Changes in the Black-White Test score Gap in the Elementary School Grades

CSE Report 715

Daniel Koretz and Young-Suk Kim

National Center for Research on Evaluation, Standards, and Student Testing, University of California, Los Angeles/ Harvard Graduate School of Education

April 2007

Center for the Study of Evaluation
National Center for Research on Evaluation, Standards, and Student Testing
Graduate School of Education & Information Studies
University of California, Los Angeles
Los Angeles, CA 90095-1522
(310) 206-1532

The work reported herein was supported under the Educational Research and Development Centers Program, PR/Award Number R305B960002, as administered by the Office of Educational Research and Improvement, U.S. Department of Education.

The findings and opinions expressed in this report do not reflect the positions or policies of the National Institute on Student Achievement, Curriculum, and Assessment, the Office of Educational Research and Improvement, or the U.S. Department of Education.

CHANGES IN THE BLACK-WHITE TEST SCORE GAP IN THE ELEMENTARY SCHOOL GRADES

Daniel Koretz and Young-Suk Kim[1]
National Center for Research on Evaluation, Standards, and Student Testing, University of California, Los Angeles/ Harvard Graduate School of Education

Abstract

In a pair of recent studies, Fryer and Levitt (2004a, 2004b) analyzed the Early Childhood Longitudinal Study – Kindergarten Cohort (ECLS-K) to explore the characteristics of the Black-White test score gap in young children. They found that the gap grew markedly between kindergarten and the third grade and that they could predict the gap from measured characteristics in kindergarten but not in the third grade. In addition, they found that the widening of the gap was differential across areas of knowledge and skill, with Blacks falling behind in all areas other than the most basic. They raised the possibility that Black and Whites may not be on "parallel trajectories" and that Blacks, as they go through school, may never master some skills mastered by Whites.

This study re-analyzes the ECLS-K data to address this last question. We find that the scores used by Fryer and Levitt (proficiency probability scores, or PPS) do not support the hypothesis of differential growth of the gap. The patterns they found reflect the nonlinear relationships between overall proficiency, θ, and the PPS variables, as well as ceiling effects in the PPS distributions. Moreover, θ is a sufficient statistic for the PPS variables, and therefore, PPS variables merely re-express the overall mean difference between groups and contain no information about qualitative differences in performance between Black and White students at similar levels of θ. We therefore carried out differential item functioning (DIF) analyses of all items in all rounds of the ECLS-K through grade 5 (Round 6), excluding only the fall of grade 1 (which was a very small sample) and subsamples in which there were too few Black students for reasonable analysis. We found no relevant patterns in the distribution of the DIF statistics or in the characteristics of the items showing DIF that support the notion of differential divergence, other than in kindergarten and the first grade, where DIF favoring Blacks tended to be on items tapping simple skills taught outside of school (e.g., number recognition), while DIF disfavoring Blacks tended to be on material taught more in school (e.g., arithmetic). However, there were exceptions to this. Moreover, because of its construction and reporting, the ECLS-K data were not ideal for addressing this

[1]Young-Suk Kim is currently at the Florida Center for Reading Research (FCRR) and Department of Childhood Education, Reading, and Disability Services, College of Education, Florida State University

question, and data better suited to the purpose might show differential divergence across areas of knowledge and skill. The paper concludes by advising secondary analysts examining this question to be wary of aspects of test design that may influence the results and to be sensitive to likely variations in findings across databases.

Few issues in educational measurement have garnered as much attention as the large mean differences in performance between racial and ethnic groups – in particular, between African American and White students.

In a pair of recent provocative studies, Roland Fryer and Steven Levitt used the Early Childhood Longitudinal Study – Kindergarten Cohort (ECLS-K) to explore changes in the characteristics of the Black-White gap between kindergarten and the third grade. In some respects, the ECLS-K provides an excellent opportunity to investigate these issues because it provides data from linked adaptive assessments administered to a nationally representative sample of more than 20,000 children tracked longitudinally, beginning in kindergarten. In the first of the studies, Fryer and Levitt (2004a) followed children from the fall of kindergarten through spring of the first grade. They found, in contrast to previous studies, that the relatively small difference between Blacks and Whites in the fall of kindergarten vanished when they controlled for a small number of covariates. However, the gap grew between then and the spring of first grade, even conditional on those covariates. In the second study, Fryer and Levitt (2004b) followed the ECLS-K sample through third grade. They reported that Blacks fall progressively further behind Whites and that this difference cannot be explained by observable student characteristics that they included in regression models. They also found a worrisome pattern in the specific skills that showed a growing gap:

> Blacks are falling behind in virtually all categories of skills tested, except the most basic. Over time, Black students lose ground in virtually every skill area, except the most basic skills that are mastered by virtually all students in the grade…. It is difficult to know precisely what conclusion to draw from these results. To the extent that the pattern of Black skill acquisition as students age follows the path of the basic skills, i.e., Black students master the material, but at a somewhat later age than White students, the patterns maybe construed as encouraging. The implication would be that Black students, although lagging Whites at any particular point in time, are on parallel trajectories. Much more troubling, it would seem, is the possibility that as the skills become more difficult, e.g., division, a nontrivial fraction of the Black students may never master the skills. (Fryer & Levitt, 2004b, pp. 18-19).

As Fryer and Levitt note, while this finding is difficult to interpret, it would be indeed troubling if the Black-White gap grows more rapidly in higher-order skills.

Do Fryer and Levitt Show Differential Divergence by Level of Skill?

To understand what the patterns uncovered by Fryer and Levitt (2004b) signify, it is necessary to examine the outcome variables they used.

