American Economic Journal: Applied Economics 2021, 13(1): 105–132
https://doi.org/10.1257/app.20180732

The Black-White Gap in Noncognitive Skills among Elementary School Children†

By Todd Elder and Yuqing Zhou*

Using two nationally representative datasets, we find large differences between Black and White children in teacher-reported measures of noncognitive skills. We show that teacher reports understate true Black-White skill gaps because of reference bias: teachers appear to rate children relative to others in the same school, and Black students have lower-skilled classmates on average than do White students. We pursue three approaches to addressing these reference biases. Each approach nearly doubles the estimated Black-White gaps in noncognitive skills, to roughly 0.9 standard deviations in third grade. (*JEL* I21, I26, J13, J15, J24)

Racial disparities in achievement and attainment are stubbornly persistent features of the US educational system. An extensive literature spanning several disciplines has studied outcomes such as test scores, graduation rates, and college attendance, finding that Black-White gaps emerge as early as age two (Scott and Sinclair 1997, Fryer and Levitt 2013). Moreover, Black-White achievement gaps widen sharply in the early years of schooling beyond what would be predicted based on differences in socioeconomic status (SES) and other observable characteristics (Jencks and Phillips 1998, Fryer and Levitt 2004, 2006).

As the literature on Black-White differences in educational outcomes has evolved, a separate but related literature has highlighted the importance of noncognitive skills in shaping adult outcomes. Heckman, Stixrud, and Urzua (2006) argued that noncognitive skills are important predictors of teenage pregnancy, tobacco and marijuana use, and participation in criminal activities.[1] Similarly, recent studies have documented the effects of noncognitive skills on labor market outcomes and a host of measures of social performance (Heckman and Rubinstein 2001; Heckman, Stixrud, and Urzua 2006; Flossmann, Piatek, and Wichert 2007; Borghans et al. 2008; Agan

*Elder: Michigan State University, 486 West Circle Drive, East Lansing, MI 48824 (email: telder@msu.edu); Zhou: Anderson School of Management, University of California at Los Angeles, 110 Westwood Plaza, Los Angeles, CA 90095 (email: yuqing.zhou.1@anderson.ucla.edu). David Deming was coeditor for this article. We thank seminar participants at Michigan State University, the University of Michigan, and the 2018 ASSA annual meetings for helpful comments and suggestions. Jeff Biddle, Brantly Callaway, Steven Haider, and Scott Imberman provided helpful comments on earlier drafts. All errors are our own.

†Go to https://doi.org/10.1257/app.20180732 to visit the article page for additional materials and author disclosure statement(s) or to comment in the online discussion forum.

[1] Heckman (2008) and Cunha, Heckman, and Schennach (2010) emphasize the complementary nature of cognitive and noncognitive skills by developing and estimating models of the joint evolution of cognitive and noncognitive skills over the life cycle.

2011; Heckman and Kautz 2012; Heckman, Pinto, and Savelyev 2013; Segal 2013; Piatek and Pinger 2016).

Despite the surge in interest in noncognitive skills within economics, little attention has been paid to documenting how these skills vary across demographic groups. In one recent exception, Bertrand and Pan (2013) finds that boys' propensity to engage in disruptive behavior stems partly from gender differences in the returns to parental inputs. Goldammer (2012) finds that noncognitive skills can explain a significant share of the Asian advantage in adult economic outcomes relative to Whites, Blacks, and Hispanics. Fryer, Levitt, and List (2015) analyzes the effects of a randomized field experiment on noncognitive test scores of each of these groups, and Urzúa (2008) explores whether racial differences in cognitive and noncognitive skills can explain racial disparities in incarceration rates.

In this paper, we use data from the 2010–2011 cohort of the Early Childhood Longitudinal Study: Kindergarten Cohort (ECLS-K:2011) to estimate Black-White gaps in noncognitive skills. The ECLS-K:2011 is especially well suited to studying noncognitive skills because it includes detailed teacher assessments of multiple aspects of each child's personality. We also use the original ECLS-K cohort of kindergarteners in the 1998–1999 school year (ECLS-K:1999) in order to track how racial gaps in noncognitive skills have evolved over time.

Our analyses produce three substantive findings. First, we find large, statistically significant Black-White gaps in several measures of noncognitive skills. We focus on measures that capture externalizing problem behaviors, self-control, internalizing problem behaviors, interpersonal skills, and "approaches to learning" (which encompasses attentiveness, task persistence, and motivation). With the exception of the "internalizing problem behavior" scale, racial gaps in these measures are large throughout the elementary school years. After conditioning on detailed controls for home and school environments, the gaps at the end of third grade are roughly one-half to three-fourths as large as the analogous test score gaps studied by Fryer and Levitt (2006), which uses a similar set of controls.

Second, the estimated gaps are remarkably stable across birth cohorts. For example, the unconditional Black-White gap in the measure of approaches to learning is 0.524 standard deviations among third graders in the ECLS-K:2011 cohort, compared to 0.545 standard deviations for the ECLS-K:1999 cohort. This finding mirrors those shown in the literature on test scores, with studies such as Neal (2006) and Magnuson and Waldfogel (2008) concluding that the Black-White test score gap has remained roughly constant since the early 1990s.

Third, and perhaps most troubling, we present evidence that our baseline estimates substantially understate true Black-White differences in noncognitive skills. We argue that this understatement arises due to the nature of teacher-provided student evaluations. The teacher questionnaires include statements such as "The student persists in completing tasks," to which teachers could choose one of four responses: "never," "sometimes," "often," or "very often." Because there is no natural, objective scale for delimiting these four choices, teachers might use relative comparisons across students to guide their responses. For example, in a classroom of driven, goal-oriented children, a teacher might respond that a particular student persists in completing tasks "sometimes," while an identically behaved student in

a classroom with less ambitious classmates may instead receive a rating of "often" (or "very often").[2]

In order to assess whether relative comparisons across students influence teacher reports of noncognitive skills, we first compare between- and within-school variation in these measures to analogous variation in arguably more objective metrics, such as achievement test scores. For example, when normed by their respective within-school variances, the between-school variance in the "approaches to learning" index is only one-fourth as large as that of third-grade math scores. We find similar patterns across every teacher-reported measure we consider, suggesting that students in relatively underperforming schools (or classrooms) receive systematically higher noncognitive skill ratings from their teachers than do otherwise identical students who attend high-performing schools. This reference bias would compress the unconditional Black-White gap in noncognitive skills if Black students disproportionately attend low-performing schools.

In order to address the effects of reference bias on the Black-White gaps in noncognitive skills, we propose three different approaches that treat those skills as latent variables. The intuition underlying each of these approaches is that the degree of sorting across schools on objective measures—such as test scores or characteristics of the home environment—is informative about the degree of sorting on the latent noncognitive skills themselves. In the first approach, we use between-school variation in objective measures to anchor the between-school variation in noncognitive skills, thereby generating estimates of the latent skill distributions. In the second, we estimate reference biases by comparing teachers' ratings of cognitive skills to item-response theory (IRT) test scores.[3] The school-level differences between these measures identify reference biases under the assumption that average test scores in a school are objective measures of the average cognitive skills of students in those schools. Finally, we generate predicted values of noncognitive skills from models that use only within-school variation, using the predictions as estimates of noncognitive skills that are free from reference bias.

For all measures of noncognitive skills, each of our approaches increases the implied Black-White gaps substantially. For example, the baseline gap in the third-grade "approaches to learning" index in ECLS-K:2011 is 0.524 standard deviations, but our methods imply gaps ranging from 0.846 to 0.973 standard deviations—roughly as large as the gaps in reading and math achievement test scores.

Finally, our results imply that previous estimates of the associations between noncognitive skills and adult outcomes are potentially misleading due to the role of

[2] The ECLS-K also includes parental ratings of measures of noncognitive skills, but both DiPrete and Jennings (2012) and Elder (2010) argue that these measures are substantially less useful than are teacher ratings. First, the parental ratings are relatively unstable across survey wave, with first-order autocorrelations of roughly 0.2, suggesting that much of the observed variation reflects noise. More generally, parental ratings are weakly correlated with observable determinants of outcomes, especially in comparison to teacher ratings. For example, parental ratings of a child's math ability are only weakly associated with variables that predict achievement test scores, such as parental education and income. As Elder (2010) argues, some degree of subjectivity in ratings appears to be unavoidable, regardless of the rater, but teacher ratings are at least based on a well-defined frame of reference (peers within the same classrooms), making them arguably preferable to parental ratings.

[3] In the ECLS-K testing batteries, children answer different test questions depending on the number of correct answers they have provided to that point on the test, so we prefer IRT scores to a simpler "overall number correct" test score measure.

reference bias. Because noncognitive and cognitive skills are positively correlated in many cases, measurement error in noncognitive skills may lead researchers to misattribute some of the impacts of noncognitive skills to cognitive skills.

I. Data and Descriptive Findings

Our analysis uses data from two cohorts of the ECLS-K, administered by the National Center for Education Statistics (NCES). The original ECLS-K cohort is a longitudinal survey that followed a nationally representative sample of roughly 18,600 children who entered kindergarten in the 1998–1999 school year. NCES resampled children in the spring of 1999, fall of 1999, and springs of 2000, 2002, 2004, and 2007. NCES also interviewed parents and teachers in each survey wave. Following NCES's convention, we refer to this survey as ECLS-K:1999.

The second cohort of the ECLS-K, denoted by NCES as ECLS-K:2011, includes 18,200 children representing the national population of kindergarten-age children in the 2010–2011 school year. NCES designed the structure of the ECLS-K:2011 to be nearly identical to the ECLS-K:1999. NCES first surveyed children in the fall of 2010, with follow-up samples in spring 2011, fall 2011, spring 2012, fall 2012, spring 2013, and spring 2014.

Both ECLS-K cohorts include detailed information on family background, home environments, and cognitive skills, including reading, math, and science IRT scores. In addition, teachers provided assessments of students' mastery of specific skills in reading, math, and science. These assessments are measured on a five-point integer scale (from zero to four) known as the Academic Rating Scale (ARS), although the interpretation of the scales differs slightly across the two cohorts.[4]

The ECLS-K includes a comprehensive set of weights designed to make analysis samples nationally representative. We report unweighted estimates below, but our central findings are insensitive to whether or not we use sample weights in all cases.

A. *Noncognitive Skills Based on Teacher and Parent Reports*

Teachers in both ECLS-K cohorts rate individual students on scales from 1 ("never") to 4 ("very often") on 24 different dimensions intended to measure social, emotional, and cognitive development. NCES does not release data on each of these 24 items individually, instead aggregating them to 5 composite scales known as Social Rating Scales (SRS).[5] For example, the "externalizing problem behaviors" scale uses information about the frequency with which a child acts impulsively, interrupts ongoing activities, fights with other children, gets angry, and argues. The "approaches to learning" scale is based on information about a child's attentiveness, task persistence, eagerness to learn, learning independence, flexibility, and

[4] In the ECLS-K:2011, a rating of zero indicates "far below grade level" and four indicates "far above grade level," while in ECLS-K:1999, zero indicates "not yet demonstrated the skill, knowledge, or behavior" and four indicates "consistent and competent demonstration of the skill, knowledge, or behavior."

