12

Subgroup Differences on Cognitive Tests in Contexts Other Than Personnel Selection

Paul R. Sackett and Winny Shen

Introduction

This chapter focuses on racial and ethnic group differences on cognitive tests of developed abilities. This includes measures of ability in the math and verbal domains, composite measures viewed as measures of *g* or of general intelligence, and measures viewed by their designers as tests of achievement in a specific domain (e.g., math, verbal, science) in contexts other than personnel selection. While industrial and organizational (I/O) psychologists are quite familiar with the common pattern of findings regarding subgroup differences on cognitive tests of developed ability and achievement in the employment context, we believe there is considerable value in putting these findings in a broader context. It is important to understand whether there is something about the employment context that contributes to the magnitude of subgroup differences or whether differences of comparable magnitude are found in other contexts, such as educational admissions or broad national samples tested for research purposes. It is useful to know whether subgroup differences are a phenomenon specific to individuals of working age or whether similar differences are found much earlier in life. It is of considerable interest to understand whether there are trends over time in the magnitude of subgroup differences: Are current differences larger, smaller, or comparable to those found, say, 10 or 20 years ago?

The focus of this chapter is limited. We focus on white–black and white–Hispanic differences on cognitive tests. We focus on these two comparisons as they represent the largest and most studied racial and ethnic subgroups in the United States. We focus on cognitively loaded tests,

given the well-established finding that such measures are among the most valid predictors of job performance, particularly of task performance (as opposed to other components of job performance such as organizational citizenship or counterproductive work behavior). Schmidt and Hunter's (1998) meta-analytic review showed that even if other predictors (e.g., work samples, structured interviews) produce similarly high levels of validity, a composite of those predictors and cognitive ability measures produces higher validity than using the other predictors alone. The combination of favorable validity evidence and the generally consistent finding of substantial white–black and white–Hispanic mean differences make the pairing of these types of tests and these racial and ethnic groups the focus of considerable attention and study. We also focus on the cognitive domain because tests in this domain are widely used for various purposes from early childhood through adulthood. Many other predictors used in the employment setting do not have a clear counterpart throughout the age spectrum (e.g., interviews, assessment centers). While a number of predictors do have some cognitive loading (i.e., positive correlation with cognitive ability), those correlations are relatively small, indicating that those predictors constitute far more than simply measures of cognitive ability. In other words, these cognitively loaded predictors are more than simply a cognitive test in a different format from the traditional paper-and-pencil multiple-choice tests prototypic of ability tests. For example, the mean interview-ability *r* is 0.27 (Berry, Sackett, & Landers, 2007), and the mean ability-situational judgment test (SJT) *r* is 0.37 for SJTs with knowledge instructions (McDaniel, Hartman, Whetzel, & Grubb, 2007).

We also do not focus systematically on gender in this chapter. Because gender differences tend to be small in the cognitive domain, we do not devote space to extensive documentation of the comparability of findings outside the employment domain. The general finding is differences around 0.10 to 0.25 standard deviations (*SD*) favoring women on measures of verbal ability and differences of similar magnitude favoring men in quantitative ability; these tend to cancel out in composite measures combining the verbal and math domains. However, there is variability across studies and among subtests and item types within the verbal and quantitative domains. For summaries and meta-analyses, see the work of Hedges and Nowell (1995), Hyde and Linn (1988), Hyde, Fennema, and Lamon (1990), and Willingham and Cole (1997).

Issues in the Cognitive Measures Used

Our focus was on tests in the cognitive domain. This included measures of ability in the math and verbal domains, composite measures viewed

as measures of *g* or of general intelligence, and measures viewed by their designers as tests of achievement in a specific domain (e.g., math, verbal, science). We viewed all of these as measures of developed ability; thus, education is a contributor for the developed ability. We acknowledge that we lacked the data needed for a detailed comparison of similarities and differences among measures as in most cases there were no data presenting correlations between the various measures examined. Our strategy was to group measures loosely by domains (e.g., math, verbal, composite) and then present results. We would not be surprised to see modest differences in the standardized mean difference, *d* values, across measures aimed at comparable populations due to differences in specific features of test content and format. We were more interested in "big picture" differences: If the white–black applicant *d* in employment settings averages about a standard deviation, are roughly similar differences found outside the employment context? Thus, we focused on commonly used cognitive ability measures, typically paper-and-pencil tests with multiple-choice format. Whether alternate testing modalities or item types produce differing findings is not a question we are able to address with our current focus on large national databases.