Fryer and Levitt's conclusions about qualitative characteristics of the growth in the achievement are based on analyses of the ECLS-K "proficiency probability scores." The proficiency probability scores (labeled PPS here for brevity) are intended to represent mastery of specific bundles of skills that are given labels such as "place value" and "multiply divide" (National Center for Education Statistics [NCES], 2004). Fryer and Levitt (2004b) seemingly interpreted this to mean that the PPS scores provide information about these specific skills independent of the estimated overall proficiency estimated for each student, but they do not.

The ECLS-K tests were scaled using a unidimensional 3-parameter logistic item response theory (IRT) model (Rock & Pollack, 2002), which yields a single overall proficiency estimate, conventionally labeled θ, for each student. To create the PPS scores, five small sets of items were chosen in both reading and mathematics to represent "agreed-on learning milestones in reading and mathematics" (Rock & Pollack, 2002. p. 3-9)—that is, proficiencies characteristic of the performance of students at different levels of θ. The goal, which was reached for most students, was that these five clusters would form a Guttman scale. Scaling of performance on these clusters was carried out as follows:

> A child was deemed proficient at any one level if he or she passed any three out of four items. An additional single item was then constructed for each of the five proficiency levels. A child was given a "1" on these supplemental items if he or she got any three out of four correct on each set of four items that marked the five proficiency levels; otherwise the score was zero. The creation of these "super items" and the subsequent estimation of their IRT parameters located the five proficiency levels on the reading score scale. This parameter estimation allows one also to *estimate a continuous measure of the child's probability of being proficient at each of the five levels using the child's IRT ability estimate score and the parameters for each of the "super items"* (Rock & Pollack, 2002, p. 3-10, emphasis added).

The probabilities estimated in the final step, italicized above, are the PPS. In other words, the PPS are an estimate of the probability of attaining mastery of each cluster, as mastery is defined above, as a function of θ and the "super item" parameters for each cluster. They are analogous to IRT number right true scores in that they are one-to-one nonlinear transformations of θ based on item parameters.

One key to the patterns shown by the PPS variables was suggested by Fryer and Levitt's reference to skills that "are mastered by virtually all students in the grade." By the fifth grade, a number of the lower-level PPS variables show no group difference or a trivial one because of ceiling effects both among Whites and Blacks (Table 1).

Table 1. Descriptive Statistics, Difference in Mean Probability, and Logit of Difference in Mean Probability, Round 6 Proficiency Probability Scores, for Whites and Blacks

	Whites (N = 6470)		Blacks (N = 1275)		Difference in mean probability	Logit of difference in mean probability
Variable	Mean	Min - Max	Mean	Min - Max		
c6r3mpb1	1.00	1.00 - 1.00	1.00	1.00 - 1.00	0.00	--
c6r3mpb2	1.00	1.00 - 1.00	1.00	1.00 - 1.00	0.00	1.65
c6r3mpb3	1.00	0.99 – 1.00	1.00	0.99 – 1.00	0.00	1.58
c6r3mpb4	1.00	0.56 – 1.00	0.99	0.56 – 1.00	0.01	1.55
c6r3mpb5	0.96	0.01 – 1.00	0.84	0.01 – 1.00	0.12	1.56
c6r3mpb6	0.85	0.00 – 1.00	0.54	0.00 – 1.00	0.31	1.56
c6r3mpb7	0.55	0.00 – 1.00	0.21	0.00 – 1.00	0.34	1.54
c6r3mpb8	0.18	0.00 – 1.00	0.03	0.00 – 1.00	0.14	1.86
c6r3mpb9	0.02	0.00 – 0.93	0.00	0.00 – 0.38	0.02	2.06

The remaining question is that of possible differences in the amount of divergence between Blacks and Whites across the remaining PPS categories. In Round 6, the gap in the means of the PPS variables increases as the skills become more difficult until the probabilities become very low, at which point the mean differences necessarily shrink (column 6 in Table 1). However, this appears to be largely a reflection of the scale, that is, proportions. Logits of the means show an essentially constant White-Black difference for all PPS variables except for the most difficult two (i.e., c6r3mpb8 and c6r3mpb9), which show modestly larger group differences (column 7 in Table 1).

However, even when the PPS variables, rescaled appropriately, show a greater disparity for some skills than others, this provides only limited information about the characteristics of the Black-White gap because of the manner in which the PPS variables are created. As the description above indicates, θ is a sufficient statistic for PPS scores—that is, the latter contain no information about individuals independent of θ. To make this concrete, Figure 1 displays the mapping of one of the third-grade mathematics PPS scores, "place value," onto the ECLS-K T scores, which are a simple linear transformation of θ. The plots on the top and right-side borders are kernel smooths of histograms of the T scores and PPS scores, respectively. This PPS had a weighted mean value of .39 (NCES, 2004, p. 3-17). The T scores show a roughly normal distribution, as one would expect, while the PPS shows a bimodal distribution, with most students showing either a very high or a very low probability of mastery of this small cluster of items. What is important for present purposes is the complete lack of scatter: the mapping of T scores to PPS scores is one-to-one. Therefore, every single student with a given estimate of θ receives the identical PPS score, regardless of any other characteristics, including race.