[5] The Social Rating Scales used by NCES are adaptations of the scales designed by Gresham and Elliot (1990). Because the scales are copyright protected, we cannot reproduce their precise wording here; we refer interested readers to Gresham and Elliott (1990).

organization. The third scale, "self-control," includes four items that measure a child's ability to control his or her behavior. The fourth scale, "interpersonal skills," uses five items that measure a child's ability to interact with others, and the last scale, "internalizing problem behaviors," includes four items that rate the presence of anxiety, sadness, loneliness, and low self-esteem. Online Appendix A provides detailed information about the creation of these variables.

The Social Rating Scales are widely used survey instruments for detecting social and behavioral problems (Gresham and Elliott 1990). As Bertrand and Pan (2013) and Neidell and Waldfogel (2011) argue, these scales are highly reliable measures of noncognitive skills, and they are arguably the most comprehensive assessments that are usable in large surveys such as the ECLS-K.[6] Bertrand and Pan (2013) shows that externalizing problem behaviors are closely linked to school suspension and illegal activities. Elder (2010) shows that among young children, the self-control scale is closely related to ADHD diagnoses and treatment, which are both associated with later educational outcomes. Cornwell, Mustard, and Van Parys (2013) finds that the "approaches to learning" scale strongly predicts students' current and future performance in school, even conditional on achievement test scores. In addition, Duckworth and Quinn (2009) argues that task persistence (grit), which is a key component of the "approaches to learning" scale, is a more powerful predictor of academic success than is IQ. Deming (2017) presents evidence on the importance of interpersonal skills for labor market outcomes, showing that the return to such skills increased sharply between the 1980s and the 2000s.

In our empirical specifications below, we restrict our samples to children who have valid teacher reports of the relevant noncognitive skill under study. For example, the third-grade ECLS:K-2011 survey includes 7,496 Black and White children. In models in which the "approaches to learning" scale is the dependent variable, we include the 7,048 children who have nonmission information for that scale. The analogous sample sizes for the externalizing problem behaviors, self-control, interpersonal, and internalizing problem behaviors scales are 7,039, 6,966, 6,989, and 7,021 children, respectively.[7] We do not exclude cases due to missing values of covariates; instead, we create missing-variable indicators and set missing values equal to the variable's respective sample mean.

B. *Descriptive Statistics*

Table 1A presents descriptive statistics for the ECLS-K:2011 sample. We standardize the noncognitive skills scales to have zero mean and unit standard deviation in the full sample, and we include all children whose race is listed as either "White, non-Hispanic" or "Black/African-American, non-Hispanic." The table also

[6] Specifically, Neidell and Waldfogel (2010) states that the ECLS-K noncognitive measures appear to have relatively high "validity as assessed by test-retest reliability, internal consistency, interrater reliability, and correlations with other, more advanced behavioral constructs (Elliott et al. 1988) and are considered the most comprehensive assessment that can be widely administered in large surveys such as the ECLS-K (Demaray et al. 1995)."

[7] The kindergarten survey includes 10,291 children, of whom 9,700, 9,667, 9,624, 9,624, and 9,653 children have nonmissing values of the approaches to learning, externalizing problem behaviors, self-control, interpersonal, and internalizing problem behaviors scales, respectively.

Table 1A—Summary Statistics by Race of Noncognitive Skills and Family and Student Characteristics: ECLS-K:2011

Variable	White (Observations = 8,489)	Black (Observations = 2,396)
Approaches to learning		
Kindergarten	0.063	−0.268
First grade	0.058	−0.358
Second grade	0.049	−0.382
Third grade	0.062	−0.462
Externalizing problem behaviors		
Kindergarten	0.034	−0.364
First grade	0.013	−0.419
Second grade	0.012	−0.462
Third grade	0.021	−0.500
Self-control		
Kindergarten	0.073	−0.325
First grade	0.069	−0.383
Second grade	0.052	−0.440
Third grade	0.054	−0.530
Interpersonal skills		
Kindergarten	0.071	−0.231
First grade	0.072	−0.317
Second grade	0.070	−0.363
Third grade	0.061	−0.425
Internalizing problem behaviors		
Kindergarten	0.000	−0.086
First grade	−0.015	−0.129
Second grade	−0.021	−0.137
Third grade	−0.013	−0.120
Family background characteristics		
Mother's education	14.507	13.295
	(2.263)	(2.190)
Father's education	14.362	13.304
	(2.408)	(2.310)
Parents married	0.713	0.267
	(0.452)	(0.442)
SES composite index	0.162	−0.573
	(0.737)	(0.703)

includes means and standard deviations of a subset of the family background and demographic characteristics that we include in the empirical analyses below, all measured in the first survey wave. Mother's and father's education is measured in years, "Parents married" is a binary variable capturing whether the child's parents were married, and the SES composite index is an NCES-created continuous measure. In our empirical models below, we also include the number of books the child has in the home, the child's birth weight in ounces, and a binary variable equal to one if the child lived with both biological parents.

We follow Fryer and Levitt (2004, 2006) in using this relatively parsimonious set of controls. As in Fryer and Levitt's (2004, 2006) analyses, our results below change little if we instead include a much more exhaustive set of controls, primarily because the kindergarten SES composite index is a powerful predictor of child outcomes. NCES created this index based on parental education, parental occupation, and

household income. For both cohorts, we standardize the index to have zero mean and unit standard deviation in our estimation samples.

As Table 1A shows, the averages of many of these variables differ markedly by race. White children outperform Black children in all five measures of noncognitive skills; we invert the scales of the externalizing and internalizing problem behavior measures so that higher scores represent "better" behavior. The differences are statistically significant in all cases. On average, White children perform better than Black children on the approaches to learning scale by 0.524 $(= 0.062 + 0.462)$ standard deviations in third grade. The third-grade gaps in the "externalizing problem behaviors," self-control, and interpersonal scales are similarly large. The Black-White gaps in the "internalizing problem behaviors" scales are smaller in all grades.

The remaining rows of the table show that Black children in the ECLS-K:2011 sample grow up in more disadvantaged households than do White children. For example, only 26.7 percent of Black children have married parents in kindergarten, compared to 71.3 percent among White children. Likewise, the Black-White difference in the SES composite index is 0.735 $(= 0.162 + 0.573)$ standard deviations.

Table 1B presents descriptive statistics for the ECLS-K:1999 sample. The central patterns are similar to those shown in Table 1A, in that in most cases, average noncognitive skills and home environments differ markedly by race. We turn next to assessing the roles of these differential home environments in explaining the gaps in noncognitive skills.

II. Baseline Estimates of Black-White Gaps in Noncognitive Skills

In order to measure Black-White gaps in noncognitive skills, we begin by estimating linear models,

$$y_{ij} = \gamma b_i + X'_{ij}\theta + \varepsilon_{ij}, \tag{1}$$

where i indexes children and j indexes schools. The vector X_{ij} denotes observed covariates, ε_{ij} denotes unobserved determinants of skills, and b_i is an indicator equaling one for Black students and zero otherwise. We limit our sample to non-Hispanic Black and White children, so the coefficient on b_i measures average Black-White conditional differences in the outcomes y_{ij}.

Table 2 presents OLS estimates of γ from specification (1). Columns 1 and 4 include estimates from models that include no controls, columns 2 and 5 add the home environment controls, and columns 3 and 6 add school indicators. All standard errors in the table are robust to heteroscedasticity and clustering within schools.

In panel A, the first three columns show estimates for the "approaches to learning" index in the ECLS-K:2011 cohort. In kindergarten, the mean difference is -0.331, and the estimate declines to -0.109 in column 2, implying that the large disparities in home environment shown in Tables 1A and 1B account for roughly two-thirds of the raw gap.[8] In contrast, column 3 shows that inclusion of the school

[8] We have also estimated specifications that include variables meant to capture parental time inputs, with little effect on the estimates. In these specifications, controls for parental time inputs include time spent reading to the

Table 1B—Summary Statistics by Race of Noncognitive Skills and Family and Student Characteristics: ECLS-K:1999

Variable	White (Observations = 9,824)	Black (Observations = 2,469)
Approaches to learning		
Kindergarten	0.108	−0.346
First grade	0.084	−0.357
Third grade	0.088	−0.458
Fifth grade	0.087	−0.451
Externalizing problem behaviors		
Kindergarten	0.066	−0.358
First grade	0.056	−0.383
Third grade	0.075	−0.552
Fifth grade	0.070	−0.554
Self-control		
Kindergarten	0.107	−0.412
First grade	0.090	−0.380
Third grade	0.088	−0.530
Fifth grade	0.082	−0.528
Interpersonal skills		
Kindergarten	0.065	−0.127
First grade	0.069	−0.164
Third grade	0.081	−0.415
Fifth grade	0.092	−0.195
Internalizing problem behaviors		
Kindergarten	−0.040	0.014
First grade	−0.020	−0.004
Third grade	0.012	−0.105
Fifth grade	−0.081	0.172
Family background characteristics		
Mother's education	13.929	12.883
	(1.932)	(1.724)
Father's education	14.025	13.096
	(2.138)	(1.864)
Parents married	0.775	0.302
	(0.417)	(0.459)
SES composite index	0.236	−0.357
	(0.736)	(0.755)

indicators increases the estimated gap to −0.242, which seemingly suggests that Black students systematically attend "better" schools—in terms of producing noncognitive skills—than do White students. Below we present an alternative explanation for why the point estimates are larger in absolute value in column 3 than in column 2: systematic differences across schools in what teacher-reported measures of noncognitive skills actually measure.[9] Specifically, reference biases may induce

child, telling stories, singing songs, helping the child create art, helping with chores, playing games, teaching nature or science, building something with the child, engaging in sports, visiting the library, going to a concert, visiting a museum, visiting a zoo, attending a sporting event, helping with homework, and helping children practice numbers.

[9] We also estimate alternative specifications in which we include indicators for teachers rather than for schools, and the resulting estimates are similar to those in column 3 in all cases. We return to this issue below in the context of the mechanisms that drive differences across schools in reported noncognitive skills.

