Assessment Issues in the Testing of Children at School Entry

Donald A. Rock and A. Jackson Stenner

Summary

The authors introduce readers to the research documenting racial and ethnic gaps in school readiness. They describe the key tests, including the Peabody Picture Vocabulary Test (PPVT), the Early Childhood Longitudinal Study (ECLS), and several intelligence tests, and describe how they have been administered to several important national samples of children.

Next, the authors review the different estimates of the gaps and discuss how to interpret these differences. In interpreting test results, researchers use the statistical term "standard deviation" to compare scores across the tests. On average, the tests find a gap of about 1 standard deviation. The ECLS-K estimate is the lowest, about half a standard deviation. The PPVT estimate is the highest, sometimes more than 1 standard deviation. When researchers adjust those gaps statistically to take into account different outside factors that might affect children's test scores, such as family income or home environment, the gap narrows but does not disappear.

Why such different estimates of the gap? The authors consider explanations such as differences in the samples, racial or ethnic bias in the tests, and whether the tests reflect different aspects of school "readiness," and conclude that none is likely to explain the varying estimates. Another possible explanation is the Spearman Hypothesis—that all tests are imperfect measures of a general ability construct, *g*; the more highly a given test correlates with *g*, the larger the gap will be. But the Spearman Hypothesis, too, leaves questions to be investigated.

A gap of 1 standard deviation may not seem large, but the authors show clearly how it results in striking disparities in the performance of black and white students and why it should be of serious concern to policymakers.

www.future of children.org

Donald A. Rock is with the Educational Testing Service. A. Jackson Stenner is chairman and CEO of Metametrics Inc. The authors thank Timothy Taylor, managing editor of the *Journal of Economic Perspectives*, for extensive contributions to the improvement of this article.

In study after study over the past ten years, researchers from a variety of fields using a variety of testing approaches have consistently found a gap between the readiness of white children and the readiness of black and Hispanic children to enter school. The concept of "readiness," however, has no obvious unit of measurement. Lacking such a tool, researchers have used a range of tests to measure different dimensions of the skills and behaviors—word comprehension, reading, math, the ability to sit still—that make a child "ready" to enter school. If a test is accurate, a child's score can be used to predict his future success or achievement. A student who is measured as more "ready" should have greater success in meeting the demands or challenges of school.

We begin by introducing the main tests that researchers have used to measure the readiness gap for children entering kindergarten. We then review the range of evidence that these studies have produced about the size of the gap. Perhaps not surprisingly, the evidence on the size of the gap differs somewhat from one study to the next, and we discuss how to interpret these differences. The articles that follow in this volume explore possible underlying causes of the readiness gap: family and neighborhood characteristics, genetic differences, neuroscience and early brain development, prenatal experiences, health of young children, and differences in parenting, child care, and early education.

How Can Readiness Be Assessed at Kindergarten Entry?

Many experts in the field suggest that it is difficult if not impossible to assess a child's academic performance accurately before age six.[1] Some studies have argued that scores on preschool or kindergarten readiness tests can predict no more than 25–36 percent of the variance in performance in early grades.[2] Even if these estimates are correct, predicting 25 to 36 percent of the variance in later achievement is not to be sneezed at. But we believe that readiness tests have improved substantially in the past decade or so and that the new tests are likely to provide a better measure of readiness. For example, kindergarten test scores in the Early Childhood Longitudinal Study, which we discuss in more detail later, predict about 60 percent of the variance in performance at third grade. Before reviewing the main tests of kindergartners' readiness to enter school, we will consider some general characteristics of these tests and how they work.

Key Characteristics of Readiness Tests

Readiness tests may be given on a *group* or *individual* basis. Group tests can be less expensive to administer. But for kindergarten students, individual tests are preferred for several reasons. Administrators are more likely to be able to get and hold the attention and cooperation of a beginning kindergartner in a one-on-one setting than in a group.[3] Small children often enjoy the individual attention they get from the test administrator, which helps make the scores more accurate. In a longitudinal study, one scheduled to have multiple retestings over several years, a sizable share of the follow-ups might require one-on-one retestings because the children scatter as time passes. Starting with a group administration and then switching to one-on-one follow-ups could cause variance in the data that would be difficult to quantify. Individualized testing gives children the time they need to finish the assessment and thus gathers relatively complete information on each child. It also allows the test to be adapted to some degree to the abilities of each child.

Indeed, the best readiness tests are *adaptive*, which means that instead of asking every child identical questions, they give children harder questions if they do well on the early questions and easier questions if they do poorly early on. Operationally, a single test form is liable to be too hard for 10–20 percent of the children in the sample and too easy for another 10–20 percent. In this case, a "floor and ceiling" problem will arise: a substantial share of children will answer all or almost all of the questions correctly, while another substantial share will answer all or almost all incorrectly. Floor and ceiling problems are the bane of all readiness tests, because they mean that the distribution of test scores at the top and bottom of the scale will barely spread out at all, thus artificially narrowing the range of student achievement. Floor and ceiling problems also make it difficult to measure whether student scores change over time, because students clustered at the top or the bottom will often remain in this pattern when retested. An adaptive test avoids these problems and allows test scores to reflect the full range of student achievement. The main disadvantage of adaptive testing is cost. It is expensive to develop a large pool of items to cover the appropriate span of abilities and to ensure that a common procedure is followed in deciding when students will receive harder or easier questions. A computer-assisted test format is often helpful in advising the administrator which items are appropriate for each child. Indeed, adaptive tests for older, computer-knowledgeable children can be administered and scored in real time at a computer terminal.

A useful test must be *reliable*, which means that it will produce essentially the same results on different occasions. Reliability can be measured in three ways: retesting, equivalent form, and internal consistency. Retesting, or giving the same test over again to the same students, raises obvious questions about how students react to being given the same test twice. But retesting that produces dramatically different results would certainly raise some flags about reliability. The equivalent form approach uses two equivalent versions of a test, which can then be compared with each other. The internal consistency approach breaks a single test into parts, which are then compared with each other. For example, the results of all even questions might be compared with those of all odd questions (the "split half" test). Or more complex mathematical formulas might be used to split up the test in many different ways and then average those results (to generate a measure known as "coefficient alpha"). Whatever the measure, reliability is assessed along a scale from 0 to 1, where 1 means that a test has perfect reliability and gives exactly the same result each time and 0 means that the results from the test at one time are completely uncorrelated with the results the next time. A reliability score of .90 or above would represent high reliability; in the .80s, medium reliability; and in the .60s or .70s, low but ac-

Floor and ceiling problems are the bane of all readiness tests, because they mean that the distribution of test scores at the top and bottom of the scale will barely spread out at all, thus artificially narrowing the range of student achievement.

ceptable reliability. A reliability score in the .50s or lower would raise serious questions about the usefulness of the test.