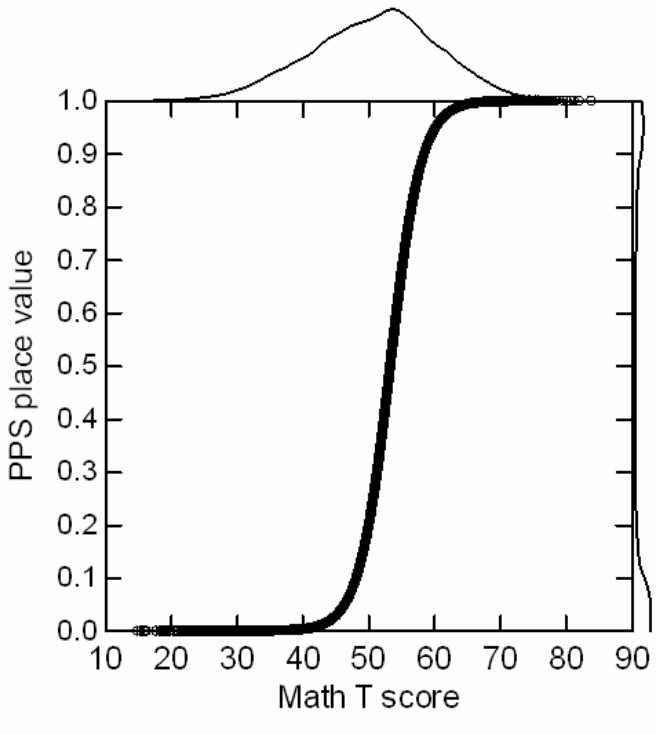

Figure 1. Mapping of the place value proficiency probability score to the mathematics T score, third grade.

The utility of the proficiency probability scores is therefore to show differences in the trajectories of performance in skill clusters as a function of θ, not to differentiate among students within a given level of θ. They are useful for characterizing the Black-White gap only in a very limited sense. Because the mapping of θ to PPS scores differs from one PPS to another—that is, from one skill cluster to another—any two groups that differ in location on the θ scale will show a larger gap on some PPS variables than on others. This variation in performance differences across PPS variables may be useful in characterizing the nature of the group difference in performance. Similarly, if two groups diverge in location as they grow older, as Fryer and Levitt (2004b) found for Blacks and Whites, the changes in their performance will vary across PPS variables, and this too may be descriptively useful. However, this information reflects only differences in location on the θ scale, *not race or any other individual characteristics other than* θ. If a subsample of Whites were drawn with the same θ distribution as the total sample of Blacks, this subsample would show precisely the same differences from the total White sample on all PPS variables as did Blacks.

To identify differential skill growth across racial groups requires a method that will differentiate between skill differences that are a function of θ, independent of race, from skill differences that are specific to race.

Research Questions

This study explored changes in the characteristics of the Black-White achievement gap by applying differential item functioning (DIF) analysis, using θ (more precisely T scores, which are a simple linear transformation of θ) as the matching criterion. That is, we evaluated racial differences in item-level performance conditional on θ. DIF analysis was conducted on all scaled items administered in the ECLS-K from kindergarten through third grade, subject to limited exclusions noted below. If specific skills contribute disproportionately to the growth of the mean difference across age, this should be reflected in DIF. Specifically, items contributing more than average to an increase in the gap should show greater DIF disfavoring black students at later ages.

For addressing the questions posed by Fryer and Levitt (2004a, 2004b), it is essential to distinguish between DIF and simple group differences in item-level performance. Simple group differences in item-level performance pose the same problem as the PPS analyses discussed above: they conflate differences in the skills

mastered by students at different levels of overall proficiency with differences in the skills mastered by students of different races, independent of overall proficiency. DIF analysis solves this problem, but at a cost: because the mean difference in θ has been removed, DIF statistics do not directly characterize the overall gap between black and white students.

Specifically, we explored several questions about DIF:

1. How common is DIF?
2. Are there consistent patterns across age in the distribution DIF?
3. Are there identifiable characteristics of items showing DIF, such as item difficulty or the characteristics of the skills required by the items? In particular, are there differences in the DIF shown by items tapping elementary cognitive skills (such as spatial visualization), knowledge commonly acquired outside of school, or items tapping skills commonly taught in school?

Data

The data used for this study are the complete ECLS-K longitudinal database, excluding Round 3 (the Fall Grade 1 sample, which was relatively small) and the very small Grade 2 bridge study. The analysis was therefore conducted on Rounds 1, 2, 4, 5, and 6. We analyzed only mathematics. The number of records with valid mathematics data ranged from 13,288 in Round 1 (Fall of Kindergarten) to 7,745 in Round 6 (Spring of Grade 5).

The ECLS-K secure database includes a variety of test scores, almost all of which, like the PPS scores, are transformations of θ. However, the database includes no information about performance at the level of individual test items. Accordingly, we obtained copyright permission from all authors or firms whose materials were used in the construction of the ECLS-K assessments, after which the NCES provided us with item-level performance data.

The ECLS-K assessments were adaptive. All students in a given round were administered a short routing test, and on the basis of their performance on the routing test, they were assigned to one of three additional test forms varying in difficulty (i.e., low, middle, and high forms). These forms were linked within and across rounds using the IRT scaling model.

Our analysis included all forms (i.e., routing, low, middle, and high forms) and all items within forms in the five rounds noted above, except when the data were too sparse. Two forms were deleted from the analysis because of the small number of valid scores for Black students: the Round 1 high form, with scores for 39 Black students, and the Round 6 high form, with scores for 120 black students. This left us with a total of 18 forms for analysis (5 rounds x 4 forms – 2 high forms). Several items were dropped when it was found that the 2-by-2 table of race by correct/incorrect had one or more very sparse cells.