TABLE 2—ESTIMATED RACIAL GAPS IN NONCOGNITIVE SKILLS, 1999 AND 2011 ECLS-K COHORTS

	ECLS-K:2011				ECLS-K:1999		
	(1)	(2)	(3)		(4)	(5)	(6)
Panel A. Approaches to learning							
Kindergarten	−0.331	−0.109	−0.242	Kindergarten	−0.454	−0.200	−0.242
	(0.026)	(0.028)	(0.040)		(0.025)	(0.035)	(0.050)
Grade 1	−0.416	−0.158	−0.284	Grade 1	−0.441	−0.178	−0.251
	(0.030)	(0.032)	(0.048)		(0.027)	(0.037)	(0.055)
Grade 2	−0.431	−0.129	−0.175	Grade 3	−0.545	−0.206	−0.350
	(0.032)	(0.034)	(0.053)		(0.034)	(0.047)	(0.073)
Grade 3	−0.524	−0.218	−0.274	Grade 5	−0.538	−0.182	−0.221
	(0.035)	(0.037)	(0.058)		(0.036)	(0.052)	(0.082)
Panel B. Externalizing problem behaviors							
Kindergarten	−0.398	−0.172	−0.260	Kindergarten	−0.424	−0.249	−0.270
	(0.027)	(0.029)	(0.042)		(0.027)	(0.035)	(0.050)
Grade 1	−0.432	−0.197	−0.274	Grade 1	−0.439	−0.222	−0.198
	(0.031)	(0.034)	(0.050)		(0.027)	(0.037)	(0.055)
Grade 2	−0.475	−0.231	−0.205	Grade 3	−0.626	−0.290	−0.354
	(0.033)	(0.036)	(0.056)		(0.034)	(0.047)	(0.072)
Grade 3	−0.521	−0.246	−0.201	Grade 5	−0.623	−0.309	−0.221
	(0.036)	(0.039)	(0.060)		(0.037)	(0.053)	(0.082)
Home environment controls		X	X			X	X
School fixed effects			X				X
Panel C. Self-control							
Kindergarten	−0.399	−0.197	−0.262	Kindergarten	−0.519	−0.330	−0.272
	(0.027)	(0.030)	(0.042)		(0.026)	(0.037)	(0.051)
Grade 1	−0.452	−0.223	−0.285	Grade 1	−0.470	−0.242	−0.241
	(0.031)	(0.034)	(0.050)		(0.027)	(0.039)	(0.056)
Grade 2	−0.492	−0.242	−0.162	Grade 3	−0.619	−0.320	−0.381
	(0.034)	(0.037)	(0.056)		(0.034)	(0.049)	(0.074)
Grade 3	−0.584	−0.284	−0.195	Grade 5	−0.610	−0.282	−0.239
	(0.036)	(0.039)	(0.061)		(0.038)	(0.054)	(0.084)
Panel D. Interpersonal skills							
Kindergarten	−0.301	−0.120	−0.274	Kindergarten	−0.192	−0.123	−0.086
	(0.027)	(0.029)	(0.041)		(0.023)	(0.035)	(0.049)
Grade 1	−0.389	−0.172	−0.262	Grade 1	−0.234	−0.173	−0.077
	(0.031)	(0.034)	(0.050)		(0.025)	(0.036)	(0.051)
Grade 2	−0.433	−0.185	−0.212	Grade 3	−0.496	−0.238	−0.349
	(0.034)	(0.036)	(0.056)		(0.035)	(0.049)	(0.075)
Grade 3	−0.486	−0.215	−0.197	Grade 5	−0.287	−0.199	0.044
	(0.036)	(0.039)	(0.060)		(0.034)	(0.051)	(0.078)
Home environment controls		X	X			X	X
School fixed effects			X				X
Panel E. Internalizing problem behaviors							
Kindergarten	−0.086	0.070	−0.024	Kindergarten	0.054	0.123	0.040
	(0.027)	(0.030)	(0.043)		(0.022)	(0.033)	(0.047)
Grade 1	−0.114	0.042	0.009	Grade 1	0.016	0.035	−0.035
	(0.031)	(0.035)	(0.051)		(0.026)	(0.036)	(0.051)
Grade 2	−0.116	0.066	0.001	Grade 3	−0.117	0.077	0.043
	(0.033)	(0.037)	(0.058)		(0.035)	(0.050)	(0.076)
Grade 3	−0.107	0.091	−0.006	Grade 5	0.252	0.204	0.030
	(0.036)	(0.040)	(0.063)		(0.035)	(0.049)	(0.074)
Home environment controls		X	X			X	X
School fixed effects			X				X

Note: Each entry in the table corresponds to an estimate from a separate regression of a measure of noncognitive skills, corresponding to equation (1) in the text.

teachers to understate students' noncognitive skills in high-achieving schools relative to students in lower-performing schools.

The estimates in panel A for grades 1, 2, and 3 in ECLS-K:2011 are similar to those from kindergarten, in that the inclusion of the home environment controls substantially reduces the estimated gaps in each grade, while the inclusion of school fixed effects increases those gaps. Although the raw gap widens monotonically between kindergarten and third grade, there are no clear temporal patterns in the estimates in column 3.

Similar patterns emerge in the ECLS-K:1999 cohort, as shown in columns 4–6. Again, the estimated gaps shrink with the inclusion of the home environment controls but widen with the inclusion of school fixed effects. The estimates in columns 3 and 6 are similar in magnitude when comparing the same grades.

Finally, panels B, C, D, and E show estimates from specification (1) based on the "externalizing problem behaviors," "self-control," "interpersonal skills," and "internalizing problem behaviors" scales as dependent variables. Most of the estimates are similar in magnitude to the analogous estimates in panel A. Unlike in panel A, including school fixed effects does not uniformly inflate the estimated gaps in panels B–E. Across all five measures, the magnitudes of the point estimates are broadly similar for the two ECLS-K cohorts, especially in kindergarten and first grade, implying that Black students made little measurable progress in catching up to White students between 1999 and 2011. Given this similarly, we focus primarily on the ECLS-K:2011 cohort hereafter.

In sum, the baseline estimates from Tables 1 and 2 show that Black elementary school students lag behind White students on several dimensions of noncognitive skills. By the end of third grade, the raw gaps in five noncognitive skills are roughly 0.5 to 0.6 standard deviations, but conditioning on controls for home and school environments reduces the estimated gaps by 45 to 66 percent. We turn next to assessing whether these estimates can be interpreted as the differences in noncognitive skills between Black and White children.

III. Teachers' Subjective Ratings of Cognitive and Noncognitive Skills

A. *Objective and Subjective Measures of Cognitive Skills*

The estimates in Table 2 are consistent with large Black-White differences in several dimensions of noncognitive skills. An important limitation of these findings, however, is that the measures of noncognitive skills available in the ECLS-K cohorts are arguably subjective, in that they are based on teacher-provided ratings. In contrast, achievement-test-based measures of cognitive skills available in the ECLS-K have not only been found to have high levels of reliability and content validity, but they are also regularly de-identified before grading, eliminating one potential source of subjectivity.[10] As a result, one might suspect that teachers' biases in favor of White students might account for at least some of the observed Black-White gaps.

[10] Recent research by Bond and Lang (2013, 2018), among others, has generated substantial debate about what achievement tests capture. Nonetheless, the achievement tests available in ECLS-K are among the most reliable

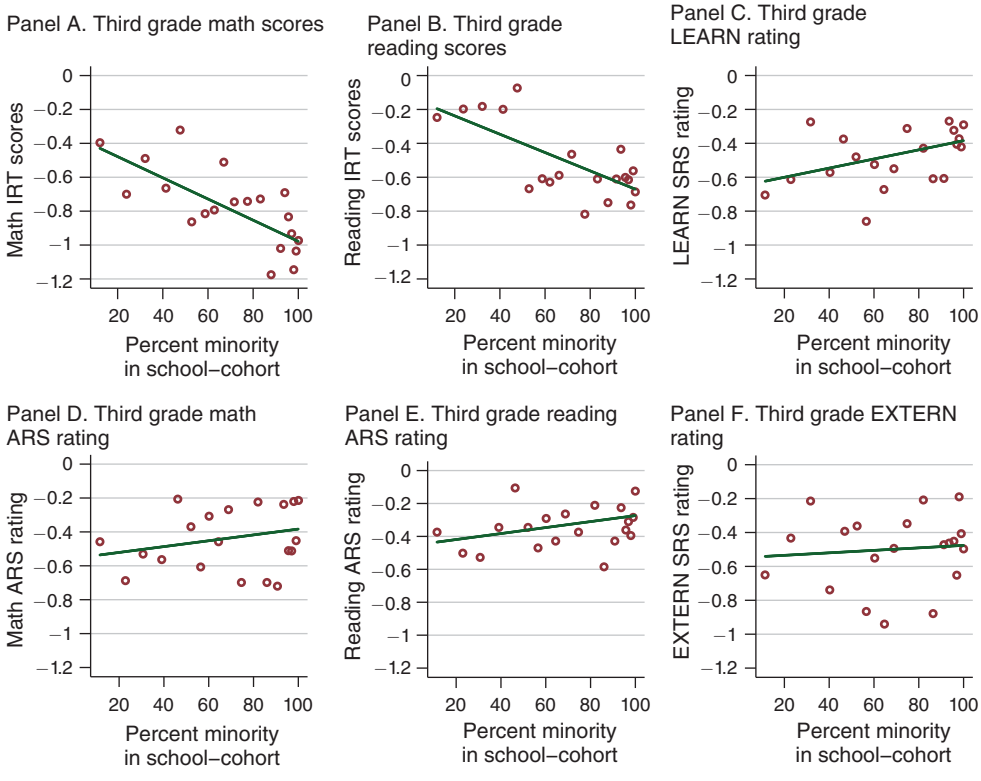

FIGURE 1. AVERAGE OBSERVED SKILLS BY SCHOOL RACIAL COMPOSITION AMONG THIRD-GRADE BLACK STUDENTS IN ECLS-K:2011

As noted above, one puzzle that emerges from Table 2 is why the estimated noncognitive skill gaps do not systematically decline after the inclusion of school fixed effects. Because of the close links between residential location and school attendance, including school indicators is roughly similar to including neighborhood indicators (Fryer and Levitt 2006), and previous research has found strong associations between skills and neighborhood characteristics. Moreover, in our own estimates based on the ECLS-K:2011, we find that the inclusion of school indicators substantially reduces Black-White gaps in achievement test scores, especially in grades 2 and 3; we include these results in online Appendix Table A1.

In order to understand how teachers form ratings of noncognitive skills, we first look to teachers' analogous ratings of cognitive skills. Figure 1 displays the association between school racial composition and six metrics of skills among third-grade Black students in ECLS-K:2011. We bin Black students into 20 equally sized categories and plot the average level of skills in each bin. For example, the top-left panel shows that average math IRT scores among Black students who attended schools with the smallest fraction of minority (Black or Hispanic) students were approximately 0.4 standard deviation units below the full-sample mean. Average scores are

of any survey-based test; NCES psychometric reports for the original ECLS-K:1999 suggest that both math and reading IRT tests have test-retest reliabilities of over 0.9 in all grades (NCES 2002).

strongly declining in the school minority share: in the bins with the highest minority shares in the school, the averages were roughly one standard deviation below the overall mean.