Issues in Estimating *d*

We used the standardized mean difference *d* as the index of group differences. This is the majority mean minus the minority mean divided by the pooled within-group standard deviation. This index expresses the group difference in standard deviation units, with zero indicating no difference, a positive value indicating a higher mean for the majority group, and a negative value indicating a higher mean for the minority group. However, in certain instances the pooled within-group standard deviation was not available, and the overall standard deviation across groups was used.

In the employment setting, adverse impact is a local issue; of interest is the impact of a given predictor in a given applicant sample. The applicant pool of a given job with a given employer may differ from the broader applicant population for a wide range of reasons (e.g., firm reputation, firm visibility, the nature of the firm's recruiting activities). Thus, while data on subgroup differences in various population samples can aid in estimating the likely adverse impact in a given situation, it must be realized that the local situation can differ. Nonetheless, there is interest in understanding the factors that influence the magnitude of *d* in various settings.

Useful insight comes from the largest meta-analysis to date of white–black and white–Hispanic differences on cognitive tests in employment

settings, conducted by Roth, Bevier, Bobko, Switzer, and Tyler (2001). First, they reported smaller *d*s for incumbent samples than for applicant samples, which would be expected in a setting in which a given cutoff excludes a higher proportion of a lower-scoring group. Second, they reported smaller *d*s for applicants for a single job than for applicants pooled across jobs. The broader the set of jobs across which applicants were pooled, the closer the pooled sample came to an estimate of the workforce population value. Applicant pools for a single job tended to show restricted range on cognitive measures relative to broad workforce samples (Sackett & Ostgaard, 1994). These issues highlight the importance of the characteristics of the sample on which *d* is estimated.

As a result, we attend to characteristics of the sample as we review various studies in nonemployment contexts. We review research in five categories: (a) studies of nationally representative samples of young adults; (b) studies of nationally representative samples of enrolled high school students, with particular attention paid to seniors as they are on the verge of workplace entry; (c) studies of the population of students taking the two major college entrance exams, namely, the Scholastic Aptitude Test (SAT) and ACT (formerly the American College Testing Program); (d) studies of the norming samples for widely used intelligence tests (e.g., Wechsler Adult Intelligence Scale [WAIS], Stanford-Binet); and (e) studies of nationally representative samples of children (preschool through grade school students).

Note that we focus our investigation on studies intended as nationally representative of the population of interest (e.g., high school seniors). There are additional studies that focused on a more limited setting, such as studies of students in a single school district, but they are outside the purview of this summary. We also note that all studies included here were large-sample studies. While *N* did vary substantially (e.g., a few thousand for the typical study up to over 1 million), the large sample sizes are such that sampling error plays a minor role in effect size estimation. Thus, we did not use sample size weighting when averaging effect size estimates.

Nationally Representative Samples of Young Adults

The U.S. military has long used the Armed Services Vocational Aptitude Battery (ASVAB) to screen recruits and to qualify them for various military occupational specialties. A composite of verbal and quantitative subtests makes up the Armed Forces Qualification Test (AFQT), which is used for initial entry decisions. As there are restrictions on military entry (e.g., a score above the 10th percentile in the national population is required), there is a need for accurate population norm data. This has resulted in

two large norming efforts in 1980 and in 1997. In both cases, attempts were made to draw nationally representative samples of youths aged 18–22, with careful attention to oversampling by race to ensure stable estimates of minority test performance. Individuals were recruited to take the test for research purposes, and it is important to note that this was a nationally representative sample of young adults and not a sample of military recruits. The white–black comparison was based on about 7,800 individuals in 1980 and 4,000 in 1997. While full details of the 1997 study results are not yet public, some initial findings focusing solely on the white–black comparison have been presented by Dickens and Flynn (2006). They converted AFQT scores to an IQ metric, from which we computed *d* values. The result was a white–black *d* of 1.23 in 1980 and 0.99 in 1997.