Methods

DIF was analyzed using logistic discriminant function analysis (Miller & Spray, 1993), which a binary group variable, *G*, is regressed on a measure of performance on the entire test, *X*, and performance on the item being analyzed, *I*:

$$(1) \quad G = \frac{1}{1+\exp\left[-\alpha_i - \beta_{1i}X - \beta_{2i}I_i - \beta_{3i}X \cdot I_i\right]}$$

$$(2) \quad G = \frac{1}{1+\exp\left[-\alpha_i - \beta_{1i}X - \beta_{2i}I_i\right]}$$

$$(3) \quad G = \frac{1}{1+\exp\left[-\alpha_i - \beta_{1i}X\right]}$$

In this case, *G* is Black versus White, and *X* is the ECLS T score, a linear transformation of the IRT θ to a within-round mean of 50 and standard deviation of 10. The interaction term in model 1 is a test for non-uniform DIF, and the term $\beta_{2i}I_i$ in model 2 is a test of uniform DIF in the absence of non-uniform DIF. Model 3 is the no-DIF case: performance on an item provides no information about group membership beyond that contributed by the total score. All items were initially assessed for both uniform and non-uniform DIF. However, instances of non-uniform DIF were relatively rare, and some of these instances appeared to reflect very small numbers of students in one cell. Therefore, the analysis reported here was restricted to uniform DIF.

DIF was screened using the common criteria suggested by Zieky (1993), but considering only the absolute size of the DIF estimate. The more lenient criterion is delta=1, corresponding to an odds ratio of 1.53 and a z-score of 0.25. The more stringent criterion is delta=1.5, corresponding to an odds ratio of 1.89 and a z-score

of 0.375. Statistical significance, adjusted for design effects, was applied as a separate criterion.

The ECLS-K sample is clustered. Correction of standard errors without jackknifing was not possible in many instances because some strata included a single PSU. The variance estimator, when there is one stage with clustering and stratification but without a finite sample correction, is as follows:

$$\hat{V}\left(\hat{Y}\right) = \sum_{h=1}^{L}\left(\frac{n_h}{n_h - 1}\right)\sum_{i=1}^{n_h}\left(y_{hi} - \bar{y}_{hi}\right)^2$$

where $h=1...L$ indexes strata, h_i indexes a PSU i within stratum h, y_{hi} is the mean across people j within PSU i, and $\bar{y}_{hi}$ is the mean of all y_{hi} across all PSUs within stratum h. Thus, the variance for strata with a single sampled PSU cannot be estimated. Given the large number of analyses (three logistic regressions per item), jackknifing would have been prohibitive. Therefore, we estimated a design effect by comparing a number of analyses conducted with and without adjustment for clustering in samples that included no strata with a single sampled PSU (using the Stata Version 9 logit and svy logit procedures). On this basis, we estimated a design effect (DEFT) of 1.25. All regressions were conducted without correction for clustering, and this design effect was applied *post hoc* to the Wald statistics. All regressions were weighted using the ECLS-K design weights.

Omitted Items

The ECLS-K assessments follow a common convention in distinguishing between omitted and not-reached items. Items following the last one a student attempts are classified as not reached, and these items were treated as if they were not administered in scaling the ECLS-K. Items that were not attempted by the student, but that are followed by items to which the student did respond, are classified as omits. Omitted items are particularly problematic for many kinds of scoring because of uncertainty about the probability of a correct response had the student responded. In the case of multiple-choice items or other formats that permit guessing, treating such items as incorrect is likely to understate proficiency because even students with no relevant knowledge would have a non-zero probability of answering correctly by means of guessing. Often omitted items are scored as fractionally correct for this reason. IRT scaling, such as that used in the ECLS-K direct cognitive assessments allows the estimation of scores for students despite

non-reached and omitted items by utilizing the information contained by the pattern of responses to the other items.

For our analyses, no such straightforward solution to missing data exits. DIF analysis of the sort we conducted requires a separate set of logistic regressions for every item, with item responses on the right-hand side. The only choices were to delete observations with missing data, or to impute some value for them. Deleting missing data could erode the representativeness of the sample. Moreover, the impact could be differential across items, and apparent variation in DIF across items could be confounded with between-item differences in sample characteristics.

In practice, however, this problem was minor, in part because of the manner in which the ECLS-K assessments were administered. Most of the materials were presented by an administrator, who was allowed to redirect the child's attention to any given item. Administrators rated each child's behavior and apparent motivational level at the end of the assessment session. "Low" motivation was described as a child "frequently saying I don't know without even trying, consistent encouragement needed," while "very low" motivation was described as "child doesn't…attempt many items, even with encouragement." The proportion of students rated as exhibiting very low motivation ranged from 1.7 percent in kindergarten to 0.9 percent in the fifth grade, and those rated as showing low motivation dropped from roughly 10 to 6 percent from kindergarten (Pollack, Najarian, Rock, & Atkins-Burnett, 2005, Table 4-1). Consistent with this, the proportion of items with omit rates greater than 1 percent was small. We identified a total of 23 such items across the 12 level-by-grade combinations. Omits were particularly rare in the mid-level and high-level forms. We identified no more than one such item in the high form in any grade, and no more than two in the middle form in any grade. These items were more common in the low form, but even in those forms we identified only three or four items in each grade.

Given these considerations, we treated omits as incorrect responses. The apparently high motivational level, combined with the fact that 'omitting' an item required a response to an administrator, not simply skipping to another item in a booklet, it seemed likely that in most cases, a student who said "I don't know" really did not know. Moreover, the small number of omits precludes substantial bias of our results from this decision.