The bottom-left panel in the figure, labeled "Third-Grade Math ARS Rating," measures students' math skills as rated by teachers' responses to the Academic Rating Scales (ARS) surveys. The ARS are designed to reflect students' proficiency scores in a subject, and in principle they should capture similar variation as that measured by IRT scores. However, in sharp contrast to the top panel, there is a weak positive relationship between school racial composition and Black students' average ARS scores. Results for reading proficiency, shown in the middle panels, are nearly identical: IRT test scores among Black students are steeply decreasing in school minority shares, while ARS scores are slightly increasing.[11]

One interpretation of the patterns shown in the left and center panels of Figure 1 is that ARS ratings and IRT test scores measure different dimensions of skills. Table 3 presents evidence suggesting that this interpretation is likely to be incorrect. Each entry in the table corresponds to an estimate from a separate regression of a measure of third-grade cognitive skills (either standardized ARS ratings or standardized IRT test scores) on a measure of home environment, using the ECLS-K:2011 sample of White students only. All estimated models include school indicators.

The table shows that ARS ratings and IRT scores have nearly identical relationships with the home environment variables among White students in ECLS-K:2011. For example, an additional year of mother's education is associated with a 0.129 standard deviation increase in ARS math ratings and a 0.121 standard deviation increase in IRT math test scores, on average. The point estimates are similar in all cases, and we reject the equality of the two coefficients at the 10 percent level in only one case, for math skills' relationship with the number of books that the child has in the home.[12]

We also note an additional implication of the results in Table 3: the similarity of the estimates for ARS ratings and IRT scores implies that teachers do not appear to evaluate students relative to expectations within schools. For example, if teachers have lower expectations for students from disadvantaged backgrounds, we would expect that ARS ratings across the SES distribution would be compressed relative to IRT scores. We do not find evidence of such compression within schools, suggesting that teachers do not use performance relative to expectations when reporting ARS ratings.

[11] It is plausible that teacher-reported subjective assessments distort not only the location of student skills but also within-school variances in those skills. This scenario is especially likely if the within-school dispersion of skills differs across schools in a systematic way, such as with respect to school minority shares. In order to explore this possibility, we created a series of figures analogous to Figure 1 except that the x-axis measures within-school dispersion in skills rather than within-school averages. In practice, we find only limited evidence that the variance of the subjective measures differs within schools from the corresponding within-school variance of the objective measures. Nonetheless, such distortions may exist; to the extent that they do, it is unclear how they would bias measured Black-White gaps in noncognitive skills, as the resulting bias would depend on how the dispersion in skills is related to within-school Black-White gaps.

[12] We note that because the distribution of p-values is uniform over the $[0, 1]$ interval under a true null, the probability of finding at least one p-value below 0.1 when testing 14 true independent hypotheses is roughly 0.651.

TABLE 3—ASSOCIATIONS BETWEEN HOME ENVIRONMENT CONTROLS AND COGNITIVE SKILLS AMONG WHITE THIRD-GRADE STUDENTS, ECLS-K:2011

	Math skills			Reading skills		
	ARS ratings (1)	IRT test scores (2)	p-value for H_0: (1)=(2) (3)	ARS ratings (4)	IRT test scores (5)	p-value for H_0: (4)=(5) (6)
Mother's education	0.129	0.121	0.433	0.136	0.127	0.388
	(0.008)	(0.008)		(0.008)	(0.008)	
Father's education	0.124	0.105	0.116	0.127	0.117	0.381
	(0.008)	(0.008)		(0.008)	(0.008)	
Parents married	0.349	0.296	0.246	0.345	0.327	0.695
	(0.039)	(0.038)		(0.040)	(0.039)	
Two-parent household	0.425	0.387	0.467	0.426	0.397	0.557
	(0.045)	(0.043)		(0.046)	(0.044)	
Number of books child has	0.002	0.003	0.035	0.003	0.003	0.952
	(0.000)	(0.000)		(0.000)	(0.000)	
SES composite index	0.486	0.432	0.185	0.505	0.471	0.362
	(0.026)	(0.025)		(0.026)	(0.026)	
Child's birth weight (oz)	0.005	0.006	0.119	0.003	0.003	0.920
	(0.001)	(0.001)		(0.001)	(0.001)	

Notes: (i) Each entry in the table corresponds to an estimate from a separate regression of a measure of cognitive skills on a measure of home environment. (ii) All estimated models also control for school fixed effects. (iii) All standard errors are robust to heteroscedasticity and clustering within schools.

If ARS ratings and IRT test scores capture the same dimensions of skills, then what explains the patterns shown in Figure 1? One potential mechanism is that ARS ratings are driven primarily by comparisons across children in the same school. In ECLS-K:2011, ARS ratings are based on an integer scale from 0 to 4, where the ratings of 0, 1, 2, 3, and 4 represent "far below grade level," "below grade level," "at grade level," "above grade level," and "far above grade level," respectively. The definition of "at grade level," however, is potentially a function of the classroom's achievement distribution. In particular, teachers might implicitly define "at grade level" as the mean of the within-classroom distribution of proficiency. The relative insensitivity of mean ARS ratings to the racial distributions of schools shown in Figure 1—despite the strong association between IRT test scores and those racial distributions—is consistent with this sort of reference bias.

B. *Implications for Noncognitive Skills*

The preceding discussion compared two measures of cognitive skills, but it likely has implications for our measures of noncognitive skills, for which we only have relatively subjective measures. To see why, we return to Figure 1. The rightmost panels show the average values of the "approaches to learning" and "externalizing problem behaviors" scales among Black students, labeled "LEARN rating" and "EXTERN rating," respectively. As was the case for the math and reading ARS ratings, both measures of noncognitive skills are weakly increasing in school minority shares. The similarity across the four teacher-provided ratings is striking, given that all are based on qualitative ratings without obvious objective scales. If ARS

ratings understate differences across schools in cognitive skill levels, then it is likely that teacher reports of noncognitive skills understate differences across schools in noncognitive skill levels.

Several previous researchers have argued that reference bias may play a significant role in teacher and parent reports of noncognitive skills. Heckman and Kautz (2012) point out that all psychological measurements are calibrated based on observed behavior and that the behaviors used to measure one trait can be influenced by the traits of others. Similarly, West et al. (2016) find that students who are randomly selected to enroll in charter schools self-report lower levels of grit and self-control than those who were not selected, presumably due to differences in peer composition and school environment. Most relevant to our analysis, Lundberg (2017, 234) writes that "[m]easures of children's noncognitive skills that are based on teacher and parent reports of externalizing behavior, lying, or the child's ability to maintain focus on an assigned task are likely to be much more sensitive than cognitive test results to incentives, expectations, and peer effects." In sum, environments appear to influence reports of noncognitive skills.[13]

Figure 2 presents additional evidence about the comparability of noncognitive skill ratings across schools in ECLS-K:2011. Unlike Figure 1, this figure focuses on home environment characteristics that predict noncognitive skills. The top panels show binned averages of the SES composite index, maternal education, and maternal marriage rates, while the bottom panels show regression-based predictions of the approaches to learning, externalizing problem behaviors, and self-control skill indices, respectively. These predicted indices are based on OLS regressions of each noncognitive skill rating on the vector of home environment variables and school indicators. In order to focus on the home environments themselves, we exclude the school indicators when we form the predicted indices. The figure shows that Black students in predominantly minority schools are substantially disadvantaged in comparison to Black students in predominantly White schools along dimensions that predict noncognitive skill ratings within schools.

Overall, the patterns shown in Figures 1 and 2 strongly suggest that teachers' opinions of what constitutes "normal" levels of achievement and behavior are systematically different in schools with disadvantaged student populations as compared to more advantaged schools. Because of these reference biases, the unconditional gaps in teacher-reported skills (both cognitive and noncognitive) might understate true Black-White disparities in these skills. We turn next to our approaches to measuring Black-White skill gaps in the context of reference bias.

[13] Schmitt et al. (2007) and Kautz et al. (2015) provide additional evidence regarding the comparability of noncognitive skills measures across groups in different environments. For example, Schmitt et al. (2007) describe the difficulties in interpreting differences across cultures. Using a cross-country survey, they find that, for example, South Korea ranks near the bottom of all countries surveyed in terms of self-reported conscientiousness, in spite of ranking first in the number of hours worked per year. Overall, average self-reported conscientiousness and hours worked were slightly negatively correlated across countries; in contrast, these measures are positively related within individual countries.

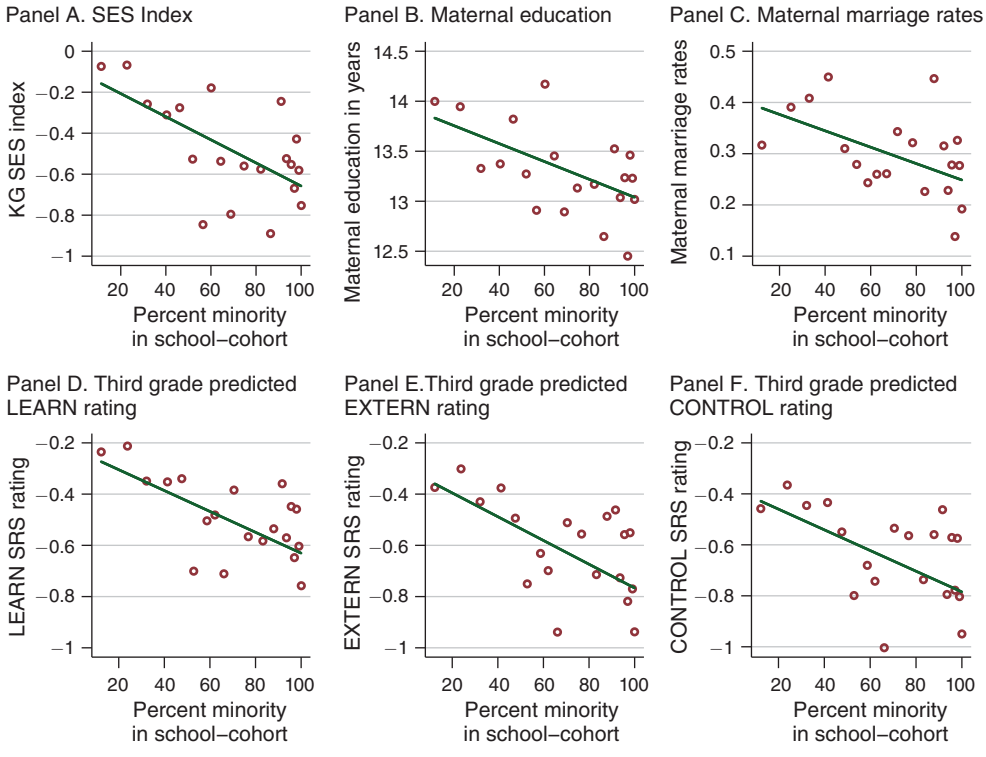

FIGURE 2. AVERAGE CHARACTERISTICS BY SCHOOL RACIAL COMPOSITION AMONG THIRD-GRADE BLACK STUDENTS IN ECLS-K:2011

IV. Estimating Latent Noncognitive Skills

In this section, we describe our approaches to refining our estimates of noncognitive skill gaps. Let y_{ijs} denote the level of skill j for student i in school s, where $j \in \{1, \ldots, J\}$ indexes dimensions of cognitive and noncognitive skills. We are interested in estimating Black-White gaps in noncognitive skills, but the presence of reference bias implies that the observed skill measures are potentially scaled differently in different schools. As a result, we view y_{ijs} as a latent variable, implying that μ_{js}, the average level of skill j within school s, is also latent. In contrast, certain measures of skills, such as test scores, arguably capture scales that do not vary by school. Thus, for achievement test scores, y_{ijs} and μ_{js} are arguably observed, not latent.[14]

[14] As argued by Heckman, Pinto, and Savelyev (2013) and Bond and Lang (2018), among others, test scores are noisy measures of latent cognitive skills. We are implicitly assuming that school-level averages eliminate all (or nearly all) of the noise in individual test scores so that school averages of those scores capture school averages of latent cognitive skills. Although this assumption may not precisely hold in practice, violations would be problematic for our estimation strategy only if deviations of school-average latent cognitive skills from test scores are systematically related to school averages of noncognitive skills.