Some have expressed concern that readiness tests may not be reliable for very young children because of their short attention spans. But individualized test assessment typically retains the attention of younger children. And very young children may be less likely than, say, seniors in high school to respond randomly or counterproductively to test questions. Brief descriptions of the major readiness tests used in this volume follow.

Peabody Picture Vocabulary Test—Revised

The Peabody Picture Vocabulary Test—Revised (PPVT-R) is an individually administered test of hearing (or receptive) vocabulary.[4] Each of two forms of the test contains five practice items and a set of 175 test items ordered by difficulty. An easy item might be "cat"; a difficult one, "carrion." All items appear in the same format: four black-and-white illustrations on a single cardboard stock plate. The examiner says a stimulus word aloud, and the examinee selects the image that best illustrates the meaning of the word. The test is adaptive, establishing a floor below which the examinee is assumed to know all word meanings, so that no more words below the floor are asked, and a ceiling above which the examinee is assumed to know no word meanings, so that no more words above the ceiling are asked. Testing typically takes between sixteen and thirty minutes, and the examinee typically responds to thirty-five to forty-five items.

The PPVT-R is a direct measure of vocabulary size. The rank order of item difficulties is highly correlated with the frequency with which the words are used in spoken and written discourse.[5] The PPVT-R was normed on a nationally representative sample of 4,200 children and 828 adults.

The PPVT-R is a widely used test, with good reliability. Reviews of its reliability conducted by the ERIC Clearinghouse on Assessment and Evaluation found split-test reliabilities ranging from the .60s to the .80s and test-retest reliabilities ranging from the .70s to the .90s.

For studies of kindergarten readiness, it is useful to test a large sample of children about whose families substantial background data are available. Two large samples of kindergarten children have taken the PPVT-R.

The first is the National Longitudinal Surveys, a set of U.S. government surveys that track people over time. The National Longitudinal Survey of Youth 1979 (NLSY79), began tracking a nationally representative sample of 12,686 young men and women aged fourteen to twenty-two in 1979. They were interviewed each year through 1994 and have been interviewed every other year since. The NLSY79 collected some data on children born to participants in the study, but in 1986 the survey began collecting much more intensive data about all children born to mothers in the NLSY79. The expanded survey administered the PPVT-R to children aged three to five (with some differences, according to the survey year).

A second large data sample of kindergartners is the Infant Health and Development Program (IHDP), a study funded by several private foundations and the U.S. government. It identified a group of 985 infants born with low birth weights in eight different cities in 1985 and tracked their development through 2000 using various tests, including the PPVT-R,

which was administered when the children were three and again when they were five.

The PPVT-R finds substantial differences in black-white readiness for kindergarten. For example, the vocabulary of black children in first grade is about half that of white first graders.[6] But two puzzles have arisen about PPVT-R findings. First, the PPVT-R often finds a larger black-white readiness gap than do other readiness tests. Second, studies using the PPVT-R on different samples of children have produced estimates of the black-white readiness gap that vary relatively widely, given that all involve nationally representative samples of children of comparable age using the same vocabulary measure. These issues will be discussed further below.

Wechsler Preschool and Primary Scale of Intelligence—Revised

The Wechsler Preschool and Primary Scale of Intelligence—Revised (WPPSI-R) is an individually administered test of general intellectual functioning for children from ages three to seven years and three months. It does not require reading or writing. The total battery contains many subtests: information, vocabulary, word reasoning, comprehension, similarities, block design, matrix reasoning, picture concepts, picture completion, object assembly, symbol search, coding, receptive vocabulary, and picture naming. Each subtest may include questions of several types. In the vocabulary subtest, for example, the child is asked to name an object (like a hammer) when she sees its picture and is asked to define a word when she hears it spoken. The test is not adaptive.

The components of the Wechsler test can be analyzed for individual patterns of learning, but readiness studies typically use an overall score based on all test components. Raw scores are converted into IQ scores with an average of 100. The IQ scores are scaled according to age groups, based on a nationally representative sample of 1,700 children in the relevant years. Reliability estimates for scores on the Wechsler test are high, typically ranging from the high .80s into the mid-.90s, depending on the kind of reliability that is reported.

The Wechsler test is often administered to learning-disabled or gifted children, but because such children are not randomly selected, their tests are of little use in researching the readiness gap. The WPPSI-R was, however, given to the children in the Infant Health and Development Program when they were five years old, thus providing a broad sample for analysis.

Stanford Binet

The Stanford-Binet Intelligence Scale, fourth edition (SB-IV), is a measure of "cognitive abilities that provides an analysis of pattern, as well as the overall level of an individual's cognitive development," according to the examiner's handbook.[7] The SB-IV is individually administered. It uses results from the vocabulary test to determine starting items for fourteen other tests, and thus is somewhat adaptive. Items in each of the fifteen tests are ordered as to difficulty. Raw scores are then converted to standard age scores for four cognitive areas: verbal reasoning, abstract/visual reasoning, quantitative reasoning, and short-term memory. The scores for each of these cognitive areas plus a composite standard age score (CSAS) are set to average 100 for each age group.

Reliability scores for the composite Stanford-Binet score as calculated by the internal consistency method (that is, dividing the test into parts and comparing the parts with each

other) range from .95 to .99. The reliability of the four cognitive area scores ranges from .80 to .97. These high correlations between the four area scores and the composite scores suggest that the cognitive area profiles are unlikely to provide reliable diagnostic information beyond that provided by the total score.

Like the Wechsler Preschool and Primary Scale of Intelligence, the Stanford-Binet test was also given as part of the Infant Health and Development Program (IHDP), in this case when the children were three years old, thus providing a substantial sample for analysis.

Woodcock–Johnson Psycho-Educational Battery—Revised

The Woodcock-Johnson—Revised (WJ-R) is an extensive battery of cognitive and academic achievement tests intended for people as young as two and as old as ninety-five. All tests are individually and adaptively administered. Seven abilities are tested and separately reported: fluid reasoning; comprehension/knowledge; visual processing; auditory processing; processing speed; long-term retrieval; and short-term memory. The standard battery then reports on four achievement clusters: broad reading, broad mathematics, broad written language, and broad knowledge. Two forms are available for the achievement tests. Raw scores are converted into grade and age equivalents.

The test manual reports high reliability. Internal consistency reliabilities for the cognitive and achievement clusters are all in the .90s. The shorter cognitive subtests that contribute to the seven ability scores have internal consistency reliabilities in the mid .70s to low .90s. The reliabilities of the achievement subtests that contribute to the broad achievement clusters are all in the high .80s and low .90s. Although alternate forms are available for the achievement clusters, these reliabilities are not reported in the manual.