Note that these are the only young adult studies representative of the population. Other studies involved college-bound populations, which are range restricted as students self-select regarding whether to take the SAT and ACT, or high school senior populations, which are restricted because the high school dropouts are not included. Sackett and Mavor (2003) documented that the white high school graduation rate has remained relatively constant, rising from 86% to 88% between the late 1970s and 2000. The black rate has risen from 65% in 1972 to over 85% by 1995. The Hispanic rate has risen from 58% in 1976 to 63% in 2000. Thus, high school dropout rates are not inconsequential and vary by race/ethnicity. As a result, school-based assessments may differ from estimates based on representative sampling for youth population. A school-based sample will miss a substantial proportion of the Hispanic population. However, while these studies are not representative, they do involve samples for which level of educational attainment is constant, thus permitting a determination of whether group differences are comparable in samples with similar or dissimilar levels of educational attainment.

Representative Samples of High School Students, With Particular Attention to Seniors

In this section, we present the results of five nationally representative studies of high school students. Four of the studies included samples of 12th graders; we view these as of particular interest as these samples represent youths at the point of transition to the world of work. Table 12.1 presents white–black and white–Hispanic effect sizes for each of these studies, separately for the math and verbal domains. The table also includes the year of the study. We present an overview of each of the studies next and then discuss our findings.

TABLE 12.1

White–Black and White–Hispanic Score Gap in High School Age Students

	Pre-1970	1975	1978	1980	1982	1984	1986	1987	1988	1990	1992	1994	1996	1998	2004
Black–white differences															
Math															
EEO Math, Grade 9	0.98														
LSAY Math, Grade 10								0.75							
NELS Math, Grade 12										0.77					
HS&B Math, Grade 12				0.87											
NAEP Math, Grade 12			1.16		1.05		1.01			0.72	0.92	0.95	0.97	1.12	1.05
EEO Math, Grade 12	1.12														
NELS Math, Grade 12											0.80				
HS&B Math, Grade 12				0.86											
Reading															
EEO Reading, Grade 9	0.96														
NELS Reading, Grade 10										0.66					
HS&B Reading, Grade 10				0.77											

NAEP Reading, Grade 12		1.30	1.30	0.84	0.56	0.74	0.92	0.70	0.72	0.77	0.70
EEO Reading, Grade 12	1.04										
NELS Reading, Grade 12							0.69				
HS&B Reading, Grade 12			0.83								
Vocabulary											
EEO Vocab, Grade 9	1.18										
HS&B Vocab, Grade 10			0.94								
EEO Vocab, Grade 12	1.24										
HS&B Vocab, Grade 12			0.82								
Hispanic–white differences											
Math											
NAEP Math, Grade 12		0.92	0.89	0.84		0.88	0.70	0.75	0.76	0.78	0.88
Reading											
NAEP Math, Grade 12		1.02	0.83	0.71	0.66	0.55	0.65	0.77	0.73	0.59	0.70

We note that Hedges and Nowell (1998) also presented an analysis of these same sets of data. Our analyses differed in several respects. First, Hedges and Nowell focused solely on white–black comparisons, while we also included white–Hispanic comparisons. Second, Hedges and Nowell included data through 1994. We were able to add more recent data, through 2004.

Equality of Educational Opportunity Math and Reading, 1965

Equality of Educational Opportunity (EEO) was a study undertaken in part due to societal concerns at the time, including the passage of the Civil Rights Act of 1964. EEO utilized a national sample of students in several grades between the 1st and 12th grades for assessments of math and reading ability. One of the primary concerns of EEO was to assess the educational opportunities for children of different backgrounds and circumstances. The results of EEO are reported in a document often known as the Coleman report (Coleman, 1966).

High School and Beyond Math, Reading, and Vocabulary, 1980 and 1982

High School and Beyond (HS&B) assessed a national sample of 10th and 12th graders on math, reading, and vocabulary, which is what is reported here (Phillips, Crouse, & Ralph, 1998). However, both cohorts were resurveyed two or three times and followed to assess the relationship between high school characteristics and educational and vocational outcomes (National Center for Education Statistics, n.d.-a).

Longitudinal Study of American Youth Math, 1987

The Longitudinal Study of American Youth (LSAY) followed a nationally representative sample of 12,686 youths (aged 14–22) beginning in 1979 to observe their employment-related outcomes (Bureau of Labor Statistics, 2001).

