recognize letters and read single words. The PIAT Reading Comprehension (PIAT-RC) is a test of a child's ability to understand sentences. The PIAT-RC is administered only if the child's score on the PIAT-RR is sufficiently high.[3]

We also examine the Peabody Picture Vocabulary Test (PPVT). The PPVT is a test of receptive vocabulary designed to assess general aptitude. The CNLSY currently administers this test to children at age four or five and age eleven, but due to variation in this policy over time, we observe PPVT scores for children as young as three. We are interested in the PPVT primarily as a measure of achievement before entering grade-school. Therefore we restrict our analysis of the PPVT to those who took the test before age five.

While the survey is a panel by year, we are interested in the racial achievement gap by grade. To convert our data to such a panel, we drop any child we observe in the same grade over multiple surveys. Because the survey was conducted biennially, this restriction binds if the child spent three years in the same grade and thus affects only a handful of individuals. We focus only on the black-white test gap, and drop members of other races. These modifications leave us with an unbalanced panel of 7,343 children born to 3,318 mothers.

The sample is not nationally representative because children born before 1982, when the mothers were age seventeen through twenty-five, are observed only during their later childhood, while those born in later years are observed only during their early childhood. To correct for this non-representativeness, we create custom weights for each grade-test designed to make that subsample nationally representative.[4] Individuals with a valid PIAT-RR raw score below the threshold for taking the RC are included in the construction of the weights for the PIAT-RC but are excluded from the analysis. This avoids putting undue weight in the early grades on a small number of low achieving students who advance to the RC due to randomly high scores. The RC results should, therefore, be interpreted as representative of the population that would have scored sufficiently well on the RR to take the RC exam in that grade. Note that we should view gaps based on the RC with caution especially in kindergarten but also in first grade because the students taking the exam are not fully representative of the overall student population.

Table 1 shows the gap on the age-adjusted percentile scale for each test in each grade,

[3] From 1986 to 1992, the threshold was a raw score of 15 on the PIAT-RR. This threshold was subsequently raised to 18.

[4] We are grateful to Jay Zagorsky of the Center for Human Resources Research for providing us with the program.

using the custom weights discussed above.[5] To ease comparison with other studies, in table 1 we follow convention and normalize the scores in each grade to have mean 0 and standard deviation 1. Because some children leave the child sample and enter the young adult sample during 8th grade, we restrict our attention to kindergarten through grade seven.

Each test tells a different story about the black-white test gap. In kindergarten blacks are .65 standard deviations behind whites on the math test. This gap rises only very slightly through seventh grade. The two reading gaps are initially only very modest but grow to roughly the magnitude of the math gap by third grade. The PPVT, administered earlier than the PIAT tests, shows a gap of over one standard deviation, larger than the gap on any PIAT test in any grade.

Taken together these tests reflect the myriad of findings in the black-white test score gap literature. The reading tests show the pattern demonstrated by Fryer and Levitt (2004, 2006) for the test administered as part of the Early Childhood Longitudinal Survey (Kindergarten sample). The PPVT gap appears similar to that in Jencks and Phillips (1998), while the PIAT-M shows a nearly constant gap that is smaller than the one observed on the PPVT.

As discussed in Bond and Lang (forthcoming), these test gaps are based on arbitrary scaling decisions. Plausible order-preserving transformations of the scales can produce startling different results. The bounds we established in that paper are very large. There we show that without placing more structure on the scales, the gap could be small to modest and decreasing from kindergarten through third grade or small to nonexistent in kindergarten but growing to substantial by the end of third grade, or somewhere in between.

In this paper, we resolve this indeterminacy by relating the scores to economic outcomes, in particular educational attainment. To do so, we construct a sample of 3,853 children who are observed in the panel after age 22 and for whom we know highest grade completed. We lose roughly one-half of our observations on each test, but still have over 1000 observations for all but the earliest PIAT-RC. We again construct custom weights so that each of these test-grade samples is nationally representative.

Table 2 repeats table 1 for this subsample. The magnitudes of the test gaps are generally similar to the full sample, though at times somewhat smaller. This probably reflects the fact that children who are 22 by 2010 were born no later than 1988 when the mothers were 23 to 31 and thus were born to relatively young mothers. By restricting the age of the mothers, we reduce the socioeconomic differences between black and white mothers.[6] Nevertheless,

[5] This scale represents the percentile of the distribution each child's raw score is in for their three month age group. Note that since we are grouping children of different ages within the same grade together, this means that younger children may have higher percentile scores than older children within the same grade despite having answered fewer questions correctly.

[6] On one measure of background, mother's AFQT percentile score, the gap between blacks and whites

the patterns mimic those in table 1: a math test gap that grows only very slowly, a growing reading gap, and a pre-schooling PPVT gap that is larger than that on any subsequent test.

Since we use this sample only to translate test scores into an education scale, the test score gap for the older sample has no direct significance. The real risk is that, because our sample with completed education was born to young mothers, the relation between test scores and educational attainment for this group may not be representative of the entire population in a way that biases our estimate of the "education test-score" gap. It will be apparent that the test scores for the older group are lower than for the sample as a whole, but it is difficult to determine whether this causes any bias in our estimates of the test score gap.

Not surprisingly given past research, we observe a racial gap in educational attainment. Table 2 also displays the difference in average educational attainment between blacks and whites for each test-grade sample. We observe gaps that are generally between .70 and .85 years of education, depending on the sample. This is somewhat higher than we observe for their parents' generation (.70) in the NLSY79 adult sample. It is unclear whether this reflects a change in the gap or the nonrepresentativeness of our older sample.

Our empirical approach depends on the assumption that measurement error is uncorrelated over time. Our model implies that the covariance between a test score in period t and all subsequent test scores $cov(s_g s_{g+2j})$ should be a constant for all $j = 1, 2, 3...$[7] and that $cov\left(s_{g+k} s_{g+k+2j}\right)$ should be nondecreasing in k for k positive. Appendix table A shows the unweighted covariance matrix of the test scores. We have relatively few years for which we can test this hypothesis. In all cases the covariance terms are much smaller than the variances. While we have not formally tested the hypothesis that $cov(s_g s_{g+2j})$ is constant for all test and grade combinations, it does not appear to be severely violated. This suggests that the correlation in measurement error induced by some individuals sharing the same scores in tests in years g and $g-2$ is unlikely to be a serious concern. We address this concern directly later in the paper.

4 Empirical Implementation

In order to obtain estimates of s_g for each grade-test combination, we use all individuals in our sample with a valid score for that grade/test and for whom we observe educational attainment after the age of 22. We then calculate average educational attainment by score

grows by about .03 standard deviations per year increase in mother's age at child's birth. Moreover, white mothers tend to be older than black mothers.