Measures of Behavioral Readiness

The tests discussed so far have focused on academic achievement—that is, skills involving words, patterns, and the like. But another important dimension of readiness for kindergarten involves behavior, such as the ability to manage one's own emotions and to work well with others.

The Achenbach System of Empirically Based Assessment offers a range of diagnostic tests for behavior. The Child Behavior Checklist (CBCL), once called the Revised Child Behavior Questionnaire, asks mothers 120 questions about how frequently they have observed various behaviors in their children over the past six months. The checklist was given to the mothers of the children in the IHDP dataset when the children were aged three and five, thus providing a broad basis for analysis. The Achenbach checklist can be used to diagnose many behavioral issues, but it commonly focuses on two broad concerns: "internalizing" behavior, such as being too fearful, anxious, unhappy, sad, or depressed; and "externalizing" behavior, such as destroying objects or having temper tantrums.

The Behavioral Problems Index (BPI), derived from the Achenbach test and other tests of child behavior, asks mothers twenty-eight questions about the frequency of behaviors they have observed in their children over the past three months.[8] Results can be used to produce internalizing and externalizing scores. The test also produces an overall composite score, which is expected to average 100. The BPI was given to the women who entered the NLSY data set in 1979 after they had become mothers, when their children were at least four years old.

Yet another approach to assessing a child's behavioral readiness is direct observation. Often a parent and child are asked to play with some toys or to solve a puzzle together. The session is videotaped. Coders who have had extensive training watch the videotapes and rate behaviors like enthusiasm, persistence, frustration, and engagement.[9]

Early Childhood Longitudinal Study—Kindergarten Battery

Until the late 1990s, the study of school readiness rested on the few tests already described (all of which were originally developed for broader or different purposes than assessing school readiness) and on the two main sources of systematic data already mentioned, the NLSY and the IHDP. Without in any way disparaging the work done with these data, researchers felt that addressing a new source of nationally representative data with up-to-date instruments for evaluation might prove extremely helpful. The result was the Early Childhood Longitudinal Study, Kindergarten Class of 1998–99 (ECLS-K), administered by the National Center for Education Statistics. The new data set began with a base year fall assessment of 21,260 kindergartners who were then reassessed in the spring of their kindergarten year and in the spring of their first and third grade years.[10] Retests are also scheduled for the spring of fifth grade.

In an effort to move away from one-dimensional cognitive assessments toward multidimensional approaches, the ECLS-K evaluates kindergartners along several dimensions in tests that are individually administered and adaptive in design.[11] The direct cognitive assessments focus on three areas: reading, mathematics, and "general knowledge" (knowledge of the social and physical world). In addition, kindergarten teachers assess both cognitive progress and social or behavioral skills, and parents assess social competence and skills. Finally, children receive a physical assessment, including measures of fine and gross motor skills. So far, the parental questions and the tests of fine and gross motor skills have not proven reliable. With the former, the main concern is that parents often have little basis for determining whether behavior is age appropriate. With the latter, the main concern is that the scores may be measuring a child's ability to comprehend the instructions as much as his motor skills. As a result, we will not discuss the parents' assessments or motor skills tests.

Cognitive tests of kindergarten readiness tend to concentrate on reading and to a lesser extent on mathematics because reading and math abilities are believed to be more modifiable by preschool programs, parental behavior, and formal schooling than some other aspects of readiness. In the ECLS-K the adaptive tests in reading and mathematics begin with a first-stage test of fifteen to eighteen test items covering the full range of difficulty. A computer calculates a score and then advises the test administrator which second-stage form is appropriate for that child. The direct cognitive assessment takes from fifty to seventy minutes.[12]

Because most entering kindergartners cannot read, the "reading" test at the kindergarten level emphasizes the child's performance on the sequential learning steps based on the phonics approach to reading development, including tasks having to do with familiarity with print, identifying upper- and lower-case letters by name, associating letters with sounds at the beginning of words, associating sounds with letters at the end of words, and recognizing common words by sight. As the ECLS-K moves through later grades, the em-

phasis in the item pool shifts toward reading comprehension skills, such as showing a more complete understanding of what is read, connecting knowledge from the text with the child's personal knowledge, and showing some ability to take a critical stance toward the text.

The ECLS-K mathematics test assesses knowledge in the following areas (in order of difficulty): identifying one-digit numerals, recognizing geometric shapes, and one-to-one counting up to ten objects; reading all one-digit numerals, counting beyond ten, and using nonstandard units of length to compare objects; reading two-digit numbers, recognizing the next number in a sequence, ordinality of objects; solving simple addition and subtraction problems; and solving simple multiplication and division problems. Again, the kindergarten test emphasizes the easier skills, and the tests in later grades shift toward the more advanced skills.

Good rating scales attempt to anchor subjective assessments by including specific descriptions of grade-appropriate performance or behaviors that are then rated on a five-point scale.

The direct cognitive measures of reading and mathematics have reliability in the low .90s—equal to or better than scores typically found in cognitive achievement tests given to older children. Moreover, it was frequently reported that the children did not want to end their assessment, largely because they enjoyed the individual attention from the test administrator. The test administrators received considerable training, including practice sessions, and the materials in the test were colorful and "game-like."

Kindergarten teachers also evaluated their students along both cognitive and behavioral dimensions. Good rating scales attempt to anchor subjective assessments by including specific descriptions of grade-appropriate performance or behaviors that are then rated on a five-point scale, with the highest number indicating that the child is proficient at the specified skill. In testing cognitive skills, the teacher evaluations follow the same general categories of reading, math, and general knowledge. The teacher social skills rating scale (TSRS) rates the kindergarten children on five socioemotional skills. "Approaches to learning" rates a child's attentiveness, task persistence, eagerness to learn, learning independence, flexibility, and organization. "Self-control" measures the child's ability to control behavior by respecting the property rights of others, controlling temper, accepting peer ideas for group activities, and responding appropriately to peer pressure. "Interpersonal skills" rates the child's behavior in forming and maintaining friendships; getting along with people who are different; helping and comforting other children; expressing feelings, ideas, and opinions in positive ways; and being sensitive to the feelings of others. "Externalizing problem behaviors" measures the likelihood that a child argues, fights, gets angry, acts impulsively, and disrupts ongoing activities. "Internalizing problem behaviors" measures anxiety, loneliness, low self-esteem, and sadness.