National Assessment of Educational Progress Math and Reading Long Term Trend, Ages 9, 13, and 17 (1975–2004)

The National Assessment of Educational Progress (NAEP) included periodic assessments of representative samples of school-enrolled youths at the ages of 9, 13, and 17 (Grades 4, 8, and 12) since the early 1970s in the areas of math, reading, and science. Our primary focus here is on the assessment at age 17; we return to the age 9 and age 13 assessments. As a school-based assessment, the age 17 assessment excluded youths who had

dropped out of high school; this constitutes a sizable proportion of the population for some subgroups. The NAEP program includes both measures that change with each administration, thus reflecting changes in school curricula, and a fixed set of measures, referred to as the *long-term trend assessment*. The NAEP long-term trend utilizes the same procedures and types of questions every time it is administered for comparability across years; therefore, changes in curriculum and instruction are not reflected in this assessment (National Center for Educational Statistics, n.d.-b). Our focus here is solely on the long-term trend data. In recent years, the NAEP assesses approximately 3,000–4,000 white/Caucasian students, 500–1,000 African American/black students, and 400–800 Hispanic students.

National Education Longitudinal Study Math and Reading, 1988, 1990, and 1992

The National Education Longitudinal Study (NELS), like HS&B, was a longitudinal survey undertaken by the National Center for Educational Statistics (n.d.-c). NELS followed a national sample of eighth graders beginning in 1988 and then surveyed these students biennially.

Results

Table 12.1 presents white–black and white–Hispanic *d* values from these studies. For the white–black comparisons, there are 16 math *d*s (mean = 0.94, *SD* = 0.14), 16 reading *d*s (mean = 0.84, *SD* = 0.21), and 4 vocabulary *d*s (mean = 1.05, *SD* = 0.20). For the white–Hispanic comparisons, there are 9 math *d*s (mean = 0.82, *SD* = 0.08), and 10 reading *d*s (mean = 0.72, *SD* = 0.13).

Of particular interest are the NAEP findings as they include assessments in multiple years from 1975 to 2004. In the math domain, the two earliest assessments (1978 and 1982) showed white–black *d*s comparable to the two most recent assessments (1998 and 2004). In the reading domain, the more recent assessment (d = 0.70) was substantially smaller than the earlier assessments (d = 1.30). A similar pattern is seen for the white–Hispanic comparison: little consistent change in the math domain but a reduction in *d* in the reading domain.

College Applicants

Table 12.2 presents white–black and white–Hispanic *d*s for the two major college admissions tests (SAT and ACT) by year. We briefly overview these two testing programs and then discuss the findings.

TABLE 12.2

White–Black and White–Hispanic Score Gap in College Admissions Tests

	1987	1988	1989	1990	1991	1992	1993	1994	1995	1996	1997	1998	1999	2000	2001	2002	2003	2004	2005	2006	2007
White–black differences																					
ACT English											0.89	0.89	0.89	0.89	0.91	0.86	0.88	0.86	0.90		0.95
ACT Math											0.86	0.88	0.88	0.90	0.90	0.92	0.90	0.90	0.94		0.92
ACT Reading											0.84	0.82	0.83	0.85	0.88	0.87	0.85	0.87	0.88		0.89
ACT Science											0.94	0.98	0.98	0.98	1.00	0.98	0.96	0.91	0.96		0.96
ACT Composite											0.98	0.98	0.98	1.02	1.04	1.02	1.00	0.98	1.00		1.02
SAT Verbal	0.96	0.94	0.95	0.91	0.91	0.91	0.90	0.91	0.91	0.91	0.91	0.91	0.92	0.94	0.96	0.97	0.98	0.98	0.98	0.92	
SAT Math	1.01	0.97	0.94	0.96	0.94	0.96	0.94	0.97	0.96	0.98	1.01	1.00	1.03	1.02	1.03	1.04	1.05	1.02	1.03	1.04	
SAT Writing										0.83	0.83	0.82	0.85	0.86	0.88	0.90	0.88	0.89	0.93		
White–Hispanic differences																					
ACT English											0.60	0.61	0.61	0.62	0.62	0.66	0.66	0.66	0.67		0.70
ACT Math											0.44	0.48	0.47	0.47	0.48	0.55	0.55	0.55	0.55		0.53
ACT Reading											0.52	0.50	0.50	0.51	0.55	0.58	0.57	0.60	0.60		0.61
ACT Science											0.57	0.59	0.58	0.58	0.61	0.65	0.64	0.62	0.64		0.63
ACT Composite											0.59	0.60	0.60	0.63	0.65	0.69	0.66	0.68	0.68		0.68
SAT Verbal						0.67	0.66	0.67	0.67	0.67	0.68	0.69	0.69	0.71	0.74	0.75	0.76	0.72	0.73	0.69	
SAT Math						0.58	0.58	0.60	0.60	0.61	0.64	0.66	0.67	0.67	0.69	0.72	0.72	0.70	0.71	0.71	
SAT Writing										0.84	0.83	0.89	0.88	0.93	0.96	1.03	1.04	1.06	1.06		