[7] We use $2j$ instead of j because tests are generally administered two years apart.

for that sample. We apply the results of this rescaling to the entire sample. We interpolate s_g for any test scores not present in the over 22 sample. This produces a score on the new scale for each individual with a valid test score on that grade-test.

Figures 1-3 show the relation between the transformed score and the base percentile on each test. While the underlying relation should be strictly increasing, not surprisingly, given the small number of observations with a particular score on a given test, there is notable imprecision in our point estimates. There is, however, a clear overall positive relation between test performance and educational attainment, s, on each test in each grade.

We first estimate the gap between blacks and whites using the s_g scales. However as discussed above, these scales over-correct for measurement error when applied to group averages and thus understate the gap. We correct this by estimating the relation between schooling and the s_g scores. If schooling were a perfect measure of achievement/ability, this would provide an estimate of how much our s_g measure understates achievement. However schooling is achievement measured with error, and so this will attenuate our correction towards zero. We correct this by using the lagged s_g values as instruments. Because the survey is given biennially, we use two year (grade) lagged test scores. For the first grade and kindergarten scores, we use the childhood PPVT s. We also use the PPVT as an instrument for the second grade PIAT-RC due to the small size and selected nature of the sample of children who advance to that test in kindergarten. Each instrument is calculated using the leave-one-out method to avoid correlation arising from the use of the individual's eventual schooling attainment in creating the s scale.[8]

We bootstrap the standard errors. Particularly in the early grades, the distribution of the bootstrap estimates tends to be skewed. Therefore, we present the 95% confidence intervals for all of our estimates, which will be valid under weaker assumptions than required for the use of the normal approximation.

The scale discussed thus far assumes that we value all years of education equally. There are many other possible choices. We do not attempt to consider the full range of alternatives, which would lead to a bounding exercise similar to that in Bond and Lang (forthcoming). Instead, we consider one quasi-monetary scale in which we scale education by the associated mean log annual earnings. While it would be more natural to relate test scores directly to wages or earnings, our sample is too young for this exercise to be informative, and we therefore rely on the indirect approach.

Using 2007 data from the American Community Survey (ACS), we calculate the average

[8] It would be possible to estimate β by using the older sample to calculate $\sigma_{Ss_{g-2}}$ and the full sample to calculate $\sigma_{s_g s_{g-2}}$. This would probably increase the precision of our estimates somewhat, but we are concerned that because the older sample is more homogeneous, calculating covariances from two different samples would be problematic.

log annual earnings by years of education for white males born in 1967.[9] The ACS and CNLSY education categories do not line up exactly, particularly among those with more than a high school diploma. We assign all CNLSY observations whom we observe with 13-15 years of education with the average log income of those in the ACS who are either college dropouts or associate's degree holders. We likewise assign those with 17 or 18 years of education the average of those in the ACS with a master's degree, and those with more than 18 years of education are assigned the average of doctoral and professional degree holders. We exclude from our calculations those who earn less than $6,000 in salary income.[10] We repeat our estimates replacing years of education with the average log earnings values to compute s_g. Early in our research, we also experimented with a scale based on mean earnings rather than mean log earnings. The results with the two measures were broadly similar, and we did not pursue this approach further.

5 Results

5.1 Estimated Achievement Gaps

Table 3 shows the test score gaps as measured by s_g for each PIAT grade-test. The bootstrapped 95 percent confidence intervals are in brackets. These scores have a clear interpretation with respect to adult outcomes: the average expected educational attainment of children with the black distribution of PPVT scores is .88 years lower than that of children with the white distribution. When measured this way, each of the PIAT tests shows a similar pattern. There is some growth in the gap over the first few years of education, but the gap stabilizes by third grade and remains roughly constant through seventh grade. Blacks, however begin much further behind in math than in reading. Based on their math tests in kindergarten, blacks are expected to obtain .55 fewer years of education than do whites, compared to a gap of only .20 years on the reading recognition test. This difference in the gap closes rapidly, so that by the third grade blacks are .67 years of expected education behind in math and .60 and .61 on the reading recognition and comprehension tests. We remind the reader that the reading comprehension results in the earliest grades should be treated with caution because many students in these grades do not perform sufficiently well on the reading recognition test to advance to the comprehension test. Therefore the results for the reading comprehension test are based on a selected sample. Nevertheless the patterns for the two reading tests are similar.

[9] We use 2007 to avoid using earnings data from the recent recession years.

[10] Many of these are small business owners whose income is calculated separately in the ACS.

These results do not account for measurement error and the corresponding over-correction when a shrinkage estimate based on a single test score is applied to a group average. Table 4 reports corrected gaps using the IV strategy discussed above. Strikingly, after correcting for excess shrinkage, the three tests show a consistent story. There is no evidence in any test that the black-white test gap grows over time. On the math test, the kindergarten black-white test gap projects that, on average, blacks will obtain 1.24 fewer years of education than whites do. This gap is substantially larger than the black-white education gap observed in the data, and is consistent with Lang and Manove (2011), who show that blacks obtain more education than whites do conditional on test scores. By seventh grade the gap has, in fact, decreased to .89 years of education, though we cannot reject that it is unchanged. The reading recognition test shows a gap of .64 years of education in kindergarten and remains flat at .68 years in seventh grade.

We are unable to estimate the gap on the reading comprehension test at kindergarten with any precision. While our estimates suggest that this test, as scaled by educational attainment, is mostly noise, we cannot precisely pin down the size of the bias this creates, and thus our confidence interval spans 10 years of education. Using the first grade as our reference point then, we again see no evidence of growth in the test gap through seventh grade. As noted above, however, this is still a somewhat selected sample. Roughly 15% of first graders do not score well enough on the PIAT-RR to take the PIAT-RC. This number is less than 1% in second grade, when we observe a .91 year education gap in performance. From this reference point, the gap falls to .72 by seventh grade, a decline similar to the one on the math test, although this change is again not statistically significant.

It is striking to compare tables 3 and 4. Measurement error and thus the implicit shrinkage on the test declines dramatically as students progress through school. On the math test, the adjustment factor is about 125 percent in kindergarten but only about 25 percent in seventh grade. Similarly, on the reading recognition test the adjustment factor goes from about 3 in kindergarten to .2 in seventh grade.

We note that as Murnane et al (2006) argued and our earlier paper (Bond and Lang, forthcoming) confirmed with other scales, the gap on the early PPVT test is much higher than on the PIAT. Our estimate of the unadjusted gap on the PPVT is .88 years of education. While this is higher than all of our unadjusted gaps, it is somewhat lower than the adjusted gap on the PIAT-M at entry and about the size of some of our early estimates of the reading gap. While we cannot adjust the PPVT gap for measurement error, one plausible explanation for the difference between the early PIAT and PPVT estimates is that the latter test suffers from much less measurement error.