Prior Screening for DIF

In one respect, the ECLS-K assessment, like most carefully constructed tests used in large-scale achievement testing, is poorly suited to answering the question posed by Fryer and Levitt (2004b) and this study: sizable DIF between racial and ethnic groups is generally avoided by design. The data are screened for DIF, and items displaying large DIF are often deleted from the operational assessment, or are dropped from the data before final scaling. This is justified both by a desire to avoid possible item bias and because of the use of unidimensional scaling models. When such techniques are applied to longitudinal data, the result can be a conservative bias in which particular skills that contribute disproportionately to divergence of performance on the latent trait are removed in whole or in part from the tested subset of the domain. However, in the case of the ECLS-K, this prior screening for DIF was inconsequential; we identified only three mathematics items deleted from scaling because of DIF. Other limitations of the ECLS-K data for the purposes of this study are discussed in the final section of this paper.

The Consistency of the Black-White Difference

Many discussions of the Black-White performance gap seemingly rest on an assumption that this differential is a reasonably uniform phenomenon that is not sample-dependent. This assumption is generally tacit, and it is not clear whether those participating in the debate would accept the assumption were it made explicit. However, both the language many scholars use to describe the gap and the methods used to analyze it often imply a unitary gap. For example, Fryer and Levitt refer to "*the* pattern of Black skill acquisition" (2004b, p. 19, emphasis added). However, it is certainly plausible that there are numerous different patterns of Black-White differences – for example, that the pattern of development varies by region or social class. This tacit assumption has important implications for the analysis and interpretation of changes in the Black-White gap over time.

Therefore, as a preliminary step before analyzing the ECLS-K data, we examined the cross-state consistency of the Black-White difference in fourth-grade mathematics on the 2000 National Assessment of Educational Progress. We selected three states that have large Black samples (to stabilize our estimates) and markedly different levels of performance: Michigan, Alabama, and Arkansas. In each state, we performed a logistic discriminant function analysis of uniform DIF, contrasting Blacks as the focal group to Whites as the reference group. The matching criterion

was NAEP scale scores. We compared the results across the three pairs of two states.

Unlike typical DIF analysis, in which the goal is to identify items for potential exclusion because of bias or other factors, our interest was to explore the consistency of the entire distribution of DIF statistics across the 143 items in the assessment. DIF analysis removes between-state variation in the size of the Black-White gap from consideration, leaving only information on the extent to which items favor or disfavor Blacks who are matched with Whites on overall proficiency. To the extent that the Black-White gap is qualitatively consistent across states independent of overall proficiency, the same items should disfavor or favor Blacks across states. Therefore, the correlation of the DIF statistics across states is a simple measure of the qualitative consistency of the Black-White gap.

All of the three of the pairwise plots were dominated by a small number of outlier items with very large DIF statistics in one or both states that obscured the more general relationship. This is illustrated by the plot of Arkansas and Michigan in Figure 2. Therefore, we removed 3 to 6 outliers from each pairwise comparison and reexamined the plots to ensure that Pearson correlations were an appropriate measure of consistency. Figure 3 shows the plot of Arkansas and Michigan with outliers removed. After removing outliers, all three of the pairwise correlations were quite small, ranging from .25 to .36 (Table 2).

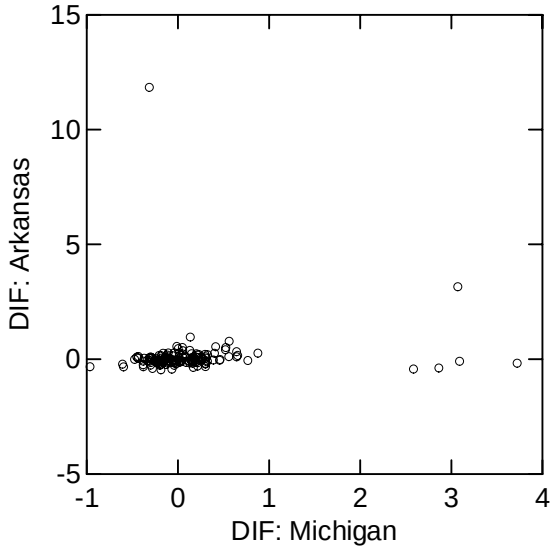

Figure 2. Scatterplot of DIF in Arkansas and Michigan, all items.

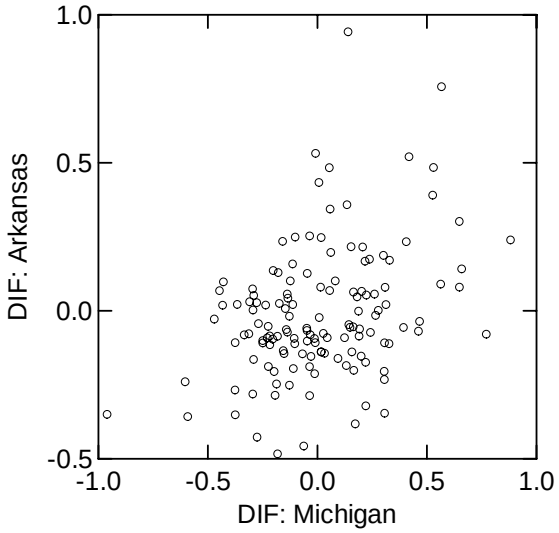

Figure 3. Scatter plot of DIF in Arkansas and Michigan, six outliers removed.

Table 2. Between-state Correlations of DIF After Removing Outliers

	Alabama	Arkansas
Alabama	1	
Arkansas	0.33	1
Michigan	0.25	0.36

While these findings represent only three states, they suggest a need to be cautious in interpreting the Black-White gap because of potentially major variations in the nature of the gap across samples. Inferences based on a nationally representative sample such as ECLS-K should be restricted to the national population, as the characteristics of the gap could be substantially different in other populations, e.g., in particular regions or states. We return to the question of the robustness of findings based on the ECLS-K in the final section of the paper.