Although these teacher ratings may seem subjective, they proved almost as reliable as

the direct cognitive scores. The teacher's rating of the child's reading development was a very respectable .87, while the teacher's rating of a child's mathematical development was .92. Similarly the teacher social ratings all had reliability close to .90, except for the measure of self-control, which had an acceptable reliability of .79.

How well are the direct cognitive ratings correlated with the teacher evaluations? Such correlations help evaluate what researchers call "construct validity," the extent to which a test measures what it is intended to measure. A measure has construct validity if it correlates well with other tests that theory suggests are measuring similar things ("convergent validity") and if it correlates relatively poorly with other tests that theory suggests are measuring different things ("discriminant validity").[13] In this case, the difficulty is that the teacher evaluations of reading and math achievement are quite highly correlated, at .83. The correlation between teacher evaluations of reading and cognitive evaluation of reading, at .60, is exactly the same as for math. Similarly, the teacher evaluation of math has only a very slightly higher correlation with the cognitive measure of math, at .54, than it does with the cognitive measure of reading, at .51.

In addition, some of the nonacademic teacher ratings of social skills, notably self-control and interpersonal skills, are more highly correlated with the academic ratings than are the corresponding test scores, which suggests a possible "halo" effect among the teacher ratings. However, the high correlation of the self-control scale and the interpersonal skills scale with the teachers' ratings of academic performance, and to a lesser extent with the tested academic performance, is also consistent with Andrew Pellegrini's theory that social skill development predicts literacy performance.[14]

The Size of the Readiness Gap

Various studies have used the tests and data sources described here to measure the readiness gap for kindergartners. Table 1 lists some selected studies that have measured academic readiness; table 2 presents studies that have measured social or behavioral readiness. The first column of each table lists the authors and the date of the study. The second column identifies the test used. The third column comments on the data used. The final columns list what are called "raw gaps" and "adjusted gaps," measured in "standard deviation units." These terms require further explanation.

Using Standard Deviation as a Common Yardstick

The human sciences in general—and psychology and education in particular—lack common, shared interchangeable metrics for expressing differences on many important constructs, like reading achievement, health risk, or depression. There are more than 200 nonexchangeable metrics for assessing how well students read.[15] Each reading test reports in a scale specific to that test—like the PPVT or the ECLS-K reading scale—but no tables exist for converting the score on one reading scale into the metric of another. How can researchers compare the results of studies done with different instruments?

To visualize the problem, consider figure 1, which shows a common pattern that arises in studies of readiness among black and white children. The darker line shows the distribution of scores for black children, the lighter line that for white children, in a study using the PPVT as the test and the NLSY79 data. The test scores have been coded so that the

Table 1. Selected Estimates of the Academic School Readiness Gap

Study	Test	Sample	White-black		White-Hispanic	
			Raw	Adjusted	Raw	Adjusted
Fryer and Levitt (2004)	ECLS-K Math test	20,000 kindergartners (ECLS-K)	0.64	0.09[a]	0.72	0.20[a]
	ECLS-K Reading test		0.40	0.12[a]	0.43	0.06[a]
	ECLS-K Math teacher assessment		0.28	0.10[b]	0.24	0.10[b]
	ECLS-K Reading teacher assessment		0.27	0.07[b]	0.35	0.18[b]
Brooks-Gunn, Klebanov, Smith, Duncan, and Lee (2003)	PPVT-R Vocabulary	315 five-year-olds (IHDP)	1.63	0.86[c]		
	WPPSI IQ	315 five-year-olds (IHDP)	1.21	0.38[c]		
	PPVT-R Vocabulary	1,354 five- to six-year-olds (NLSY child data)	1.15	0.73[c]		
Phillips, Brooks-Gunn, Duncan, Klebanov, and Crane (1998)	PPVT-R Vocabulary/IQ	Five- and six-year-olds (NLSY)	1.14	0.95[d]		
	PPVT-R Vocabulary/IQ	Five-year-olds (IHDP)	1.71	0.69[d]		
	WPPSI IQ	Five-year-olds (IHDP)	1.28	0.26[d]		

Sources: Roland G. Fryer and Steven D. Levitt, "Understanding the Black-White Test Score Gap in the First Two Years of School," *Review of Economics and Statistics,* vol. 86, no. 2 (May 2004): 447–64; Jeanne Brooks-Gunn, Pamela K. Klebanov, Judith Smith, Greg J. Duncan, and Kyunghee Lee, "The Black-White Test Score Gap in Young Children: Contributions of Test and Family," *Applied Developmental Science* 7, no. 4 (2003): 239–52; Meredith Phillips, Jeanne Brooks-Gunn, Greg J. Duncan, Pamela Klebanov, and Jonathan Crane, "Family Background, Parenting Practices, and the Black-White Test Score Gap," in *The Black-White Test Score Gap*, edited by Christopher Jencks and Meredith Phillips (Brookings, 1998), pp. 103–45.

Notes: To standardize the score differentials, we used 16 as the standard deviation on the Stanford-Binet and 15 as the standard deviation on the PPVT-R and the WPPSI, unless the author gave the actual standard deviation for the entire sample. ECLS-K is the Early Childhood Longitudinal Study-Kindergarten Cohort; IHDP is the Infant Health and Development Program; EHS is the Early Head Start Research and Evaluation Program; NLSY is the National Longitudinal Survey of Youth Child Supplement.

a. Controls for composite measure of socioeconomic status, a quadratic in the number of children's books, sex, age attending kindergarten, birth weight, mother's age at birth, and WIC participation.

b. Same as note a with the addition of teacher fixed effects.

c. Controls for family income, female headship, mother's education, mother's age at birth, and home environment.

d. Controls for family income, female headship, mother's educational attainment, neighborhood socioeconomic status, home learning environment, and home warmth.

average score for white and black children combined is 50. The median score for blacks (that is, the score that half the children are above and half below) is 40; the median score for whites is 52. Most children, however, are not exactly at the middle, but are rather above or below it, and so graphs of scores on readiness tests typically take on a hill, or bell, shape, with relatively few children at the extremes and more clustered near the middle of the distribution. The gap between the median white and black scores is 12 points—but who knows what that means compared with any other vocabulary or readiness scale?