Consistent with this interpretation, the covariance between the PPVT and the two reading tests (see appendix table A) increases sharply between kindergarten and third grade from .21 to .39 for reading recognition and from .14 to .37 for reading comprehension. Note that this is only possible if the PIAT reading tests are doing a better job of capturing skills already acquired by the time the children took the PPVT.[11] In contrast the correlation between the PPVT and math PIAT is roughly constant, going from .34 to .35.

Similarly, we might expect the correlation between child's test score and mother's performance on the Armed Forces Qualifying Test, often used as a measure of general intelligence, would decline as children progress through school. In fact, this correlation increases from kindergarten to second grade for each of the PIAT tests (not shown). While greater measurement error on the kindergarten test than on the second grade test is not the only possible explanation for this regularity, it is surely one of the simplest.

These results show the achievement gap when test scores are calibrated using education and treating all years of education as equally valuable. It is natural to ask whether the results would be similar using other important metrics such as wages or earnings. Unfortunately, the sample of respondents in the CNLSY for whom we have wage data is small and not representative. Therefore, as discussed above, we instead scale education by the earnings associated with each level of education, a non-linear transformation of the education scale.

Table 5 shows the measurement-error corrected results from this exercise. The results confirm the patterns obtained when using completed education to scale the test scores. There is little evidence of a growing achievement gap between blacks and whites. The math test suggests that, given their performance in kindergarten, blacks will earn roughly 17% less than whites do and shows no significant change through seventh grade. While the size of the gap fluctuates across grades, any evidence for a change in the gap is in the direction of blacks catching up rather than falling behind.

The gaps implied by the reading tests are similar and, if anything, lower than those derived from the math test. Still in neither case does table 5 suggest that the gap grows as children progress through school.

Nevertheless, there is also a striking difference between the results in tables 4 and 5. Assuming even a 10 percent return to education, the gaps in math in table 4 suggest a (log) wage gap on the order of .08 to .1. The math gaps in table 5 are all above this range as are the slightly smaller estimated reading gaps. Measured by this dollar metric, the test score gap appears substantially larger.

To address the concern that our results are driven by differences between blacks and

[11] A possible explanation for this increased ability to capture these skills is that they are more correlated with the more advanced skills of third graders than with the sorts of skills generally developed by the end of kindergarten.

whites in both test scores and educational attainment, tables 6 and 7 repeat tables 4 and 5, but use only whites in the calculation of the rescaled test scores. Our estimated gaps are similar whether we include or exclude blacks in the re-scaling of the scores although the latter are less precise.

5.2 Unified Measure of Achievement

Thus far we have considered using adult outcomes as a way to scale the individual subject tests. In this subsection we combine information from all three tests to estimate a single measure of achievement in each grade by forming a conditional expectation of future achievement

$$E[A_{ig}|T_{ig}] = h(T_{ig})$$

where T is the set of tests available for student i and h is the conditional expectation function. Analogous to our earlier discussion, we do not observe achievement directly but observe eventual educational attainment, which reflects achievement in grade g. Following the theory laid out previously, if we can estimate h, we can use instrumental variables to create corrected achievement gaps for each grade.

We estimate h using a multivariate kernel Nadarya-Watson regression estimator. For a set of test scores T, the estimator creates weights for each observation based on the closeness of its test scores to T. The estimator then uses these weights to form a weighted average of the outcome variable (in our case, education). Thus we can generate an expected outcome conditional on the full set of tests.

The weights depend on the choice of kernel function and bandwidth. We select a multivariate Gaussian kernel. For each point, the kernel weights observations around the point so that the density is multivariate normal. The choice of kernel is inconsequential; however the bandwidth is not (Blundell and Duncan, 1998). In a multivariate Gaussian kernel, the bandwidth essentially determines the variance of the density. For bandwidth selection, we follow Silverman's (1986) rule of thumb, so that the bandwidth is proportional to the variance of the distribution in the data.

As previously noted, many children do not advance to the reading comprehension test during the first two years of school. To account for this, we estimate the conditional expectations separately for those who did and did not take the RC exam. In the remaining grades, the very small sample of children who do not advance to the reading comprehension exam is dropped from the analysis.

Table 8 displays the results of this exercise. The first column shows our achievement gap estimates using only the differences in conditional expectations, not adjusting for excess shrinkage. We see the familiar pattern of a rising initial achievement gap. Based on their performance on all tests in kindergarten, blacks are projected to obtain .41 fewer years of education than whites do. This gap rises quickly, however, to .73 years in second grade and remains roughly constant thereafter. In this respect, the results are more similar to those in table 3 for math than for either of the reading tests. This may reflect the poor ability of the early reading tests to predict educational attainment.

However, once we correct for excess shrinkage, the growth in the gap again disappears. The estimates in column two project a future racial difference in educational attainment of .82 years in kindergarten, with little change through seventh grade. Once again, the projected education gap is at least as high if not higher than the actual education gap, consistent with Lang and Manove (2011).

The results in table 8 are broadly consistent with those in table 4. In every grade except 7th, the estimated gap when we use all three tests lies within the range of the gaps produced by using each of the three tests individually. However, the confidence intervals are consistently tighter when we use all three tests, and our estimates appear meaningful even for kindergarten and first grade.

Moreover, the gap averages about .9 years of education, almost exactly what we obtain using the early PPVT test. This suggests that the difference between the results using the PPVT and PIAT tests may not be their content but simply greater measurement error in the latter although we cannot test this directly

In table 9, we repeat the exercise but instead scale the tests to represent the education-predicted mean log incomes of each score. We find similar results to those of table 8. The achievement gap remains steady at about a 12% income difference, which is on par with that shown for the math achievement tests in table 5. Our estimates are also generally more precise than in table 5, though the improvement in precision is not nearly as substantial as with the education-scaled scores.

5.3 Correlated Measurement Error

As discussed above, in some cases more than one individual gets the same pair of scores on, for example, the first and third grade math tests. Suppose that Linda and Mike both got 28th percentile in first grade and 36th percentile in third grade. Then both Linda and Mike's eventual education enter the calculation of the mean education associated with a 36 in third grade. Moreover, when we instrument for Mike's third grade education-scaled score with the mean education of everyone else with a 28 in first grade, Linda's education will also enter that calculation. This creates correlated measurement error in finite samples.[12]

To cast light on the importance of this small sample bias for our sample, we ran four simulations in which we took our actual data and added additional error to the education levels. We added a mean zero normal error with standard deviations of 1, 2, 3 and 4. Since the standard deviation of education conditional on test scores is a little less than 2 in most grades, we in effect experimented with increasing the sampling variance by 50-500 percent.