Results

To interpret the results described here, it is necessary to keep in mind two characteristics of DIF.

First, DIF statistics do not indicate simple performance differences, either favoring or disfavoring Black students. Rather, they are group differences in item-level performance conditional on a measure of proficiency. In these analyses, the matching criterion is the IRT estimate of overall proficiency, θ, and therefore, DIF indicates that an item is easier or more difficult for Black students than for Whites, relative to other items in the test. Most items that showed DIF favoring Black students were nonetheless more difficult for Blacks than for Whites. For example, in the Round 1 routing form, the item that showed the strongest DIF favoring Blacks—a counting item—was nonetheless slightly harder for Blacks than for Whites (weighted p-values of .54 and .61, respectively). A few items did show simple differences in favor of Black students (for example, a number-identification item in the Round 1 mid-level form), but this should not be inferred from DIF statistics favoring Black students.

Second, given that the method used here matches students on total score (θ), if DIF appears at all, it must necessarily appear in both directions, that is, both disfavoring and favoring Blacks. Imagine a case in which a test comprises 40 items, all but one of the items are consistent in terms of the performance difference between the two groups, and the remaining item is differentially difficult for Blacks. This anomalous item affects the estimation of θ, making the θ estimates for Blacks

somewhat lower than they would be if the item were omitted. Thus, the effect of matching students on θ is to slightly attenuate the DIF observed for the anomalous item, while generating offsetting DIF favoring Blacks on the other items (In some DIF models, the average of the DIF estimates is constrained to zero; in our models, the averages are only approximately zero.). In this hypothetical example, the offsetting DIF would be spread over 39 items and would therefore be very small in comparison, and it would escape detection using conventional screening criteria. This illustrates that the number of items identified as showing DIF need not be symmetrical. Nonetheless, both the attenuation and the offsetting DIF always exists, and it affects the interpretation of the DIF coefficients.

How Common is DIF?

The frequency of DIF varied over the 18 forms we analyzed. The frequency of DIF using the more lenient screen of delta=1 was reasonably consistent across the low, middle, and high forms, in Rounds 1 through 4, ranging from 27 to 35 percent of items (Table 3). In the two kindergarten waves (Rounds 1 and 2), DIF was abundant in the routing forms, affecting 56 and 69 percent of items.

If the growing Black-White gap were accompanied by greater differentiation by skills, as Fryer and Levitt (2004b) suggest, one might expect an increase in the frequency of DIF as children age. Items tapping some skills (they suggested those measuring lower-order skills) would become easier for Blacks, conditioning on θ, while others would become relatively harder.

What we find is the reverse: DIF is *less* frequent in Rounds 5 and 6, with no items in the Round 5 routing form showing DIF (using the more lenient screen) and only one item of 17 showing DIF in the Round 6 low form.

Table 3. Proportion of Items Showing DIF, Lenient Screen, After Excluding Items with Small Cell Counts

	Round				
Form	1	2	4	5	6
Routing	0.56	0.69	0.31	0.00	0.22
Low	0.29	0.29	0.29	0.28	0.06
Middle	0.35	0.26	0.35	0.21	0.33
High	—	0.27	0.29	0.21	—

Are There Trends in the Variance of DIF?

By the same token, if the growing Black-White gap reflects more rapidly growing differences in some skill areas than others, one might expect that the variance of the DIF statistics for items showing DIF would increase, as performance in some areas falls progressively farther behind than skills in other areas.

Although the standard deviation of the DIF statistics varied somewhat by form and round, there was no clear pattern in this variation. Calculated across all forms within round, the standard deviation of the DIF statistic for items exceeding the more lenient criterion ranged from .54 to .63 in all rounds other than Round 4, with no clear trend across rounds. Round 4 was anomalous, with a standard deviation across forms of .73, a result of a very large standard deviation (.97) within the Round 4 routing form.

Is DIF Correlated with Item Difficulty?

If Black students are falling farther behind in higher skills than in more basic skills, one might expect a progressively stronger relationship between item difficulty and DIF as students' age. In this case, item difficulty was indexed by the logit of the item p-value (to avoid potential for bias in the correlation because of the compression of the p-value scale at the extremes), and the regression analyses were structured so that a negative DIF coefficient indicated conditional performance disfavoring Blacks. Therefore, a tendency for Blacks to perform worse on more difficult items would be reflected in a positive correlation. Correlations were calculated only for items that manifested DIF in excess of the more lenient screening criterion.

These correlations were inconsistent in sign and varied markedly across forms and rounds (Table 4). The correlations were all positive for the routing forms and negative for the high forms, but the variability among the correlations makes this

pattern suspect. There was no consistent tendency for the correlations to become more positive across age groups.

Table 4. Correlation Between DIF and the Logit of Item p-values

	Round				
Form	1	2	4	5	6
Routing	0.50	0.62	0.99	—	0.10
Low	-0.12	-0.70	0.81	0.46	—
Middle	-0.55	-0.80	-0.42	0.51	-0.21
High	—	-0.34	-0.22	-0.44	—

What are the Characteristics of Items Showing DIF?

Across all forms and age groups, the items in the ECLS-K mathematics test are diverse. Many of the items assess material that is a focus of instruction in the primary grades, such as basic arithmetic operations. A number of items, particularly in the youngest age groups and lower-level forms, tap skills that one might expect most children to learn outside of school, such as counting and number recognition. A modest number might be classified as measuring basic cognitive skills, such as comparing the length of figures, estimating the number of items in a picture, and distinguishing among dimensions of figures (such as size and shape) in a classification task.