Statisticians have a tool called the standard deviation for measuring the spread of a bell-shaped distribution.[16] A standard deviation tells how far a distribution is spread out around the average score—the numerical scale used to measure the scores doesn't matter. To put it another way, imagine that in fig-

Table 2. Selected Estimates of the Behavioral School Readiness Gap

Authors	Test	Sample	White-black		White-Hispanic	
			Raw	Adjusted	Raw	Adjusted
Magnuson (2004)	Approaches to learning	20,000 kindergartners, teacher reports	.36		.21	
	Self-control	(ECLS-K)	.38		.13	
	Externalizing behavior		–.31		.01	
	Internalizing behavior		–.06		–.05	
Chase-Lansdale, Gordon, Brooks-Gunn, and Klebanov (1997)	Internalizing behavior (Achenbach CBCL)	642 five-year-olds, maternal reports (IHDP)		–.30[a]		
	Externalizing behavior (Achenbach CBCL)			–.20[a]		
	Internalizing behavior (BPI)	699 five- to six-year-olds, maternal reports (NLSY-CS)		–.01[a]		
	Externalizing behavior (BPI)			–.22[a]		

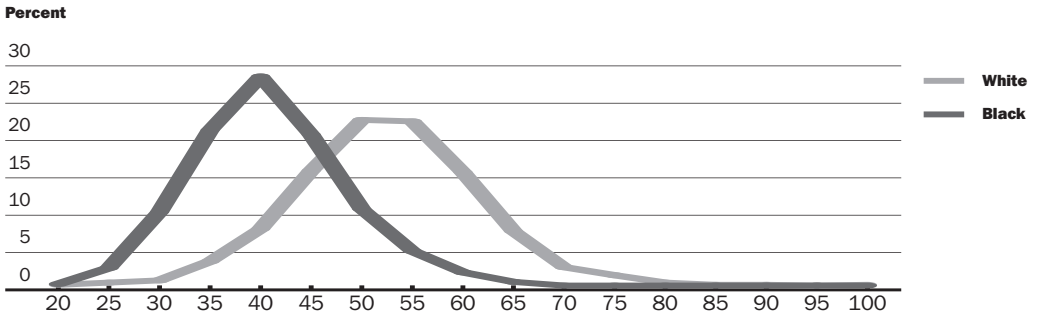
Sources: Katherine Magnuson, analyses prepared for this article from the Early Childhood Longitudinal Study-Kindergarten Cohort, School of Social Work, University of Wisconsin (2004); P. Lindsay Chase-Lansdale, Rachel A. Gordon, Jeanne Brooks-Gunn, and Pamela K. Klebanov, "Neighborhood and Family Influences on the Intellectual and Behavioral Competence of Preschool and Early School-Age Children," in *Neighborhood Poverty*, vol. 1, *Context and Consequences for Children*, edited by Jeanne Brooks-Gunn, Greg L. Duncan, and J. Lawrence Aber (New York: Russell Sage, 1997), pp. 79–118.
Notes: ECLS-K is the Early Childhood Longitudinal Study-Kindergarten Cohort; IHDP is the Infant Health and Development Program; NLSY-CS is the National Longitudinal Survey of Youth—Child Supplement.
a. Controls for gender, family income, female headship, mother's age at birth, mother's employment, age, and school status.

ure 1, all the scores on the horizontal axis were multiplied by a factor of 10, or 20, or any number you choose. The scores themselves would change, and the measure of the gap between the peaks of the white and black distributions would change, but the number of standard deviations between the two peaks would be exactly the same. Thus, instead of expressing the readiness gap in terms of scores on a particular test, which cannot readily be compared with scores on other tests, researchers can express the readiness gap in terms of standard deviations. In figure 1, the standard deviation is 10 points, so a gap of 12 points means 1.2 standard deviations.

Using standard deviations to compare distributions is based on the underlying assumption that the hill shapes of the distributions are the same. This assumption is not literally true. But it remains useful for researchers, because it creates a "scale free" measure of effects that allows comparisons across studies with different numerical scales.[17]

Now look back at table 1 and the column showing the white-black "raw" gap, the gap between the averages for white and for black children before scores are adjusted to take into account such factors as the age or education of a child's mother, family income, or whether the child was born at low birth weight. By this measure, the studies listed in table 1 typically find a white-black gap of more than 1 standard deviation, with many of the estimates roughly similar to the gap illustrated in figure 1. But the estimates of the white-black raw gap at entrance to kindergarten using the ECLS-K data are substantially lower, often hovering at about 0.5 standard deviation. Finally, the highest estimates of the raw gap in the table are generated

Figure 1. Vocabulary Scores for Three- and Four-Year-Olds, by Race



Source: Christopher Jencks and Meredith Phillips, eds., *The Black-White Test Score Gap* (Brookings, 1998).
Notes: The data are from National Longitudinal Survey of Youth Child study, 1986–94. For blacks, N = 1,134; for whites, N = 2,071. The figure is based on black and white three- and four-year-olds who took the Peabody Picture Vocabulary Test-Revised. The test is the standardized residual, coded to a mean of 50 and a standard deviation of 10, from a weighted regression of children's raw scores on their age in months, age in months squared, and year-of-testing dummies.

using the PPVT, some of which are substantially greater than 1 standard deviation. The studies listed in table 2 find a much smaller gap in behavioral readiness, with the raw gap often in the range of 0.0 to 0.3 standard deviation. Some measures even find a negative gap in behavioral readiness, meaning that black or Hispanic children were more behaviorally ready for kindergarten on this dimension than white children.

How Much Does 1 Standard Deviation Matter?

Should a gap of, say, 1 standard deviation in reading ability be considered a big difference? To what extent should policymakers take note of a white-black achievement gap that averages 1 standard deviation?

Statisticians often work with what they call a "normal" distribution, the bell-shaped distribution produced by many random observations, such as flipping 100 coins and seeing how many times heads comes up or rolling two dice and seeing how often each total comes up. A rule of thumb for normal distributions is that 68 percent of all scores will be within 1 standard deviation above or below the mean score, while 95 percent of all scores will be within 2 standard deviations of the mean. In that spirit, consider the situation in which the gap between the peak of the hill-shaped distributions of scores for white and black children is 1 standard deviation. Under the assumptions that the two distributions have the same standard deviation and that both distributions are "normal," the following six statements about the degree of overlap between the two distributions will all hold true.[18]

First, randomly selecting one black child and one white child and comparing their scores will show the white child exceeding the black child 76 percent of the time and the black child exceeding the white child 24 percent of the time. Second, 84 percent of white children will perform better than the average black child, while 16 percent of black children will perform better than the average white child. Third, if a class that is evenly divided by race is divided into two equal-sized groups based on ability, then black students will compose roughly 70 percent, and whites 30 percent, of the students in the lower performing group. Fourth, if a school district

chooses only the top-scoring 5 percent of students for "gifted" courses, such classes will have thirteen times more whites than blacks. Fifth, assume that a school district's student body mimics the national racial distribution (17 percent black, 83 percent white and other). The district chooses the lowest-scoring 5 percent of all students for a special needs program. Although 17 percent of the district's children are black, 72 percent of the special needs students will be black. Finally, assume that a reading textbook is written so that the average white student will read it at a 75 percent comprehension rate. The implied comprehension rate for the average black student will be 53 percent, virtually guaranteeing that such a reader will not engage with the text.[19]

These statements strongly suggest that a gap of 1 standard deviation is quite important in terms of student performance and should be of serious concern to policymakers. Indeed, even a gap of 0.5 standard deviation will result in striking differences between races, especially in matters like how many students are assigned to gifted or to remedial classes.