We conducted the simulation 100 times and compared the mean estimates with our actual estimates. The differences caused by this increase in the sampling error were sufficiently modest that in no case were we able to reject that the simulations produced estimates that, on average, were equal to those obtained with the actual data. And the differences between the mean simulated and actual coefficients were also visually modest, suggesting that small sample bias due to correlated measurement error is not a major concern.[13]

5.4 Achievement Gaps and Sociodemographics

One of the key findings in Fryer and Levitt (2004, 2006) was that the early test gap could be "explained" by a small set of sociodemographic controls. Our earlier work (Bond and Lang, forthcoming) showed that while the gaps after controlling for sociodemographic factors were still sensitive to scale choice, they were much more robust than the raw gaps. In tables 9 and 10, we explore the impact of sociodemographics on our "education-scaled" test gaps.

We select a set of sociodemographic controls from the CNLSY to account for differences in the early childhood environment. We include mother's education and age at first birth, and the child's birth weight. We also include a set of controls for the child's home environment from age 0-2: log family income, log hours worked per week by the mother, whether the child ever lived in a household below the poverty line and categorical variables for number of books in the household, amount of cuddly and plush toys, frequency with which the mother reads to the child, whether the child sees a father-figure daily, and frequency of eating dinner with both parents. When we had multiple observations of these variables between age 0 and 2 we used the mean for income and hours worked, and the median category for the categorical variables.[14] From the year in which the test is administered we control for whether the child sees a father-figure daily and whether there are ten or more children's books in the

[12] Asymptotically there will be lots of such pairs but their mean deviation from expected education will go to 0, so the IV estimator is consistent.

[13] In one case in the experiment which added N(0,16) error, there was a noticable difference between the mean estimate of the experiment and table 4, but the variance around this estimate was much too large to be meaningful.

[14] If children had a median category in between two discrete categories, a new category was created for them.

household, as well as family income and mother's hours worked and poverty status. This set of controls is based on the ones used in the CNLSY by Lang and Sepulveda (2008) to closely match those used by Fryer and Levitt (2004, 2006) in the ECLS-K although it is probably somewhat more extensive that the latter.

We compute the education-scaled test scores and their measurement error corrections as before and then add these controls to our regression to estimate the controlled test gap. Tables 10 and 11 show the results for the education- and mean log income-scaled test scores, respectively. While we lack precision in our estimates of the education-scaled gaps, there is no evidence that the controlled gap increases with schooling. Our estimates using the mean log income-scaled test scores are more precise and tell the same story. Relative to table 5, our controls reduce the gap on every test and in every grade, sometimes quite substantially. In fact, at no point using this scale is the test gap in reading recognition statistically significant once we control for early childhood environment.

One must always be careful in the interpretation of achievement gaps conditional on sociodemographics. As pointed out by Jensen (1969), environment may reflect heritable factors. However, our results in table 11, in particular, suggest that the frequently observed racial test gaps may reflect a common effect of environment on test scores, and not a specific race-based environmental disadvantage.

6 Summary and Conclusions

The comparison of tables 1, 3 and 4 leads us to a strong conclusion. The apparent growth of the black-white test score gap from kindergarten through third grade is likely to be an artifact of measurement error. Fryer and Levitt (2004, 2006) report this growth for the Early Childhood Longitudinal Survey using their base scale. Our earlier work (Bond and Lang, forthcoming) confirms this finding and replicates it using the base scale from a sample from the CNLSY similar to the one in the current paper. We find the same growing gap when we use percentiles (table 1) or our transformation based on eventual educational attainment (table 3) instead of the base scale. Yet when we use instrumental variables to adjust the estimate for the effect of measurement error on the shrinkage associated with rescaling by adult outcomes (table 4), the pattern disappears. Indeed when we use this approach on all three PIAT tests in combination, the gap is large and consistent with the gap on the early PPVT.

We emphasize again that our finding of a constant gap in grades K through 7 does not mean that the gap neither widens nor narrows according to some other metric. It could be that the gap between low-performing and high-performing students narrows or grows over these years. This is difficult, if not impossible, to determine when we only have an ordinal achievement measure. Our results say only that the racial gap does not widen or narrow relative to the education level or the mean log earnings associated with that education level that would be predicted on the basis of early test scores.

In particular, our results neither support nor contradict the view that "it is all over by first grade." They are consistent with this view but also with one in which the later environment is important but in which early test scores predict later environment.

Despite this limit on their interpretation, we believe that our results advance our understanding of the black-white test score gap by elucidating the importance of measurement error and by emphasizing that its evolution does not have a strictly racial component after kindergarten.

References

Blundell, Richard and Alan Duncan, “Kernel Regression in Empirical Microeconomics,” *Journal of Human Resources,* 33 (1998), 62-87.

Bond, Timothy and Kevin Lang, “The Evolution of the Black-White Test Score Gap in Grades K-3: The Fragility of Results,” *Review of Economics and Statistics*, forthcoming.

Boyd, Donald, Hamilton Lankford, Susanna Loeb and James Wyckoff, “Measuring Test Measurement Error: A General Approach,” National Bureau of Economic Research Working Paper 18010, 2012.

Cunha, Flavio and James J. Heckman, “Formulating, Identifying, and Estimating the Technology of Cognitive and Noncognitive Skill Formation,” *Journal of Human Resources* 43 (2008), 738-782.

Cunha, Flavio, James J. Heckman, and Susanne M. Schennach, “Estimating the Technology of Cognitive and Noncognitive Skill Formation,” *Econometrica* 78 (2010), 883-931.

Farber, Henry S. and Robert Gibbons, “Learning and Wage Dynamics,” *Quarterly Journal of Economics* 111 (November 1996):1007-47.

Fryer, Roland G., Jr. and Steven D. Levitt, “Understanding the Black-White Test Score Gap in the First Two Years of School,” *Review of Economics and Statistics* 86 (2004), 447-64.

Fryer, Roland G., Jr. and Steven D. Levitt, “The Black-White Test Score Gap Through Third Grade,” *American Law and Economics Review* 8 (2006), 249-81.

Jencks, Christopher and Meredith Phillips, “The Black-White Test Score Gap: An Introduction,” in Christopher Jencks and Meredith Phillips (Eds.), *The Black-White Test Score Gap* (Washington, DC: Brookings Institution Press, 1998).

Jensen, Arthur R., “How Much Can We Boost IQ and Scholastic Achievement?” *Harvard Educational Review* 39 (1969) 1-123.

Junker, Brian, Lynn Steurle Schofield and Lowell J. Taylor, “The Use of Cognitive Ability Measures as Explanatory Variables in Regression Analysis,” *IZA Journal of Labor Economics*, 1:4 (October 2012).