To focus on the potential role of reference bias, we model observed teacher-reported skills as follows:

$$\tilde{y}_{ijs} = \beta_j b_i + \gamma_j X_s + \kappa_{js} + \varepsilon_{ijs}, \tag{2}$$

where $\tilde{y}_{ijs}$ represents student i's reported level of skill j when that student attends school s, b_i is again an indicator equaling 1 if a student is Black and 0 otherwise, X_s captures measures of school-specific factors that affect skills, and κ_{js} is a school-specific measure of reference bias—it is constant for all students within a school for a skill j. We ignore individual controls and intercepts for simplicity, and we define ε_{ijs} to be orthogonal to b_i, X_s, and κ_{js}.

Our baseline estimates captured Black-White differences in these reported skill measures:

$$\begin{aligned} &E\big(\tilde{y}_{ijs} \big| b_i = 1\big) - E\big(\tilde{y}_{ijs} \big| b_i = 0\big) \\ &\quad = \beta_j + E\big(\gamma_j X_s + \kappa_{js} \big| b_i = 1\big) - E\big(\gamma_j X_s + \kappa_{js} \big| b_i = 0\big). \end{aligned} \tag{3}$$

However, our primary goal is to recover the Black-White differences in latent skills, i.e., the observed skills purged of the reference bias terms κ_{js}:

$$\begin{aligned} &E\big(y_{ijs} \big| b_i = 1\big) - E\big(y_{ijs} \big| b_i = 0\big) \\ &\quad = \beta_j + E\big(\gamma_j X_s \big| b_i = 1\big) - E\big(\gamma_j X_s \big| b_i = 0\big). \end{aligned} \tag{4}$$

Because κ_{js} is constant within school-skill pairs, the latent within-school variance of skill j, σ_j^2, is identical to the corresponding observed within-school variance:

$$\sigma_j^2 \equiv E\big(y_{ijs} - \mu_{js}\big)^2 = E\big(\tilde{y}_{ijs} - \tilde{\mu}_{js}\big)^2, \quad \text{where } \tilde{\mu}_{js} = E\big(\tilde{y}_{ijs} \big| s\big). \tag{5}$$

We pursue several strategies in order to estimate latent Black-White skill gaps. In the first, we impose additional structure on the variances of skills.

CONDITION 1: *The distributions of latent mean skills across schools, $F\big(\mu_{js}\big)$, are equivalent for all skills j up to a proportionality factor given by σ_j; i.e., $F\big(\mu_{js}/\sigma_j\big)$ does not depend on j.*

Condition 1 implies that the ratio of the within-school variance of a skill to the corresponding between-school variance is constant across all skills j. It also implies that, for a given school s, the normalized "skill level" of a school, $C_s \equiv \big(\mu_{js} - \mu_j\big)/\sigma_j$, is constant across all skills, where μ_j denotes the population mean level of skill j.[15]

[15] Alternatively, one could implement the approaches described in this section focusing on classrooms rather than schools as the unit of analysis. In practice, we have found that our point estimates are relatively insensitive to this choice, but estimating classroom-specific reference biases rather than school-specific reference biases leads to less precise estimates of skill gaps.

To get a sense of how we use Condition 1 in practice, consider a school with an average mathematics IRT test score that lies 0.5 standard deviations above the overall population mean μ_j. Condition 1 implies that the average levels of all skills among students in that school lie 0.5 standard deviations above their corresponding population means. If, instead, the school average of the observed "approaches to learning" index lies only 0.1 standard deviations above the overall sample mean, then Condition 1 implies that reference bias has distorted the teacher reports in that school to the extent that they are downward biased by $0.4\sigma_j$. To recover the latent index, we would add $0.4\sigma_j$ to the observed approaches to learning scores for all students at that school, leaving all other features of the distribution unchanged; we then treat this new measure as our estimate of the latent index.

In sum, our estimation procedure is as follows:

Step 1: Choose a benchmark skill k, such as math IRT scores, to calculate $\hat{C}_s = (\hat{\mu}_{ks} - \hat{\mu}_k)/\hat{\sigma}_k$ for each school s.

Step 2: For each subjective skill $j \neq k$, generate estimates $\hat{\mu}_{js}$ using the estimates $\hat{C}_s$ from Step 1 and the skill-specific estimates $\hat{\sigma}_j$ and $\hat{\mu}_j$:

$$\hat{\mu}_{js} = \hat{C}_s \times \hat{\sigma}_j + \hat{\mu}_j.$$

Step 3: Estimate the individual latent measures for skill j:

$$\hat{y}_{ijs} = \tilde{y}_{ijs} - (\tilde{\mu}_{js} - \hat{\mu}_{js}).$$

Step 4: Standardize each estimated latent measure to have an identical sample variance to that of the corresponding observed measure. This allows us to compare the magnitudes of Black-White gaps in the latent measures to the analogous Black-White gaps in the observed measures shown in Table 2.

Before implementing this approach, we first provide additional evidence that the scales of reported noncognitive skills vary across schools. Table 4 presents the ratio of between-school variance to total variance in several observed skill measures in ECLS-K:2011. Panel A shows results for the teacher-reported scales, including three of our measures of noncognitive skills and the two ARS ratings. The first entry in column 1 shows that between-school variation accounts for only 8.5 percent of the total sample variation in the kindergarten "approaches to learning" index. The remaining rows in panel A show analogous estimates for the other four teacher-reported skill assessments and grades. In all cases, between-school variation accounts for little of the overall variation in any of the indices, consistent with Figure 1 above.

Panels B and C show that, in comparison to the teacher-reported assessments, between-school variation accounts for a much larger share of the total variation in relatively objective measures of skills. Panel B shows results for IRT test scores. In all grades, the estimated contribution of between-school variance to IRT test scores is roughly three to four times as large as that of the analogous teacher ARS

Table 4—The Ratio of Between-School Variance to Total Variance of Skill Measures in ECLS-K: 2011

	Kindergarten (1)	1st grade (2)	2nd grade (3)	3rd grade (4)
Panel A. Teacher assessments				
Approaches to learning	0.085	0.066	0.069	0.079
Externalizing problem behaviors	0.116	0.085	0.106	0.144
Self-control	0.127	0.101	0.088	0.110
Math ARS rating	0.063	0.093	0.099	0.078
Reading ARS rating	0.040	0.117	0.070	0.092
Panel B. IRT test scores				
Math	0.249	0.248	0.296	0.310
Reading	0.212	0.277	0.314	0.316
Panel C. Predicted teacher assessments				
Approaches to learning	0.322	0.311	0.307	0.309
Externalizing problem behaviors	0.324	0.281	0.271	0.263
Self-control	0.323	0.314	0.291	0.262
Math ability	0.340	0.347	0.330	0.334
Reading ability	0.336	0.338	0.327	0.329

Notes: (i) Each entry in the table corresponds to an estimate of the ratio of the between-school variance of a particular skill measure to its total variance. (ii) Predicted teacher assessments are generated from linear regressions of teacher assessment on the home environment variables and school fixed effects, forming predicted values based on the estimated coefficients on the home environment variables.

ratings. Panel C shows estimates for the predicted teacher assessments described in Section IV. The estimates imply that there is much more sorting on the characteristics that predict noncognitive skills than on those skills themselves.

The intuition underlying Condition 1 is that the degree of sorting across schools on objective measures—such as test scores—is informative about the degree of sorting on the latent noncognitive skills. However, we acknowledge that the degree of sorting across schools may be considerably stronger for some skills than for others. The "skill level" of a school, C_s, is identified in the presence of a single objective measure, such as math IRT scores. When there are multiple objective measures—such as math and reading IRT scores—C_s is overidentified. In practice, we will estimate $\hat{C}_s$ for both reading and math IRT scores and use the simple average of the two for each grade level. Below, we discuss the sensitivity of our results to alternatives such as using only one of the IRT scores or including characteristics of the home environment.

We turn next to our second approach to estimating latent Black-White skill gaps. Specifically, we gauge the strength of reference bias in a school by comparing the average of objective measures of cognitive skills in that school to the corresponding subjective (teacher-provided) measures of cognitive skills. For example, under the assumption that math IRT test scores are free of reference bias, the difference between the within-school averages of those scores and math ARS ratings identifies κ_{js}, the reference bias in teacher-reported math skills. We then apply the following condition.

CONDITION 2: *The reference bias in teacher reports of student skills,* κ_{js}*, does not vary across j.*

Under Condition 2, we rely solely on the difference between subjective and objective measures of cognitive skills to generate an estimate of κ_{js}. Consider again a school with an average mathematics IRT test score that lies 0.5 standard deviations above the overall population mean. If that school's average teacher-reported mathematics ARS score lies only 0.2 standard deviations above the population mean of ARS scores, we would conclude that the average teacher reports in that school understate students' math aptitude by 0.3 standard deviations. Condition 2 then implies that teacher reports in that school also understate students' average noncognitive skills by 0.3 standard deviations. We would then add 0.3 standard deviation units to the observed noncognitive skills for all students at that school, leaving all other features of the distribution unchanged; we treat this new measure as our estimate of the latent skill index.[16]

The identifying assumptions embedded in Conditions 1 and 2, while not identical, share several similar features. Like Condition 1, Condition 2 generates multiple estimates of latent noncognitive skills when there are multiple objective measures available. Specifically, one could estimate κ_{js} using either math aptitude (i.e., comparing IRT test scores to teachers' subjective math ratings) or reading aptitude. In practice, we will again use both reading and math IRT scores and take the simple average of the two for each grade level. Below, we discuss the sensitivity of the estimates to this choice.