These distinctions are not clear-cut. Many test items require multiple skills, and it is not always apparent from inspection what skills children will bring to bear on answering one. Moreover, some schools devote time to teaching skills that some students learn at home (such as counting), and some students learn outside of school material that is the focus of direct instruction in school. Nonetheless, variation in DIF across these categories of items may offer clues about changes in the Black-White difference as children age.

We examined the content and demands of every item in all forms and rounds that showed DIF, either favoring or disfavoring Black students, using the more lenient screening criterion of delta ≥ 1. We are prohibited from revealing the specific content of any items, but we characterized each of these items in more general terms sufficient for our purposes—for example, "identify a specific two-digit number," "two-digit addition without carrying, presented without pictorial representation,"

and "spatial visualization: how many of each shape required to construct a given polygon."

We found no clear patterns that were consistent across all rounds of data, and within rounds, there were exceptions to every pattern we identified. However, a number of patterns appeared in portions of the database that warrant further exploration.

In the fall of kindergarten (Round 1), eight items were identified as showing DIF favoring Black students[2] (Items appearing in more than one form in a round are counted only once.). Six of the eight were counting or number-identification items. The two exceptions were addition problems in the mid-level form. Thirteen items showed DIF disfavoring Blacks. Seven of these tapped skills taught in school; six were addition items, while a seventh required understanding a bar graph. The remaining six were diverse and included estimating the number of items in a picture, generalizing the notion of half of a shape, identifying a simple polygon (one typically learned outside of school), and comparing the length of two objects. Taken together, these hint that at kindergarten entry, Black students are more disadvantaged on mathematics content taught directly in school, but this pattern is not consistent.

Results from the spring of kindergarten (Round 2) also showed some indication that Black students are less disadvantaged on material taught outside of school, but this pattern seemed to be a function of performance level. Across forms, a total of nine items showed DIF favoring Black students. All four in the routing form and both in the low form (one of which was also administered in the routing form) were counting or number-identification items. A total of 10 items in these two forms showed DIF disfavoring Blacks. These were diverse, but four tapped arithmetic operations taught primarily in school. In contrast, in the high form, both items showing DIF favoring Blacks and four of the six disfavoring Blacks measured material taught in school—arithmetic and interpretation of a bar graph.

By the spring of the first grade (Round 4), the pattern was less clear. Once again, simple counting and number identification items in the routing and low-level forms showed DIF favoring Black students, but these items were so easy for both groups that the differences are not especially informative. All three of these items in

[2] Recall that the Round 1 high form was dropped from the analysis because of the small number of Black students administered that form.

the routing form—all number-recognition items—had weighted p-values in both groups of .90 or more. Leaving these extremely easy items aside, there was no obvious pattern in the items showing DIF favoring and disfavoring Blacks. Both material taught in school, such as addition, and other skills are found in both sets of items.

The characterization of some items as reflecting material typically taught in school requires a caveat: the ECLS-K includes some items that are most often taught in school at a later grade than the one in which the assessment was administered. In a few cases, we found DIF disfavoring Black students on items of this sort—for example, a multiplication item in Round 2 (fall of kindergarten) and a division problem in Round 4 (spring of the first grade). In these cases, the causes of the performance difference may lie outside of school, notwithstanding the content of the items.

In the spring of the third and fifth grades, no clear pattern appeared. In the third grade, arithmetic items appeared among those favoring both groups. A number of the items disfavoring Blacks tapped geometric skills and spatial visualization, such as showing an understanding symmetry and indicating which of several shapes is constructed with only straight sides. The relatively few items showing DIF in the fifth-grade sample measured skills taught in school. For example, one item that showed DIF strongly favoring Black students asked students to find the length of one side of a figure from the area and the length of another and required dividing a 3-digit number by a two-digit number. Another in the same form that showed substantial DIF disfavoring Black students was a word problem that required calculating the perimeter of a square and subtracting single-digit numbers.

Discussion

Fryer and Levitt (2004b) raised the troubling possibility of a differential growth in the gap between White and Black students, with Blacks falling farther behind on more advanced skills. However, the ECLS-K data do not provide evidence of this differential growth.

Fryer and Levitt's concerns were aroused by variations in the Black-White difference across PPS. These variations, however, do not provide evidence of differential divergence between Blacks and Whites. First, the variations they found in the Black-White differences across PPS variables appear to reflect the nonlinear

relationships between θ and the PPS variables and ceiling effects in the latter. But even if this were not the case, variations in the gap across PPS variables, by construction, only re-express the overall mean difference on θ and can tell one nothing about qualitative differences in the performance of Blacks and Whites at similar levels of proficiency. The concern raised by Fryer and Levitt cannot be addressed using these particular scores, and the ECLS-K database does not provide scores more suited to this use. Therefore, in the case of the ECLS-K data, their concern must be explored using item-level data.

Patterns of differential item functioning in the ECLS-K data also provide little evidence of the differential growth in the gap that Fryer and Levitt raised. DIF is less common in the later rounds than the early ones, and the variance of DIF does not change consistently with age. The relationship between DIF and item difficulty was inconsistent and did not increase with age. We found only one potentially relevant pattern in the characteristics of the items showing DIF favoring and disfavoring Blacks: in both kindergarten waves, there was some tendency for Black students to be at less of a disadvantage in simple skills often learned outside of school (e.g., number recognition) than on skills primarily taught in school (e.g., arithmetic items). This pattern appeared in the routing and low forms in the spring of first grade also, but in this case, the relevant items were answered incorrectly by so few students of either race that their import is questionable. We found no obvious pattern in the later grades.