Lang, Kevin and Michael Manove, “Education and Labor Market Discrimination,” *American Economic Review*, 101(4) (June 2011): 1467-1496.

Lang, Kevin and Carlos E. Sepulveda, “The Black-White Test Score Differential,” unpublished (2008).

Murnane, Richard J., John B. Willett, Kristen L. Bub and Kathleen McCartney, “Understanding Trends in the Black-White Achievement Gaps during the First Years of School,” *Brookings-Wharton Papers on Urban Affairs* (2006), 97-135.

Neal, Derek A., and William R. Johnson. 1996. “The Role of Premarket Factors in Black-White Wage Differences.” *Journal of Political Economy,* 104(5): 869-95.

Silverman, B.W., *Density estimation for statistics and data analysis.* Vol. 26. Chapman & Hall/CRC, 1986.

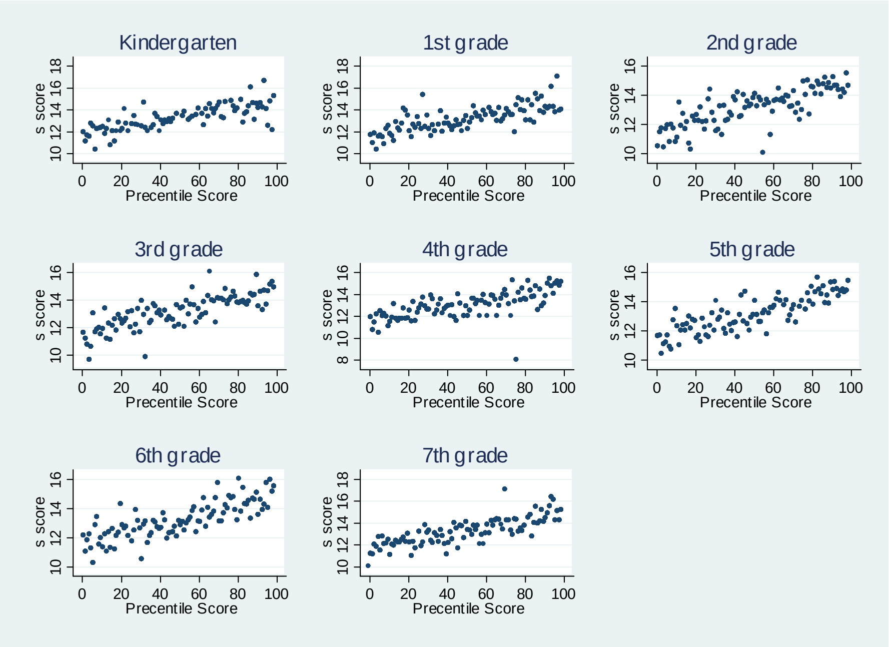

Figure 1: PIAT:M Transformation

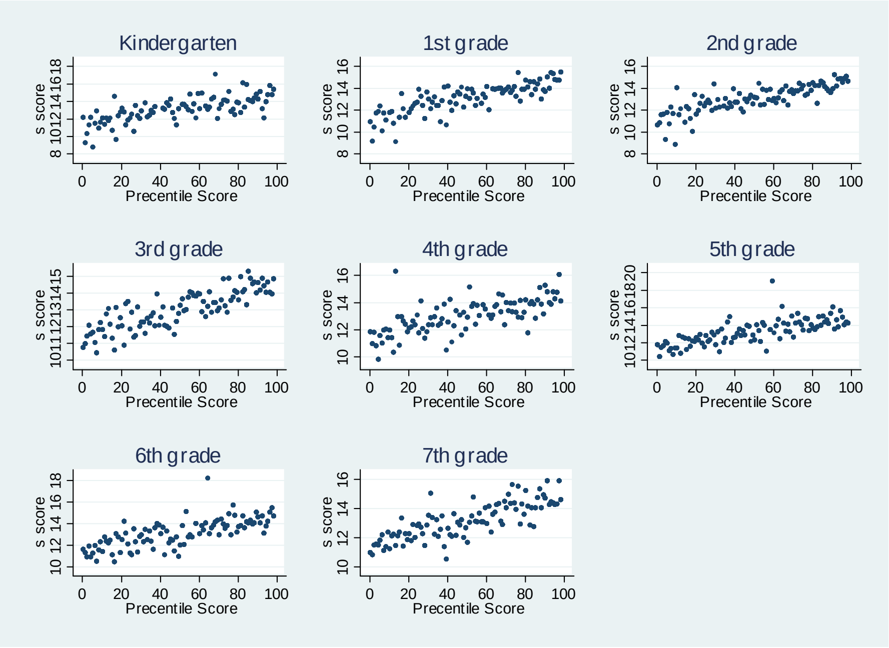

Figure 2: PIAT:RR Transformation

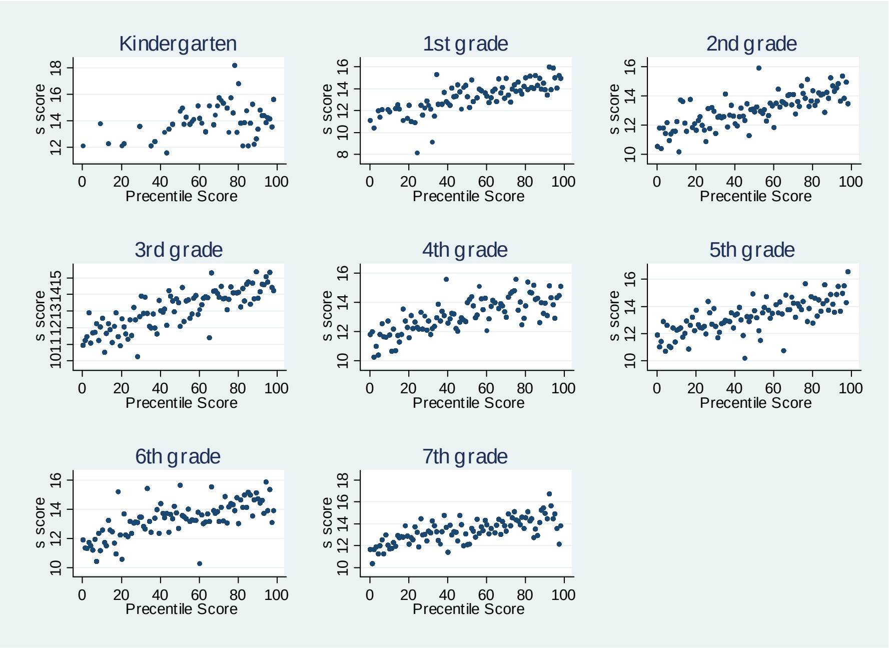

Figure 3: PIAT:RC Transformation

Table 1: Descriptive Statistics - Full Sample

Pre Age-5 PPVT			
Test Gap	1.15		
	(0.05)		
Observations	3657		
	(1)	(2)	(3)
	Math	Read-RR	Read-RC
Kindergarten			
Test Gap	0.65	0.19	0.19
	(0.06)	(0.06)	(0.13)
Observations	2877	2835	1221
First Grade			
Test Gap	0.66	0.42	0.39
	(0.05)	(0.05)	(0.06)
Observations	2893	2888	2478
Second Grade			
Test Gap	0.74	0.60	0.61
	(0.05)	(0.05)	(0.05)
Observations	2858	2885	2781
Third Grade			
Test Gap	0.73	0.62	0.67
	(0.04)	(0.04)	(0.04)
Observations	2889	2861	2811
Fourth Grade			
Test Gap	0.79	0.65	0.66
	(0.04)	(0.04)	(0.05)
Observations	2864	2729	2702
Fifth Grade			
Test Gap	0.71	0.57	0.61
	(0.04)	(0.04)	(0.04)
Observations	2734	2785	2755
Sixth Grade			
Test Gap	0.81	0.64	0.72
	(0.04)	(0.04)	(0.04)
Observations	2590	2594	2572
Seventh Grade			
Test Gap	0.74	0.59	0.69
	(0.04)	(0.05)	(0.04)
Observations	2475	2477	2466

SOURCE: Children of the National Longitudinal Survey of Youth. Test gaps are difference between average white and average black perecentile score measured in standard deviations. Custom weights are used so that each test-grade sample is nationally representative.

Table 2: Descriptive Statistics - Over 22 Sample

Pre Age-5 PPVT			
Test Gap		1.23	
		(0.08)	
Observations		1866	
	(1)	(2)	(3)
	Math	Read-RR	Read-RC
Kindergarten			
Test Gap	0.63	0.10	0.06
	(0.07)	(0.08)	(0.16)
Education Gap	0.86	0.85	1.13
	(0.13)	(0.13)	(0.19)
Observations	1480	1446	661
First Grade			
Test Gap	0.67	0.40	0.40
	(0.06)	(0.06)	(0.08)
Education Gap	0.76	0.78	0.83
	(0.11)	(0.11)	(0.13)
Observations	1544	1536	1281
Second Grade			
Test Gap	0.69	0.52	0.52
	(0.06)	(0.06)	(0.06)
Education Gap	0.76	0.72	0.70
	(0.12)	(0.12)	(0.12)
Observations	1638	1633	1572
Third Grade			
Test Gap	0.67	0.57	0.61
	(0.06)	(0.06)	(0.06)
Education Gap	0.85	0.86	0.86
	(0.12)	(0.12)	(0.12)
Observations	1587	1580	1552
Fourth Grade			
Test Gap	0.72	0.58	0.60
	(0.05)	(0.06)	(0.06)