V. Estimates of Latent Noncognitive Skill Gaps

In Table 5, we present estimates of the Black-White gaps in third-grade observed and latent noncognitive skills in ECLS-K:2011. The top row of column 1 replicates the baseline estimates shown in the first column of Table 2, showing the raw gap in observed noncognitive skills. The second row shows the baseline estimates from models that include the home environment controls. Column 2 presents analogous gaps in the estimated latent noncognitive skills, as estimated using Condition 1, while column 3 uses Condition 2. We label these estimates as "Method 1" and "Method 2," respectively.

For each of the five noncognitive skills, the estimated latent gaps are substantially larger than are the corresponding observed gaps. For example, in panel A, columns 1 and 2 of the top row imply that the raw gap in the latent "approaches to learning" index is 0.322 standard deviations larger (-0.846 versus -0.524) than is the gap in the observed index. Including home environment controls reduces the estimates gaps in both columns, but the conditional gap in column (2) is considerably larger than the analogous gap in column 1 (-0.496 versus -0.218).[17] In all cases, the

[16] Similar to our implementation of Condition 1, we again rescale by $\sqrt{R_j/R_k}$, where R_j and R_k represent the reliabilities of the subjective and objective skills, respectively.

[17] In all cases, the inclusion of the home environment controls reduces the size of the latent gaps more than the corresponding observed gaps. For example, comparing the top two rows of column 1 in panel A shows that the inclusion of the controls reduces the estimated gap by 0.306 standard deviations, from -0.524 to -0.218, while the corresponding reduction in column 2 is 0.350 standard deviations, from -0.846 to -0.496. We conjecture that this phenomenon arises because a portion of the observed gaps reflects variation across schools in what constitutes "normal" behavior, and this variation has the opposite-signed relationship to the control variables as the within-school variation—students in high-SES schools tend to have lower observed ratings than identically behaved students in

Table 5—Estimates of Racial Gaps in Latent Third-Grade Noncognitive Skills, ECLS-K: 2011

	(1)	(2)	(3)	(4)
Panel A. Approaches to learning				
Raw gaps	−0.524	−0.846	−0.893	−0.973
	(0.035)	(0.041)	(0.044)	(0.028)
Including controls	−0.218	−0.496	−0.474	
	(0.037)	(0.042)	(0.045)	
Panel B. Externalizing problem behaviors				
Raw gaps	−0.521	−0.821	−0.886	−1.078
	(0.036)	(0.041)	(0.045)	(0.028)
Including controls	−0.246	−0.479	−0.497	
	(0.039)	(0.044)	(0.045)	
Panel C. Self-control				
Raw gaps	−0.584	−0.830	−0.946	−1.110
	(0.036)	(0.042)	(0.046)	(0.028)
Including controls	−0.284	−0.480	−0.529	
	(0.039)	(0.045)	(0.046)	
Panel D. Interpersonal skills				
Raw gaps	−0.486	−0.814	−0.847	−1.113
	(0.036)	(0.042)	(0.038)	(0.028)
Including controls	−0.215	−0.482	−0.467	
	(0.039)	(0.044)	(0.045)	
Panel E. Internalizing problem behaviors				
Raw gaps	−0.107	−0.686	−0.471	−0.997
	(0.036)	(0.042)	(0.045)	(0.028)
Including controls	0.091	−0.385	−0.187	
	(0.040)	(0.045)	(0.046)	
Baseline estimates	X			
Adjustment via Method 1		X		
Adjustment via Method 2			X	
Adjustment via Method 3				X

Notes: (i) Each entry in the table is an estimate from a separate regression of a measure of noncognitive skills on an indicator equaling 1 for black children and 0 for white children. (ii) Standard errors in columns 2–4 are derived from 200 bootstrap replications using stratified resampling within schools. (iii) In column 2, we adjust estimates using Method 1, which imposes that the ratio of the within-school to between-school variances of skills is constant across all skills. In column 3, we use Method 2, which imposes that school-level reference biases do not vary across skills. In column 4, we estimate gaps in regression-based predictions of noncognitive skills rather than the teacher-reported skills themselves.

estimates in column 3 are similar in magnitude to those in column 2, suggesting that the adjustments are robust to the variation in assumptions underlying the two approaches.[18]

low-SES schools. The estimated latent scales purge this source of variation, with the result that the home environment controls have more explanatory power in the latent scales than in the observed scales. The partial R^2 values of the home environment controls support this conjecture; for example, in column 1 the partial R^2 of the home environment controls is 0.128, compared to 0.172 in column 4.

[18] The estimates in columns 2 and 3 are based on $\hat{C}_s$, $\hat{\kappa}_{js}$, and/or the estimated within- and between-school variance of reported noncognitive skills measures. The ECLS-K datasets use two-stage stratified sample designs in which students are randomly sampled within schools, and the resulting school-level samples include fewer than 30 students in almost all cases. For inference, we account for the resulting estimation error in our estimates of $\hat{C}_s$ and $\hat{\kappa}_{js}$ by using a stratified bootstrap procedure. Specifically, we randomly sample students with replacement in

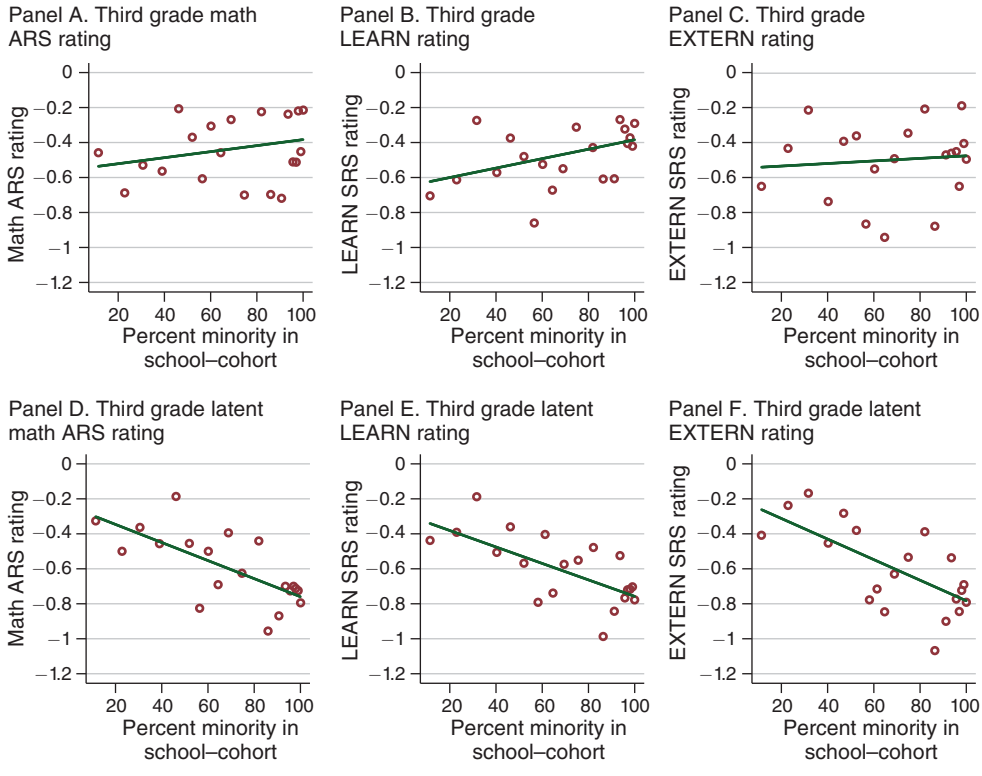

FIGURE 3. AVERAGE SKILLS BY SCHOOL RACIAL COMPOSITION AMONG THIRD-GRADE BLACK STUDENTS IN ECLS-K:2011

Note: In the bottom panels of the figure, we plot estimated latent measures of skills based on Method 1, which imposes that the ratio of the within-school to between-school variances of skills is constant across all skills.

Finally, in column 4, we present Black-White gaps in the predicted noncognitive skills indices, which we label as "Method 3" hereafter. In order to use only within-school variation, we exclude the school indicators when we form the predicted indices. For all five skills, the resulting estimated Black-White gaps are even larger than those based on Conditions 1 and 2. Note that the home environment controls are perfectly collinear with the predicted indices, so the table does not include estimates for models that include home environment controls.[19]

In Figure 3, we show how observed and estimated latent skills vary by school racial composition. The top panels show the average skills across 20 racial composition bins for 3 observed teacher-reported skills—math, approaches to

each school, with the resulting within-school bootstrap samples equal in size to the original samples. In each replication, we estimate the latent skill measures using Steps 1–4 described above for Condition 1 (and use the analogous steps for Condition 2), and we then estimate Black-White gaps in the estimated latent skills. We use 200 bootstrap replications in all cases.

[19] Our adjustments based on Method 3 differ from those based on Methods 1 and 2 in one important respect. Specifically, the adjustments in Methods 1 and 2 shift noncognitive skills for every student in a given school by the same amount, as Conditions 1 and 2 both imply that reference bias in a particular skill is constant within a school. Method 3, on the other hand, does not impose this assumption of constant reference biases within a school.

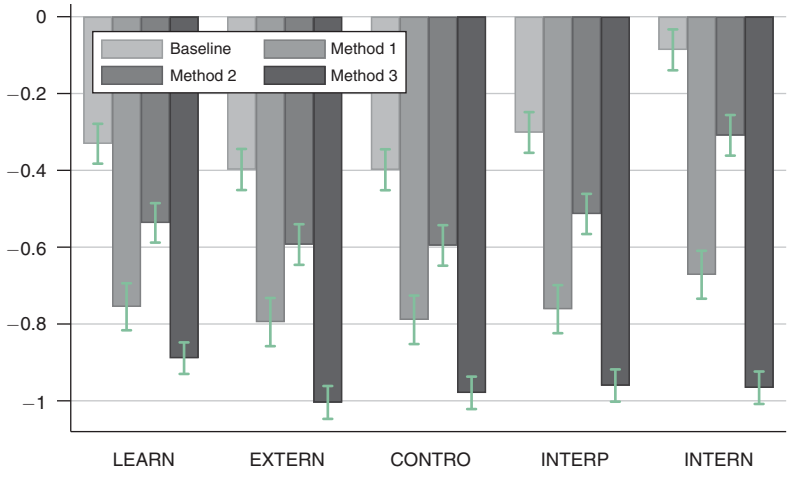

FIGURE 4. ESTIMATES OF NONCOGNITIVE SKILL GAPS IN KINDERGARTEN BY ESTIMATION METHOD IN ECLS-K:2011

Notes: The estimates for Method 1 impose that the ratio of the within-school to between-school variances of skills is constant across all skills. Method 2 imposes that school-level reference biases are constant across skills. Method 3 uses as dependent variables the regression-based predictions of noncognitive skills rather than the teacher-reported skills themselves.

learning, and externalizing problem behaviors—reproducing the findings from Figure 1. The bottom panels replace the observed skills with the corresponding estimated latent skills based on Method 1, which are steeply decreasing in school minority shares (results based on Method 2 are similarly downward-sloping). For example, the gradient in the estimated latent ARS math rating differs sharply from the observed ARS math rating and is instead similar to the gradient in IRT math scores shown above in Figure 1.