Raw Gap versus Adjusted Gap

Two columns in table 1 are labeled "raw gap," one referring to the gap between whites and blacks and the other to that between whites and Hispanics. As noted, the raw gap is calculated by looking at the distributions for white students and for either black or Hispanic students and calculating the difference between the mean scores, measured in terms of standard deviations, without making any further adjustments.

Two other columns are labeled "adjusted gap." The adjusted gap is the raw gap adjusted statistically to take into account different factors that might affect scores. For example, the 2003 study by Jeanne Brooks-Gunn and others listed in table 1 accounts for family income, whether a woman is the head of the family, the mother's level of education, the mother's age at the child's birth, and aspects of the home environment. The adjusted gap calculates how much one would expect a white and black (or Hispanic) student to differ even if both had the same family income, the same type of head of household, mothers of the same education and age, and the same home environment. Different studies use different data on the child and family, so one study's adjusted score will account for different factors than another's. The specific factors taken into account in the adjusted scores are listed in the notes to tables 1 and 2.

The adjusted gap often substantially reduces the raw gap, although how much it does so varies across test instruments and studies. This pattern suggests that influences outside school, such as family background, health, and neighborhood, can have important effects on a child's academic readiness for school. In some of the calculations using the ECLS-K data in table 1, these other factors can almost completely account for the raw gap in white-black academic scores. In most, however, some gap in academic scores remains even after adjustment. In table 2, the adjusted scores are often near zero or even negative, suggesting that outside factors can more than explain any behavioral readiness gap.

Can the Differing Estimates of Readiness Be Reconciled?

No one would reasonably expect the gaps in school readiness between white, black, and Hispanic students to be the same in every study, regardless of the particular test and the data used. What factors might help explain and interpret some of the differences across tests? In particular, why does the most recent

and seemingly up-to-date study, the ECLS-K, produce a substantially lower measure of the readiness gap than do other tests and data?

Sample Characteristics

When two studies differ, a first obvious question is whether they are based on different data. But the data from both the NLSY and the ECLS-K are chosen to be nationally representative, so they should show no systematic difference. And the IHDP data set, although it was not chosen to be nationally representative, is a large enough group and has been studied for long enough that it is unlikely to have a buried flaw that would call results into question. Many of the studies of kindergarten readiness discussed here struggle with such issues as how to make good comparative measurements with children who do not speak English as a first language, or are blind, or perhaps have a condition like cerebral palsy that makes it difficult to finish the test, and to address these issues they make various adjustments. But although differences in the samples certainly explain some of the variation around the edges, they seem unlikely to account for substantial variation.

Racial or Ethnic Bias in the Tests?

A common concern is that the readiness gap measured between white and minority children may be caused by systematic bias in the test; for example, perhaps certain vocabulary words are more commonly used in white families than in black or Hispanic ones. There are many ways to check for racial or ethnic bias.

One straightforward approach is to look at groups of white and minority children who have the same overall scores on the test. These children should also have essentially the same breakdown of right and wrong answers on each question on the test. Otherwise, "differential item functioning" exists, and an item on the test may be sorting by race or ethnicity rather than ability.

A related concept is construct bias; that is, whether a test measures what it purports to measure. A test is construct biased if items tend to be more familiar to one group than another, so that the characteristics of the test question help to explain why whites, blacks, or Hispanics find the questions hard or easy to answer. More than thirty years of intense examination of the possibility of construct bias, with particular focus on white-black differences, has failed to demonstrate that they are due to construct bias in achievement tests.[20]

Prediction bias might arise if a school district used a "school readiness battery" administered in kindergarten to predict third grade reading proficiency and found that the ability of the test to predict later proficiency differed for blacks, Hispanics, and whites. In general, though, achievement test items like reading, vocabulary, mathematics, social studies, and science function the same for blacks and whites. That is, test scores on achievement tests predict similarly for blacks and whites—and indeed, at the high school level, they have a slight tendency to overpredict black outcomes in college grades and workplace performance (rather than underpredict, as would be expected if there were prediction bias).[21] Thus, claims of prediction bias for achievement tests are, for the most part, not sustainable.

Another possibility is that even if the test instruments themselves are not racially or ethnically biased, the broader social context in which these tests and their uses are embed-

ded may lead to racial or ethnic gaps in outcomes. Claude Steele and Joshua Aronson have conducted studies that show that calling a test "a diagnostic measure of ability" produces in black students a "stereotype threat," resulting in poorer test performance. The black-white gap is markedly reduced when the test does not bear the label "intellectual ability." Steele and Aronson caution against generalizing these findings beyond high achievers at a prestigious university and call for further study of the central hypothesis and its many implications.[22] In particular, it is not clear whether this issue would affect kindergartners.

Are the Tests Different Ways of Measuring a Common Underlying Readiness?

There is little evidence that distinctions such as verbal versus nonverbal, group administered versus individually administered, spatial versus numerical, or paper-and-pencil versus performance test explain the pattern of gap size estimates. Differences in the readiness gap across the tests can to some extent, however, be explained by the Spearman Hypothesis. This hypothesis states that all tests are imperfect measures of a general ability construct, commonly known as *g*. The more highly a given test correlates with *g*, the larger will be the black-white readiness gap.[23]

Highly specific school-related tasks, like those involving handwriting or auditory memory span, have lower correlations with general ability (*g*). But tests that involve reasoning with figures or vocabulary tests like the PPVT-R correlate highly with *g*. When a test combines multiple task types into a composite, as do all the tests reviewed above (other than the PPVT-R), the composite score correlates more highly with *g* than do the specific subtests—in keeping with the Spearman Hypothesis. In effect, composite scores average out the specific contributions of particular task types, leaving what is common among them—that is, general ability, *g*. Researchers have tested the Spearman Hypothesis repeatedly over the past twenty years by looking at the common factors across the intelligence tests, and the hypothesis has successfully predicted the pattern of black-white differences in thirteen studies using a broad array of cognitive tests.[24]

But the "vocabulary" construct measured by the PPVT-R seems to pose a challenge to the Spearman Hypothesis. Even though one would expect vocabulary to be highly correlated with general ability (*g*), it is only one measure and thus should presumably produce a smaller black-white readiness gap than do composite scores. But as noted, the PPVT-R produces some of the highest estimates of the readiness gap. Further, theories of vocabulary acquisition emphasize that words with high frequency in written and oral discourse are learned first, and words with low frequency are learned later; that is, children learn words primarily because they are exposed to them.[25] And the order of vocabulary acquisition is highly invariant for advantaged and disadvantaged populations. Perhaps the greater exposure to words in some way exaggerates differences in underlying general ability, but the reasons why vocabulary tests often produce a larger readiness gap than composite achievement tests remain to be investigated.