Education Gap	0.77	0.78	0.76
	(0.11)	(0.11)	(0.11)
Observations	1612	1580	1587
Fifth Grade			
Test Gap	0.68	0.52	0.60
	(0.06)	(0.06)	(0.06)
Education Gap	0.84	0.84	0.82
	(0.11)	(0.12)	(0.12)
Observations	1562	1558	1539
Sixth Grade			
Test Gap	0.74	0.56	0.63
	(0.06)	(0.06)	(0.06)
Education Gap	0.85	0.85	0.85
	(0.12)	(0.12)	(0.12)
Observations	1447	1446	1434
Seventh Grade			
Test Gap	0.71	0.57	0.67
	(0.06)	(0.06)	(0.05)
Education Gap	0.67	0.69	0.69
	(0.11)	(0.11)	(0.11)
Observations	1453	1450	1444

Table 3: Raw Difference in Expected Grade Completion conditional on Test Score

	(1) Math	(2) Read-RR	(3) Read-RC
Pre-Age 5 PPVT		0.88	
		[1.12, 0.65]	
Kindergarten	0.55	0.20	0.26
	[0.33, 0.72]	[0.08, 0.36]	[-0.02, 0.52]
First Grade	0.50	0.35	0.32
	[0.37, 0.66]	[0.19, 0.46]	[0.17, 0.48]
Second Grade	0.72	0.58	0.48
	[0.56, 0.96]	[0.36, 0.77]	[0.26, 0.60]
Third Grade	0.67	0.60	0.61
	[0.52, 0.84]	[0.49, 0.78]	[0.46, 0.75]
Fourth Grade	0.70	0.56	0.58
	[0.57, 0.88]	[0.40, 0.71]	[0.44, 0.77]
Fifth Grade	0.69	0.47	0.51
	[0.54, 0.85]	[0.30, 0.61]	[0.33, 0.63]
Sixth Grade	0.70	0.58	0.60
	[0.55, 0.90]	[0.41, 0.79]	[0.40, 0.76]
Seventh Grade	0.71	0.54	0.57
	[0.54, 0.85]	[0.38, 0.70]	[0.44, 0.75]

Difference between average white and average black predicted education conditional on test score for each grade-test combination. Bootstrapped 95 percent confidence intervals in brackets. Conditional predicted education computed for those who are observed over age 22 and applied to the full sample. All results are weighted to be nationally representative.

Table 4: Measurement Error Adjusted Difference in Ability in Units of Predicted Education

	(1) Math	(2) Read-RR	(3) Read-RC
Kindergarten	1.24	0.64	1.32
	[0.65, 2.07]	[0.17, 1.68]	[-3.04, 7.41]
First Grade	1.01	0.88	0.64
	[0.57, 1.54]	[0.38, 1.40]	[0.25, 1.15]
Second Grade	1.05	0.81	0.91
	[0.50, 1.52]	[0.35, 1.37]	[0.43, 1.48]
Third Grade	1.02	0.65	0.69
	[0.47, 1.55]	[0.41, 0.96]	[0.23, 1.09]
Fourth Grade	1.05	0.57	0.71
	[0.68, 1.56]	[0.29, 0.78]	[0.12, 1.06]
Fifth Grade	0.81	0.60	0.67
	[0.52, 1.08]	[0.32, 0.75]	[0.36, 0.87]
Sixth Grade	0.91	0.74	0.81
	[0.62, 1.18]	[0.50, 1.07]	[0.48, 1.06]
Seventh Grade	0.89	0.68	0.72
	[0.54, 1.12]	[0.43, 0.90]	[0.37, 1.19]

Difference between average white and average black predicted education conditional on test score for each grade-test combination corrected for measurement error by instrumental variables. Bootstrapped 95 percent confidence intervals in brackets. Conditional predicted education computed for those who are observed over age 22 and applied to the full sample. All kindergarten and first grade tests, and the second grade Read-RC use predicted education conditional on test score for the PPVT as an instrument, while the remaining tests use that measure lagged two grades. All results are weighted to be nationally representative.