For brevity, we have focused thus far on third-grade skill gaps. Tables A2–A4 in the online Appendix present analogous estimates for kindergarten, first grade, and second grade. In each grade, the estimated latent skill gaps in columns 2–4 are substantially larger than the baseline gaps in all cases. Figures 4–7 also present the unconditional gaps graphically for kindergarten, first, second, and third grade, respectively, with overlaid 95 percent confidence intervals.

Figures 4–7 additionally highlight that the estimated latent gaps are sensitive to whether we use Methods 1, 2, or 3 to construct them. The variation is strongest in kindergarten and weakens in later grades, but it does not entirely disappear by third grade. This sensitivity underscores that each of our approaches to estimating latent gaps involves strong assumptions that may be unlikely to hold in practice. For example, Condition 1 implies that the distributions of latent mean skills across schools are equivalent across for all skills, but in reality, the degree of sorting across schools may be considerably stronger for some skills than for others.

Furthermore, our estimates of latent skill gaps vary not only across approaches but within approaches as well. For example, to use Method 2 in order to estimate school-specific reference biases κ_{js}, we could use only math aptitude, only reading aptitude, or a combination of both. The estimated latent "approaches to learning"

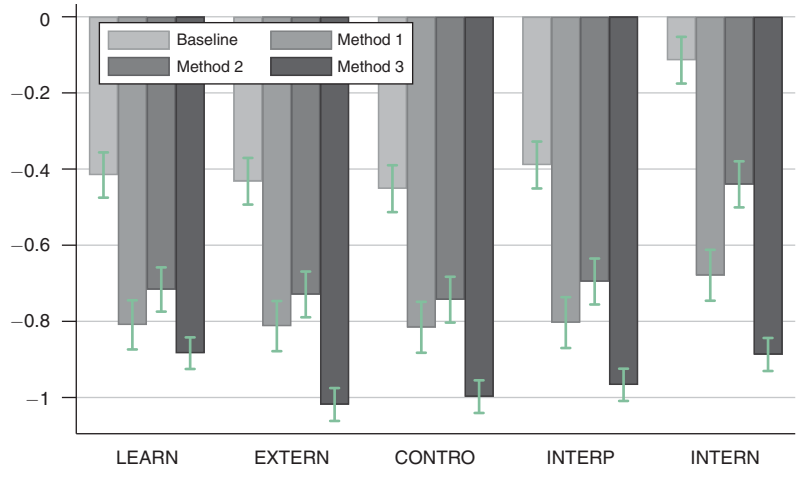

FIGURE 5. ESTIMATES OF NONCOGNITIVE SKILL GAPS IN FIRST GRADE BY ESTIMATION METHOD IN ECLS-K:2011

Notes: The estimates for Method 1 impose that the ratio of the within-school to between-school variances of skills is constant across all skills. Method 2 imposes that school-level reference biases are constant across skills. Method 3 uses as dependent variables the regression-based predictions of noncognitive skills rather than the teacher-reported skills themselves.

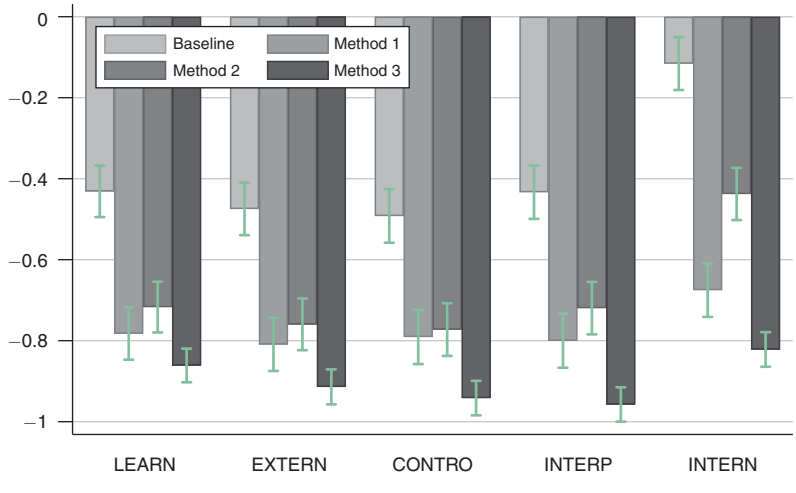

FIGURE 6. ESTIMATES OF NONCOGNITIVE SKILL GAPS IN SECOND GRADE BY ESTIMATION METHOD IN ECLS-K:2011

Notes: The estimates for Method 1 impose that the ratio of the within-school to between-school variances of skills is constant across all skills. Method 2 imposes that school-level reference biases are constant across skills. Method 3 uses as dependent variables the regression-based predictions of noncognitive skills rather than the teacher-reported skills themselves.

gap shown in Table 5, −0.893, uses the school average of reading and math ARS ratings minus the school average of reading and math IRT test scores to estimate κ_{js}. If we instead used only the math ARS ratings and IRT test scores, the implied Black-White gap is −0.940. Alternatively, if we used only the reading ARS

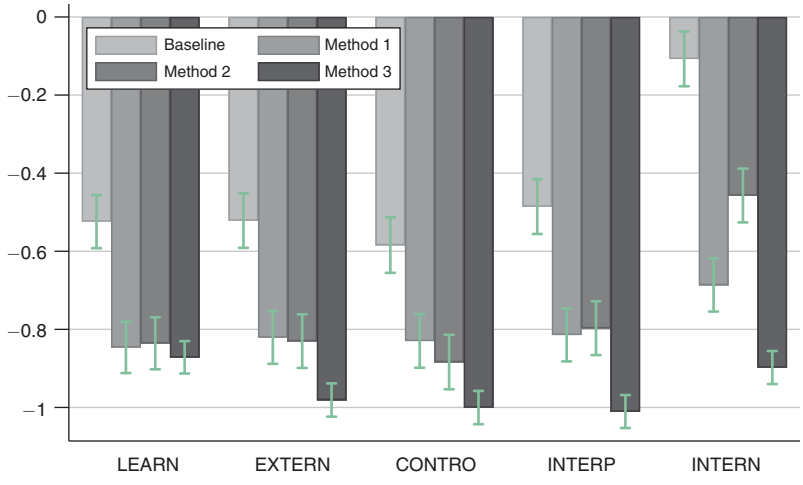

Figure 7. Estimates of Noncognitive Skill Gaps in Third Grade by Estimation Method in ECLS-K:2011

Notes: The estimates for Method 1 impose that the ratio of the within-school to between-school variances of skills is constant across all skills. Method 2 imposes that school-level reference biases are constant across skills. Method 3 uses as dependent variables the regression-based predictions of noncognitive skills rather than the teacher-reported skills themselves.

ratings and IRT test scores, the implied Black-White gap is −0.847. Method 1 produces similar variability, with analogous point estimates of −0.846, −0.915, and −0.755 when we use both math and reading scores, math scores only, and reading scores only, respectively, to estimate school-specific skill levels C_s.

Although we cannot test the assumptions underlying Methods 1 and 2 directly, we can assess their predictive power relative to the "no reference bias" assumption inherent in the baseline estimates. Specifically, in Table 6 we show how applying our methods to teacher-reported cognitive skill ratings affects these ratings' associations with future IRT test scores.

The first column in panel A of the table shows estimates from three different models of third-grade math IRT scores. The first model includes only the unadjusted teacher-provided first-grade ARS math ratings, while the second includes only the corresponding adjusted ratings, estimated via Method 1. The third includes both of these measures. As the estimates show, the adjusted ratings are much stronger predictors of future test scores than are the unadjusted ratings. Specifically, the estimated effect of the adjusted rating in the first column changes modestly, from 0.809 to 0.980, when the unadjusted rating is included. In contrast, the estimated effect of the unadjusted rating declines from 0.683 to −0.193 when the adjusted rating is included.

Column 2 is identical to column 1 except that the adjusted ratings are estimated using Method 2 instead of Method 1. The estimates show similar patterns to those in Column 1. Columns 3 and 4 repeat this exercise using third-grade reading IRT scores and first-grade reading ARS ratings. In every case, the coefficient on the unadjusted ratings is negative when the adjusted ratings are included. One interpretation of these results is that the variation in the unadjusted ratings that is orthogonal to the adjusted ratings captures only reference bias, which is negatively associated with future test scores.

Table 6—The Association between Adjusted and Unadjusted ARS Ratings and Future Test Scores in ECLS-K: 2011

	Mathematics		Reading	
	(1)	(2)	(3)	(4)
Panel A. Third-grade IRT test scores				
Unadjusted first grade ARS rating only	0.683	—	0.693	—
	(0.008)	—	(0.008)	—
Adjusted rating only	0.809	0.766	0.800	0.766
	(0.009)	(0.010)	(0.009)	(0.009)
Including both adjusted and unadjusted ratings				
Unadjusted first grade ARS rating	−0.193	−0.215	−0.133	−0.158
	(0.016)	(0.021)	(0.016)	(0.020)
Adjusted rating	0.980	0.965	0.919	0.912
	(0.017)	(0.022)	(0.017)	(0.021)
Panel B. First-grade IRT test scores				
Unadjusted KG ARS rating only	0.656	—	0.754	—
	(0.008)	—	(0.007)	—
Adjusted rating only	0.768	0.744	0.813	0.814
	(0.010)	(0.009)	(0.009)	(0.009)
Including both adjusted and unadjusted ratings				
Unadjusted KG ARS rating	−0.075	−0.274	0.159	−0.042
	(0.016)	(0.021)	(0.015)	(0.018)
Adjusted rating	0.833	0.997	0.675	0.853
	(0.017)	(0.021)	(0.017)	(0.020)
Adjustment via Method 1	X		X	
Adjustment via Method 2		X		X

Notes: Each entry in the table represents an estimate from a regression of IRT test scores on teacher-reported ARS ratings of cognitive skills. The estimates in the "Including both adjusted and unadjusted ratings" panels come from specifications that include both the raw teacher-reported ARS ratings and the adjusted ratings, using either Method 1 or Method 2. Standard errors are derived from 200 bootstrap replications using stratified resampling within schools.

Panel B presents analogous estimates from models in which the dependent variables are first-grade IRT test scores and the independent variables are kindergarten ARS ratings. The findings mirror those shown in the top panel, in that the coefficients on the unadjusted ARS ratings decline dramatically when the adjusted ratings are included.