What about the ECLS-K?

The readiness gap as measured by the ECLS Kindergarten sample is consistently smaller than that detected by the other methods, whether using raw or adjusted scores. Why might this be so? The ECLS test was designed more recently, with many useful up-

dates in its methodology and administration, and it has a larger and more recent database. These factors might contribute to a smaller measure of the readiness gap.

Another possibility that fits with the Spearman Hypothesis, however, is that the version of the ECLS-K test given to kindergarten students is less correlated with general ability, *g,* than is the version given later to, say, third graders. Remember that the ECLS-K test evolves and looks different for different age levels. In kindergarten the ECLS-K reading test involves basic phonics and decoding tasks; by third grade, the emphasis of the reading test has shifted toward comprehension, with a heavy word-meaning component. As the ECLS-K assessment moves on from basic skill processes in kindergarten to product outcomes in third grade it finds a larger black-white readiness gap. Indeed, the ECLS-K readiness gap as of third grade is much closer to that found by other test instruments. It is possible that the lower ECLS readiness gap at the kindergarten level may reflect the specific way it tests kindergartners.[26] At the same time, student scores on the ECLS kindergarten test are very highly correlated with their scores on the test in third grade, suggesting that the two tests are not in fact measuring different constructs. Clearly, the reasons why the ECLS-K test generates smaller estimates of the racial and ethnic gaps in school readiness are not well understood and are worthy of serious future study, because of the important implications for education policy.

Student scores on the ECLS kindergarten test are very highly correlated with their scores on the test in third grade, suggesting that the two tests are not measuring different constructs.

Future Directions for Research on the Readiness Gaps

Future research on the school readiness gaps among black, white, and Hispanic children will depend to a large extent on the availability of new data and the uses of new methods. Data from the ECLS Kindergarten 1998–99 cohort have invigorated research in this area. And ECLS is also now tracking a sample of 10,600 children born in 2001 whom it plans to follow through first grade. The new study seems certain to provide further evidence about the size and underlying causes of the racial and ethnic readiness gaps. Researchers should also be on the lookout for situations in which a large group of kindergarten-age children, such as the IHDP group, might usefully be administered an achievement test.

Another approach is to use different methods. A relatively new line of thought emphasizes a kind of cognitive measurement that is highly correlated with general ability, *g*. "Choice reaction time" is the time it takes the subject to react to a light stimulus by moving her index finger from a home base to one or more of eight lights arranged in a semicircle. Total reaction time is decomposed into the milliseconds it takes the examinee to remove her index finger from the home base after the stimulus light is activated and the time it takes *after* removing the index finger to touch the stimulus switch. The two times are experimentally independent. The procedure is simple, can be used for all ages, requires no memory component, and is highly reliable.

And the time it takes a subject to remove a finger from the home base is remarkably highly correlated with cognitive test composites.[27] Some tantalizing links also exist between reaction time and vocabulary development.

Most data sets described in this paper are longitudinal—that is, they track groups of children over time. Such an approach is obviously useful for investigating the determinants and effects of school readiness. But it is not the only possible approach. For example, if assessors are interested in a snapshot of the status of the children at a specific time, a single cross-sectional study can be less costly and less complex than a longitudinal study.

Yet another approach is to conduct an experiment by assigning children to different government intervention programs and having each intervention test the children's school readiness. For example, the federal government has supported the Early Head Start Research and Evaluation Project (EHS), which has studied seventeen Head Start programs around the United States since the late 1990s using a methodology in which 3,000 children were randomly assigned either to Early Head Start or to a control group. The first phase of the study focused on children from birth to age three, but a second phase from 2001 to 2004 is tracking children from the time they leave Early Head Start until they enter kindergarten. The project is evaluating prekindergarten children using many of the tools already discussed: the PPVT, the Woodcock Johnson Psycho-Educational Test Battery, the Achenbach Child Behavior Checklist, analysis of videotaped problem-solving and play sessions, and others. These data will surely generate a wave of studies of kindergarten readiness, often with policy implications, in the next few years. Of course, experimental evidence of this sort need not be collected nationwide; such experiments can also be carried out at the state or metropolitan levels.

Future research on the readiness gap at kindergarten will prove useful, but it seems highly unlikely to overturn the conclusion that the raw readiness gaps, between white and black children in particular but also between white and Hispanic children, are real and large. The remainder of this issue is devoted to exploring possible explanations for this very serious problem, along with their policy implications.

Endnotes

1. Lori Shepherd, Sharon Lynn Kagan, and Emily Wurtz, eds., *Principles and Recommendations for Early Childhood Assessments* (Washington: National Education Goals Panel, February 1998).

2. Samuel J. Meisels, "Can Head Start Pass the Test?" *Education Week* 22, no. 27 (March 19, 2003): 44; Anthony D. Pellegrini and Carl D. Glickman, "Measuring Kindergarteners' Social Competence," *Young Children* (May 1990): 40–44.

3. Sally Atkins-Burnett, Brian Rowan, and Richard Correnti, "Administering Standardized Achievement Tests to Young Children: How Mode of Administration Affects the Reliability of Standardized Measures of Student Achievement in Kindergarten and First Grade," paper presented at the annual meeting of the American Educational Research Association, April 2001 (available at www.sii.soe.umich.edu/papers.html).

4. Although the studies reviewed in this issue use the PPVT-R, the test has recently been revised, and studies now in the field use the PPVT-III. For discussion, see Jeanne Brooks-Gunn, Pamela K. Klebanov, Judith Smith, Greg J. Duncan, and Kyunghee Lee, "The Black-White Test Score Gap in Young Children: Contributions of Test and Family," *Applied Developmental Science* 7, no. 4 (2003): 239–52.

5. A. Jackson Stenner, Malbert Smith, and Donald S. Burdick, "Toward a Theory of Construct Definition," *Journal of Educational Measurement* 20, no. 4 (1983): 304–15.

6. George A. Miller and Patricia M. Gildea, "How Children Learn Words," *Scientific American* 257, no. 3 (1987): 94–99.

7. Elizabeth P. Hagen, Elizabeth A. Delaney, and Thomas F. Hopkins, *Stanford-Binet Intelligence Scale—Examiner's Handbook: An Expanded Guide for Fourth Edition Users* (Chicago: Riverside Publishing Company, 1987).