Table 5: Difference in Ability in Education-Predicted Log Earnings

	(1) Math	(2) Read-RR	(3) Read-RC
Kindergarten	0.17	0.09	0.15
	[0.07, 0.40]	[0.00, 0.41]	[-0.36, 0.94]
First Grade	0.12	0.11	0.08
	[0.06, 0.17]	[0.04, 0.19]	[0.03, 0.14]
Second Grade	0.13	0.12	0.12
	[0.06, 0.20]	[0.06, 0.23]	[0.05, 0.23]
Third Grade	0.12	0.09	0.09
	[0.05, 0.18]	[0.05, 0.13]	[0.03, 0.16]
Fourth Grade	0.13	0.07	0.08
	[0.07, 0.20]	[0.02, 0.10]	[-0.01, 0.13]
Fifth Grade	0.11	0.08	0.09
	[0.06, 0.15]	[0.04, 0.10]	[0.05, 0.12]
Sixth Grade	0.11	0.09	0.10
	[0.07, 0.16]	[0.06, 0.14]	[0.05, 0.13]
Seventh Grade	0.12	0.10	0.10
	[0.07 0.15]	[0.05, 0.13]	[0.05, 0.17]

Difference between average white and average black mean log-earnings of predicted education conditional on test score for each grade-test combination corrected for measurement error by instrumental variables. Bootstrapped 95 percent confidence intervals in brackets. Conditional predicted education computed for whites who are observed over age 22 and applied to the full sample. All kindergarten and first grade tests, and the second grade Read-RC use log-earnings predicted education conditional on test score for the PPVT as an instrument, while the remaining tests use that measure lagged two grades. All results are weighted to be nationally representative.

Table 6: Measurement Error Adjusted Difference in Ability in Units of Predicted White Education

	(1) Math	(2) Read-RR	(3) Read-RC
Kindergarten	1.14	0.62	1.26
	[0.58, 2.17]	[0.12, 2.27]	[-4.23, 8.81]
First Grade	1.01	0.82	0.74
	[0.56, 1.55]	[0.31, 1.32]	[0.28, 1.26]
Second Grade	1.07	0.87	0.79
	[0.57, 1.57]	[0.29, 1.53]	[0.35, 1.63]
Third Grade	0.97	0.64	0.66
	[0.52, 1.59]	[0.40, 0.99]	[0.24, 1.15]
Fourth Grade	1.12	0.53	0.72
	[0.69, 1.58]	[0.28, 0.76]	[0.02, 1.04]
Fifth Grade	0.79	0.55	0.63
	[0.51, 1.10]	[0.29, 0.75]	[0.33, 0.82]
Sixth Grade	0.84	0.74	0.75
	[0.54, 1.12]	[0.48, 1.01]	[0.43, 1.02]
Seventh Grade	0.87	0.66	0.68
	[0.51, 1.13]	[0.36, 0.93]	[0.38, 1.20]

Difference between average white and average black predicted education for whites conditional on test score for each grade-test combination corrected for measurement error by instrumental variables. Bootstrapped 95 percent confidence intervals in brackets. Conditional predicted education computed for whites who are observed over age 22 and applied to the full sample. All kindergarten and first grade tests, and the second grade Read-RC use predicted education conditional on test score for the PPVT as an instrument, while the remaining tests use that measure lagged two grades. All results are weighted to be nationally representative.

Table 7: Difference in Ability in Education-Predicted White Log Income

	(1) Math	(2) Read-RR	(3) Read-RC
Kindergarten	0.16	0.09	0.14
	[0.06, 0.41]	[-0.11, 0.47]	[-0.35, 0.81]
First Grade	0.12	0.11	0.09
	[0.05, 0.19]	[0.03, 0.20]	[0.03, 0.17]
Second Grade	0.13	0.14	0.10
	[0.06, 0.20]	[0.06, 0.27]	[0.03, 0.22]
Third Grade	0.12	0.09	0.09
	[0.04, 0.21]	[0.06, 0.13]	[0.02, 0.17]
Fourth Grade	0.13	0.06	0.09
	[0.07, 0.20]	[0.02, 0.10]	[-0.02, 0.14]
Fifth Grade	0.11	0.08	0.09
	[0.07, 0.15]	[0.04, 0.10]	[0.04, 0.12]
Sixth Grade	0.11	0.09	0.09
	[0.06, 0.15]	[0.05, 0.13]	[0.05, 0.13]
Seventh Grade	0.11	0.09	0.10
	[0.06, 0.15]	[0.05, 0.13]	[0.04, 0.17]

Difference between average white and average black mean log-earnings of predicted education for whites conditional on test score for each grade-test combination corrected for measurement error by instrumental variables. Bootstrapped 95 percent confidence intervals in brackets. Conditional predicted education computed for whites who are observed over age 22 and applied to the full sample. All kindergarten and first grade tests, and the second grade Read-RC use log-earnings predicted education conditional on test score for the PPVT as an instrument, while the remaining tests use that measure lagged two grades. All results are weighted to be nationally representative.

Table 8: Difference in Predicted Education Using All Tests

	(1) Unadjusted	(2) IV Adjusted
Kindergarten	0.41	0.82
	[0.23, 0.55]	[0.43, 1.12]
First Grade	0.49	0.93
	[0.33, 0.65]	[0.64, 1.25]
Second Grade	0.73	0.98
	[0.56, 0.88]	[0.64, 1.25]
Third Grade	0.75	0.88
	[0.64, 0.91]	[0.70, 1.15]
Fourth Grade	0.79	0.93
	[0.66, 0.98]	[0.73, 1.19]
Fifth Grade	0.71	0.79
	[0.52, 0.82]	[0.54, 0.90]
Sixth Grade	0.80	0.97
	[0.69, 1.01]	[0.77, 1.22]
Seventh Grade	0.74	0.85
	[0.56, 0.88]	[0.62, 1.08]

Difference between average white and average black predicted education conditional on all test scores for each grade-test combination. Bootstrapped 95 percent confidence intervals in brackets. Column 2 estimates are corrected for measurement error by instrumental variables. Conditional predicted education computed for those who are observed over age 22 using a multivariate kernal regression and applied to the full sample. Kindergarten and first grade use the predicted education conditional on test score for the PPVT as an instrument, while the remaining grades use that measure lagged two grades. All results are weighted to be nationally representative.