In sum, we view our estimates based on Methods 1–3 as complements to the baseline estimates rather than as replacements for them. The assumptions underlying each of these approaches are strong and unlikely to hold in practice. Nonetheless, the evidence provided in Table 6 is consistent with the view that those assumptions are a better approximation to reality than the implicit assumption underlying the baseline results: that the standards that teachers use to evaluate achievement and behavior are invariant to the composition of a school's student body.

VI. Discussion and Conclusions

Using two nationally representative datasets from the ECLS-K, we find significant differences in observed measures of noncognitive skills between White and Black elementary school students, even after controlling for a large set of background variables. The raw gaps in these skills are roughly 0.5 standard deviation units by the end of third grade.

Given that observed noncognitive skills are based on subjective judgments by teachers, we conjectured that the large estimated Black-White gaps may stem from teachers' biases against Black students. We instead found evidence suggesting that the baseline estimates substantially understate true Black-White disparities. Teachers appear to base their responses on the skills of "typical" students in their classrooms, at least in part, so a student in a high-achieving school will tend to have lower teacher-reported skills than an identical student in a low-achieving school. Because White students are disproportionately likely to attend high-achieving schools, this phenomenon compresses the unconditional Black-White gap in observed noncognitive skills.

In order to correct for the effect of reference bias on our baseline estimates, we adopt three approaches that treat the underlying skills as latent variables. The first approach assumes that the distributions of latent mean skills across schools are common for all skills up to a proportionality factor given by the within-school standard deviation of skills, which allows us to recover the latent distributions from the corresponding observed distributions. The second imposes that the magnitude of reference bias is identical across all teacher-reported skills, while the third assumes that the degree of sorting across schools in noncognitive skills is identical to the degree of sorting in observable home-environment characteristics that are correlated with those skills.

We then estimate Black-White gaps in our estimated latent skills. For each of the skills we consider, the estimated latent gaps are substantially larger than the corresponding observed gaps. In third grade, the raw gaps in four of the five latent skills are roughly 80 percent of the corresponding full-sample standard deviations.

Finally, our findings imply that reference bias is negatively related to school-level achievement levels. As a result, studies that use subjective measures of skills without accounting for such bias arguably understate the impacts of noncognitive skills on adult outcomes.

The magnitudes of our estimated Black-White gaps are disheartening, but they highlight a potentially important cause of racial inequality in the United States. A nascent literature has established that noncognitive skills have large impacts on adult outcomes and are much more malleable than are cognitive skills after age five (see Kautz et al. 2017). As a result, our estimates imply that interventions aimed at reducing noncognitive skill gaps among young children might be an effective policy tool for ameliorating Black-White disparities in adult outcomes.

REFERENCES

Agan, Amanda Y. 2011 "Non-cognitive Skills and Crime." https://www.iza.org/conference_files/CoNoCoSk2011/agan_a6558.pdf.

Bertrand, Marianne, and Jessica Pan. 2013. "The Trouble with Boys: Social Influences and the Gender Gap in Disruptive Behavior." *American Economic Journal: Applied Economics* 5 (1): 32–64.

Bond, Timothy N., and Kevin Lang. 2013. "The Evolution of the Black-White Test Score Gap in Grades K–3: The Fragility of Results." *Review of Economics and Statistics* 95 (5): 1468–79.

Bond, Timothy N., and Kevin Lang. 2018. "The Black-White Education Scaled Test-Score Gap in Grades K–7." *Journal of Human Resources* 53 (4): 891–917.

Borghans, Lex, Angela Lee Duckworth, James J. Heckman, and Bas ter Weel. 2008. "The Economics and Psychology of Personality Traits." *Journal of Human Resource* 43 (4): 972–1059.

Cornwell Christopher, David B. Mustard, and Jessica Van Parys. 2013. "Noncognitive Skills and the Gender Disparities in Test Scores and Teacher Assessments: Evidence from Primary School." *Journal of Human Resources* 48 (1): 236–64.

Cunha, Flavio, James J. Heckman, and Susanne M. Schennach. 2010. "Estimating the Technology of Cognitive and Noncognitive Skill Formation." *Econometrica* 78 (3): 883–931.

Demaray, Michelle K., Stacey L. Ruffalo, John Carlson, R.T. Busse, Amy E. Olson, Susan M. McManus, and Amy Leventhal. 1995. "Social Skills Assessment: A Comparative Evaluation of Six Published Rating Scales." *School Psychology Review* 24: 648–71.

Deming, D.J. 2017. "The Growing Importance of Social Skills in the Labor Market." *Quarterly Journal of Economics* 132 (4): 1593–1640.

DiPrete, Thomas A., and Jennifer L. Jennings. 2012. "Social and Behavioral Skills and the Gender Gap in Early Educational Achievement." *Social Science Research* 41 (1): 1–15.

Duckworth, Angela Lee, and Patrick D. Quinn. 2009. "Development and Validation of the Short Grit Scale (GRIT–S)." *Journal of Personality Assessment* 91 (2): 166–74.

Elder, Todd E. 2010. "The Importance of Relative Standards in ADHD Diagnoses: Evidence Based on Exact Birth Dates." *Journal of Health Economics* 29 (5): 641–56.

Elder, Todd, and Yuqing Zhou. 2021. "Replication data for: Black-White Gap in Noncognitive Skills among Elementary School Children." American Economic Association [publisher], Inter-university Consortium for Political and Social Research [distributor]. https://doi.org/10.38886/EE117301V2.

Elliott, Stephen N., Frank M. Gresham, Terry Freeman, and George McCloskey. 1988. "Teacher and Observer Ratings of Children's Social Skills: Validation of the Social Skills Rating Scales." *Journal of Psychoeducation Assessment* 6 (2): 152–61.

Flossmann, Anton L., Rémi Piatek, and Laura Wichert. 2007. "Going Beyond Returns to Education: The Role of Noncognitive Skills on Wages in Germany." https://www.researchgate.net/profile/Laura_Wichert/publication/228342336_Going_beyond_returns_to_education_The_role_of_noncognitive_skills_on_wages_in_Germany/links/55a774c808aeceb8cad6283c/Going-beyond-returns-to-education-The-role-of-noncognitive-skills-on-wages-in-Germany.pdf.

Fryer, Roland G., Jr., and Steven D. Levitt. 2004. "Understanding the Black-White Test Score Gap in the First Two Years of School." *Review of Economics and Statistics* 86 (2): 447–64.

Fryer, Roland G., Jr., and Steven D. Levitt. 2006. "The Black-White Test Score Gap through Third Grade." *American Law and Economics Review* 8 (2): 249–81.

Fryer, Roland G. Jr., and Steven D. Levitt. 2013. "Testing for Racial Differences in the Mental Ability of Young Children." *American Economic Review* 103 (2): 981–1005.

Fryer, Roland G., Jr., Steven D. Levitt, and John A. List. 2015. "Parental Incentives and Early Childhood Achievement: A Field Experiment in Chicago Heights." NBER Working Paper 21477.

Goldammer, Christian. 2012."Racial Gaps in Cognitive and Noncognitive Skills: The Asian Exception." Unpublished.

Gresham, Frank M., and Stephen N. Elliott. 1990. *Social Skills Rating System Manual*. Circle Pines, MN: American Guidance Service.

Heckman, James J. 2008. "Schools, Skills, and Synapses." *Economic Inquiry* 46 (3): 289–324.

Heckman, James J., and Tim Kautz. 2012. "Hard Evidence on Soft Skills." *Labour Economics* 19 (4): 451–64.

Heckman, James, Rodrigo Pinto, and Peter Savelyev. 2013. "Understanding the Mechanisms through Which an Influential Early Childhood Program Boosted Adult Outcomes." *American Economic Review* 103 (6): 2052–86.

Heckman, James J., and Yona Rubinstein. 2001. "The Importance of Noncognitive Skills: Lessons from the GED Testing Program." *American Economic Review* 91 (2): 145–49.

Heckman, James J., Jora Stixrud, and Sergio Urzua. 2006. "The Effects of Cognitive and Noncognitive Abilities on Labor Market Outcomes and Social Behavior." *Journal of Labor Economics* 24 (3): 411–82.

Jencks, Christopher, and Meredith Phillips. 1998. "The Black-White Test Score Gap: An Introduction." In *The Black-White Test Score Gap*, edited by Christopher Jencks and Meredith Phillips, 1–51. Washington, DC: Brookings Institution Press.

Kautz, Tim, James J. Heckman, Ron Diris, Bas ter Weel, and Lex Borghans. 2017. "Fostering and Measuring Skills: Improving Cognitive and Non-cognitive Skills to Promote Lifetime Success." NBER Working Paper 20749.

Kautz, Tim, Wladimir Zanoni, and Chapin Hall. 2015. "Using School Administrative Data to Measure Non-Cognitive Skills" Unpublished.

Lundberg, Shelly. 2017. "Chapter 6—Noncognitive Skills as Human Capital." In *Education, Skills, and Technical Change—Implications for Future US GDP Growth*, edited by Charles R. Hulten and Valerie A. Ramey, 219–50. Chicago: University of Chicago Press.

Magnuson, Katherine, and Jane Waldfogel, eds. 2008. *Steady Gains and Stalled Progress: Inequality and the Black-White Test Score Gap*. Russell Sage Foundation.

National Center for Education Statistics (NCES), US Department of Education. 2002. "Early Childhood Longitudinal Study—Kindergarten Class of 1998–99 (ECLS–K), Psychometric Report for Kindergarten through First Grade." NCES Working Paper 2002-05.

Neal, Derek. 2006 "Why has black–white skill convergence stopped?" *Handbook of the Economics of Education* 1 (2006): 511-76.

Neidell, Matthew, and Jane Waldfogel. 2010. "Cognitive and Noncognitive Peer Effects in Early Education." *Review of Economics and Statistics* 92 (3): 562–76.

Piatek, Rémi, and Pia Pinger. 2016 "Maintaining (Locus of) Control? Data Combination for the Identification and Inference of Factor Structure Models." *Journal of Applied Econometrics* 31 (4): 734–55.

Schmitt, David P., Jüri Allik, Robert R. McCrae, and Verónica Benet-Martínez. 2007. "The Geographic Distribution of Big Five Personality Traits: Patterns and Profiles of Human Self-Description across 56 Nations." *Journal of Cross-Cultural Psychology* 38 (2): 173–212.

Segal, Carmit. 2013. "Misbehavior, Education, and Labor Market Outcomes." *Journal of the European Economic Association* 11 (4): 743–79.

Urzúa, Sergio, 2008. "Racial Labor Market Gaps: The Role of Abilities and Schooling Choices." *Journal of Human Resources* 43 (4): 919–71.

West, Martin R., Matthew A. Kraft, Amy S. Finn, Rebecca E. Martin, Angela L. Duckworth, Christopher F.O. Gabrieli, and John D.E. Gabrieli. 2016. "Promise and Paradox: Measuring Students' Non-cognitive Skills and the Impact of Schooling." *Educational Evaluation and Policy Analysis* 38 (1): 148–70.