The lack of evidence of differential growth in the ECLS-K, however, does not necessarily indicate that Fryer and Levitt's concern is misplaced, and that the growth in the Black-White gap is consistent across different types of skills. The ECLS-K was not designed to address the question Fryer and Levitt posed, and it is not well suited to this purpose. This is a common dilemma: assessment designs that are desirable for one purpose are often undesirable for others. For example, some of the methods used in the National Assessment of Educational Progress to improve aggregate estimates of performance make the assessment unsuitable for providing individual scores. Although the ECLS-K assessments were deliberately designed to tap a wide range of skills indicated by prior research to be important, they were not to discern differential growth across those skills. This is apparent from numerous aspects of the assessment design, including the imposition of a single unidimensional scoring model across types of skills and ages, the prescreening for DIF, the lack of subscales representing categories of knowledge and skill that might be expected to show

differential growth patterns, and the use of the PPS variables—an adaptation of the IRT number right true score—to obtain qualitative information about performance at different levels of θ. The result is a large risk of a Type II error when the data are used to address the question Fryer and Levitt raised. This is not a criticism of the ECLS-K, but simply a recognition that one assessment design cannot serve all masters.

In the ideal case, an assessment designed specifically to address Fryer and Levitt's concern would look very different. One reasonable approach would be to construct the test around clusters of skills that one might expect to show interactions between population group and age, and to report performance on those clusters. One reporting option would be the scaling procedure used with the National Assessment, in which *a priori* subsets of the data, such as "number sense, properties, and operations," and "measurement," are scaled separately; and the overall proficiency estimates are then calculated as a weighted composite of the subscales. It would be necessary to represent these clusters at different ages, avoiding ceiling and floor effects to the extent feasible. It would still be necessary to screen items for DIF, but items that show large DIF might be retained. Some of these choices, however, such as selection of certain types of items for inclusion in later waves, might make the assessment less useful than the ECLS-K for some of the purposes for which the ECLS-K tests were designed.

In practice, we are unlikely to have any data approaching the ideal for addressing the important concern that Fryer and Levitt raised because of the effort and money required to create representative longitudinal achievement data. Therefore, the primary question for analysts is how best to use databases created for other purposes to address these questions. The results of this study make it clear that using these extant databases for this purpose will require careful attention to the construction and limitations of the available tests, and the scores used to report them.

Another reason for caution in addressing the question Fryer and Levitt raised is the risk of results that are not robust across databases. For example, Rock and Stenner (2005) noted that the ECLS-K assessment shows a considerably smaller initial Black-White difference than many other tests. Murnane, Willett, Bub, and McCartney (2006) showed that several of Fryer and Levitt's findings do not replicate when using the National Institute of Child Heath and Human Development Study of Early Child Care and Youth Development (NICHD SECCYD): the gap in the latter

database grows larger with age, and a substantial part of the gap at young ages is not predicted by the covariates used by Fryer and Levitt (2004a, 2004b).

This variation across tests is to be expected, and to some degree, it would persist even if we had assessments specifically designed to address Fryer and Levitt's question. Tests are only small samples of performance, and findings can vary substantially across different acceptable samples. This variation can be both particularly sizable and especially difficult to deal with when tests are used to measure growth. Substantial variations can arise not only because of differences in choices about content and difficulty, but also because of differences in the methods used to scale the tests and link them across age groups. These effects can be large and are not entirely predictable. Moreover, because the characteristics of the Black-White gap can vary among subpopulations, differently constructed national probability samples might yield appreciably different results even if a single test were administered to them. Replication using multiple data sources is therefore especially important, and in its absence, interpretation should reflect the substantial likelihood of test-specific findings.

References

Fryer, R. G., and Levitt, S. D. (2004a). Understanding the black-white test score gap in the first two years of school. *The Review of Economics and Statistics, 86*(2), 447-464.

Fryer, R. G., and Levitt, S. D. (2004b). The black-white tesT score gap through third grade. Draft available from the fist author at www.economics.harvard.edu/faculty/fryer/papers/fryer_levitt_ecls2.pdf, last accessed on February 11, 2007. (Forthcoming in *American Law and Economic Review* (special issue on Brown *v.* Board of Education).

Miller, T. R., & Spray, J. A. (1993). Logistic discriminant function analysis for DIF identification of polytomously scored items. *Journal of Educational Measurement, 30*, 107-122.

Murnane, R. J., Willett, J. B., Bub, K. L., and McCartney, K. (2006). Understanding trends in the black-white mathematics achievement gap during the first years of school, *Brookings-Wharton Papers on Urban Affairs,* 97-135.

National Center for Education Statistics (2004). *User's Manual For The ECLS-K Third Grade Public-Use Data File And Electronic Code Book*. Washington, D. C.: Author (NCES 2004-0001).

Pollack, J. M., Najarian, M., Rock, D. A., and Atkins-Burnett, S. (2005). *Early Childhood Longitudinal Study, Kindergarten Class of 1998–99 (ECLS-K) Psychometric Report for the Fifth Grade.* Washington, D. C.: National Center for Education Statistics (2006-036).

Rock, D. A., and Pollack, J. M. (2002). *Early Childhood Longitudinal Study-Kindergarten Class of 1998–99 (ECLS–K), Psychometric Report for Kindergarten Through First Grade.* Washington, D. C.: National Center for Education Statistics (Working Paper 2002-05).

Rock, D. A., and Stenner, A. J. (2005). Assessment issues in the testing of children at school entry. *The Future of Children, 15*(1), 15-34.

Zieky, M. (1993). Practical questions in the use of DIF statistics in test development. In P. W. Holland and H. Wainer, *Differential Item Functioning*, Hillsdale, N.J., 349-364.