8. For the genesis of the Behavioral Problems Index, see James L. Peterson and Nicholas Zill, "Marital Disruption, Parent-Child Relationships, and Behavioral Problems in Children," *Journal of Marriage and the Family* 48, no. 2 (May 1986). For a discussion of how the BPI is used in the NLS, see Center for Human Resource Research, *NLSY79 Child and Young Adult Data Users Guide* (Ohio State University, December 2002), especially pp. 91–94.

9. The article by Jeanne Brooks-Gunn and Lisa Markman in this issue describes in more detail how this approach was used in one study of 2,000 three-year-olds, with data from the Early Head Start Research and Evaluation Project (EHS). These data are also discussed further at the end of the present article.

10. National Center for Education Statistics, *U.S. Dept. of Education, ECLS-K Base Year Data Files and Electronic Codebook* (2001).

11. Susan M. Benner, *Assessing Young Children with Special Needs: An Ecological Perspective* (New York: Longman, 1992); Everett Waters and Alan L. Sroufe, "Social Competence as a Developmental Construct," *Developmental Review* 3 (1983): 79–97; Anthony Pellegrini, Lee Galda, and Donald L. Rubin, "Context in Text: The Development of Oral and Written Language in Two Genres," *Child Development* 55 (1984): 1549–55.

12. Frederick M. Lord and Melvin R. Novick, *Statistical Theories of Mental Test Scores, with Contributions by Alan Birnbaum* (Reading, Mass.: Addison-Wesley, 1968); Frederick M. Lord, *Applications of Item Re-*

sponse Theory to Practical Testing Problems (Hillsdale, N.J.: Lawrence Erlbaum Associates, 1980). See also Benner, *Assessing Young Children* (see note 11).

13. Donald T. Campbell and Donald W. Fiske, "Convergent and Discriminant Validation by the Multi-Trait Multi-Method Matrix," *Psychological Bulletin* 56 (1959): 81–105.

14. Pellegrini, Galda, and Rubin, "Context in Text" (see note 11).

15. A. Jackson Stenner and Benjamin D. Wright, "Readability, Reading Ability, and Comprehension" (paper presented at the Association of Test Publishers Hall of Fame induction for Benjamin D. Wright, San Diego, 2002), in *Making Measures,* edited by Benjamin D. Wright and Mark H. Stone (Chicago: Phaneron Press, 2004).

16. The mathematical formula for calculating standard deviation works like this: (1) calculate the average of the scores; (2) calculate the difference between each individual score and the average; (3) square these differences from the average, and then add them up; (4) take the square root of the total. This calculation will give the number of points that are equal to 1 standard deviation for this group of scores.

17. Space does not permit a full treatment of the soundness of all assumptions underlying the standard deviation as a common unit of effect. However, we did compare the standard deviations for five well-known reading tests that were linked to a common scale and found they ranged from a low of .94 to a high of 1.13. This modest variability across grades and tests provides a context for evaluating the variability in estimates of the black-white achievement gap across various studies and instruments reported in this volume.

18. For purposes of this discussion we made the usual simplifying assumptions of bivariate normality, homogeneity of variance, and equal sample sizes in the two groups. Furthermore, we assume that the 1 standard deviation difference is in construct measures, not test score performances, which are uncorrected for measurement error.

19. Using data from the NCES-NAEP website, we estimate that 1.0 standard deviation on NAEP is equivalent to 220L (220 Lexiles). A back check on this number is to average four norm-referenced achievement test (NRT) standard deviations. The RMSA standard deviation for the four NRTs is 229L. Comprehension rate is modeled as the difference between reader ability and text readability. A difference of 225L between a targeted reader (75 percent comprehension rate) at fourth grade and the average black fourth grader implies a 53 percent comprehension rate for a "book bag" of fourth grade textbooks. See Lexile.com, the Lexile Calculator.

20. For background on construct validity, see Stenner, Smith, and Burdick, "Toward a Theory of Construct Definition" (see note 5). For discussion of the evidence, see Richard E. Nisbett, "Race, Genetics, and IQ," in *The Black-White Test Score Gap,* edited by Christopher Jencks and Meredith Phillips (Brookings, 1998). The psychometric literature has largely given up on the term "bias" in favor of the less emotionally charged terms "differential item functioning" and "differential instrument functioning."

21. See Thomas J. Kane, "Racial and Ethnic Preferences in College Admissions," Frederick E. Vars and William G. Bowen, "Scholastic Aptitude Test Scores, Race, and Academic Performance in Selective Colleges and Universities," and William R. Johnson and Derek Neal, "Basic Skills and the Black-White Earning Gap," in *The Black-White Test Score Gap,* edited by Jencks and Phillips (see note 20).

22. Claude M. Steele, "Race and the Schooling of Black America," *Atlantic Monthly* (April 1992): 68–78; Claude M. Steele, "A Threat in the Air: How Stereotypes Shape the Intellectual Identities and Performance of Women and African Americans," *American Psychologist* (June 1997): 613–29; Claude M. Steele and Joshua Aronson, "Stereotype Threat and the Intellectual Test Performance of African Americans," *Journal of Personality and Social Psychology* 69, no. 5 (1995): 797–811.

23. Arthur R. Jensen, *Bias in Mental Testing (*New York: Free Press, 1980), especially p. 146-147; Arthur R. Jensen, "Spearman's Hypothesis Tested with Chronometric Information-Processing Tasks," *Intelligence* 17 (1993): 47–77.

24. Arthur R. Jensen, "Psychometric *g* and Achievement," in *Policy and Perspectives on Educational Testing*, edited by Bernard R. Gifford (Boston: Kluwer Academic, 1993), pp. 117–227.

25. Betty Hart and Todd R. Risley, *The Social World of Children Learning to Talk* (Baltimore: Brooks, 1999). See Stenner, Smith, and Burdick, "Toward a Theory of Construct Definition" (see note 5), and Miller and Gildea, "How Children Learn Words" (see note 6), for introductions to exposure theory. See Jensen, *Bias in Mental Testing* (see note 23) for an introduction to education theory.

26. Meredith Phillips, James Crouse, and John Ralph, "Does the Black-White Test Score Gap Widen after Children Enter School?" in Jencks and Phillips, *The Black-White Test Score Gap* (see note 20), pp. 229–71.

27. Jensen, *Bias in Mental Testing* (see note 23). See also William Hick, "On the Rate of Information," *Quarterly Journal of Experimental Psychology* 4 (1952): 11–26.