Table 9: Difference in Education-Predicted Log Income Using All Tests

	(1) Unadjusted	(2) IV Adjusted
Kindergarten	0.05	0.11
	[0.00, 0.08]	[0.01, 0.20]
First Grade	0.07	0.12
	[0.03, 0.11]	[0.05, 0.21]
Second Grade	0.09	0.12
	[0.06, 0.12]	[0.05, 0.21]
Third Grade	0.10	0.12
	[0.07, 0.13]	[0.08, 0.18]
Fourth Grade	0.10	0.11
	[0.07, 0.14]	[0.04, 0.17]
Fifth Grade	0.09	0.11
	[0.04, 0.12]	[0.05, 0.14]
Sixth Grade	0.10	0.12
	[0.07, 0.14]	[0.07, 0.17]
Seventh Grade	0.10	0.12
	[0.06, 0.13]	[0.07, 0.18]

Difference between average white and average black mean log-earnings of predicted education conditional on all test scores for each grade-test combination. Bootstrapped 95 percent confidence intervals in brackets. Column 2 estimates are corrected for measurement error by instrumental variables. Conditional predicted education computed for those who are observed over age 22 using a multivariate kernal regression and applied to the full sample. Kindergarten and first grade use the predicted education conditional on test score for the PPVT as an instrument, while the remaining grades use that measure lagged two grades. All results are weighted to be nationally representative.

Table 10: Conditional Difference in Ability in Units of Predicted Education

	(1) Math	(2) Read-RR	(3) Read-RC
Kindergarten	0.61	-0.04	0.55
	[-0.12, 1.52]	[-0.86, 1.21]	[-5.41, 5.49]
First Grade	0.38	-0.19	0.01
	[0.34, 0.83]	[-1.07, 0.38]	[-0.66, 0.41]
Second Grade	0.53	0.11	0.06
	[0.02, 1.11]	[-0.53, 0.53]	[-0.79, 0.55]
Third Grade	0.63	0.22	0.34
	[0.13, 1.23]	[-0.06, 0.60]	[0.01, 0.73]
Fourth Grade	0.39	0.22	-0.02
	[-0.00, 0.98]	[-0.13, 0.49]	[-0.46, 0.43]
Fifth Grade	0.50	0.31	0.69
	[0.17, 1.03]	[-0.12, 0.74]	[0.30, 1.25]
Sixth Grade	0.26	0.18	0.26
	[-0.20, 0.73]	[-0.26, 0.61]	[-0.21, 0.76]
Seventh Grade	0.48	0.08	0.26
	[0.07, 0.83]	[-0.50, 0.36]	[-0.14, 0.71]

Opposite of coefficient on black indicator in regression on predicted education conditional on test score for each grade-test combination corrected for measurement error by instrumental variables. Boostrapped 95 percent confidence intervals in brackets. Each regression includes controls for mother's education and age at first birth, child's birthweight, and household conditions at age 2 including log family income, log mother's hours worked, books, frequency of mother reading to child, mother's philosophy on children's learning, amount of toys in the household, whether the child sees the dad daily, and frequency of eating dinner with both parents Conditional predicted education computed for those who are observed over age 22 and applied to the full sample. All kindergarten and first grade tests, and the second grade Read-RC use predicted education conditional on test score for the PPVT as an instrument, while the remaining tests use that measure lagged two grades. All results are weighted to be nationally representative..

Table 11: Conditional Difference in Ability in Education-Predicted Log Income

	(1) Math	(2) Read-RR	(3) Read-RC
Kindergarten	0.09	-0.02	0.00
	[-0.01, 0.27]	[-0.17, 0.16]	[-0.88, 0.89]
First Grade	0.04	-0.02	-0.00
	[-0.06, 0.10]	[-0.16, 0.06]	[-0.10, 0.06]
Second Grade	0.06	0.02	0.01
	[-0.01, 0.14]	[-0.08, 0.09]	[-0.10, 0.09]
Third Grade	0.07	0.03	0.04
	[0.00, 0.15]	[-0.01, 0.08]	[0.00, 0.10]
Fourth Grade	0.04	0.03	0.00
	[-0.01, 0.12]	[-0.01, 0.07]	[-0.05, 0.06]
Fifth Grade	0.07	0.04	0.09
	[0.02, 0.14]	[-0.02, 0.07]	[0.04, 0.17]
Sixth Grade	0.04	0.03	0.04
	[-0.02, 0.10]	[-0.03, 0.09]	[-0.02, 0.11]
Seventh Grade	0.06	0.01	0.04
	[-0.00, 0.11]	[-0.08, 0.06]	[-0.02, 0.10]

Opposite of coefficient on black indicator in regression on log-earnings of predicted education conditional on test score for each grade-test combination corrected for measurement error by instrumental variables. Bootstrapped 95 percent confidence intervals in brackets. Each regression includes controls for mother's education and age at first birth, child's birthweight, and household conditions at age 2 including log family income, log mother's hours worked, books, frequency of mother reading to child, mother's philosophy on children's learning, amount of toys in the household, whether the child sees the dad daily, and frequency of eating dinner with both parents Conditional predicted education computed for those who are observed over age 22 and applied to the full sample. All kindergarten and first grade tests, and the second grade Read-RC use log-earnings of predicted education conditional on test score for the PPVT as an instrument, while the remaining tests use that measure lagged two grades. All results are weighted to be nationally representative..

Table A: Test Score Covariance Matrix

	PPVT	Grade K	Grade 1	Grade 2	Grade 3	Grade 4	Grade 5	Grade 6	Grade 7
PPVT	0.81								
MATH									
Grade K	0.34	1.04							
Grade 1	0.32		1.02						
Grade 2	0.39	0.52		1.31					
Grade 3	0.35		0.52		1.23				
Grade 4	0.34	0.45		0.66		1.20			
Grade 5	0.37		0.51		0.71		1.34		
Grade 6	0.40	0.46		0.64		0.66		1.29	
Grade 7	0.36		0.50		0.65		0.74		1.26
READING RECOGNITION									
Grade K	0.21	1.07							
Grade 1	0.28		1.09						
Grade 2	0.33	0.40		1.11					
Grade 3	0.39		0.67		1.23				
Grade 4	0.30	0.37		0.66		1.23			
Grade 5	0.35		0.61		0.77		1.50		
Grade 6	0.34	0.37		0.70		0.69		1.39	
Grade 7	0.35		0.61		0.77		0.81		1.43
READING COMPREHENSION									
Grade K	0.13	0.78							
Grade 1	0.29		1.05						
Grade 2	0.35	0.19		1.07					
Grade 3	0.37		0.49		1.24				
Grade 4	0.29	0.12		0.52		1.16			
Grade 5	0.32		0.43		0.56		1.14		
Grade 6	0.41	0.20		0.54		0.55		1.28	
Grade 7	0.39		0.37		0.56		0.54		1.16

Covariances are calculated using all available observations for each individual cell and are unweighted. Covariances of tests taken 1, 3, 5, or 7 years apart are not shown because the sample is surveyed every two years.