ACT English, Math, Reading, Science, Verbal, and Composite (1997–2005, 2007)

The ACT is a standardized college admissions test taken by graduating high school seniors. Annually, approximately 1 million students take the ACT; the data presented represent the white–black and white–Hispanic data across all test takers for that particular year or the annual population of ACT test takers for three subgroups studied here (ACT, 2008). The ACT is made up of several subsections (e.g., English, Math, Reading, Science, and Verbal), which form an overall composite score. The ACT only represents potentially college-bound students.

SAT Verbal, Math, and Writing (White–Black 1987–2006, White–Hispanic 1992–2006)

The SAT is also a standardized college admissions test taken by graduating high school seniors. Annually, approximately 1.5 million students take the SAT, and the data presented represent the SAT test-taking population (Kobrin, Sathy, & Shaw, 2007). The SAT reports separate verbal, math, and writing scores. Like the ACT, the SAT only represents potentially college-bound students.

Results

The Table 12.2 findings show a grand mean white–black *d* of 0.93 (*SD* = 0.06) and white–Hispanic *d* of 0.66 (*SD* = 0.13) across SAT and ACT subtests. The white–black data do not show evidence of a time trend, with the possible exception of the SAT Writing subtest, for which *d* was 0.83 in 1995 and 0.93 in 2005. In contrast, there is a pattern of increasing *d* values over time for all SAT and ACT subtests for the white–Hispanic comparison. The largest change is seen for the SAT Writing subtest, for which *d* was 0.84 in 1995 and 1.06 in 2005. It is important to note that the population of test takers can change from year to year, and thus a change in *d* is difficult to interpret.

Norming Samples for Intelligence Tests

Dickens and Flynn (2006) obtained unpublished information from test publishers about white–black *d*s from norming samples for various IQ tests. We summarize their findings here. Comparable data about white–Hispanic differences are not available in the published literature; hence, this section

focuses solely on white–black differences. We note that Dickens and Flynn focused solely on the total IQ score rather than on individual subtests.

Adult Samples: Wechsler Adult Intelligence Scale (WAIS Revised, WAIS Third Edition), 1978 and 1995

Both samples were standardization samples for a new version of the WAIS test. The standardization samples for the WAIS Revised (WAIS-R) in 1978 included both non-Hispanic white and Hispanic as white (Dickens & Flynn, 2006). The WAIS-R standardization sample consisted of 1,880 individuals aged 16–74 and the WAIS Third Edition (WAIS-III) standardization sample consisted of 1,250 individuals aged 16–89 (Kane, 2000). Both samples were representative by age group of the U.S. population at the time.

Dickens and Flynn (2006) reported white–black *d*s of 1.01 and 0.92 for the 1978 and 1995 norming samples, respectively. They reported a separate analysis for individuals aged 25 and under to determine whether young adult samples differed from the full sample; in this subsample, *d*s of 1.00 and 0.89 were obtained for the two norming samples, respectively.

Full Age Range, Adult and Child: Stanford-Binet (1985, 2001)

The Stanford-Binet is designed and normed for use from ages 2 through 85+. The data here are based on two standardization samples, which match the breakdown of the census at the time. There are some slight differences between the 1985 (SB-4) and the 2001 (SB-5) standardization samples and versions of the tests, such that the test in 2001 was more highly *g* loaded (+12%; Jensen, 1992), and special education and limited English proficiency students, in which blacks were more highly represented, were included (Dickens & Flynn, 2006). The Stanford-Binet uses adaptive testing through a routing subtest that allows a better estimate of appropriate starting points on other subtests and can be used with young children to the elderly (DiStefano & Dombrowski, 2006). The SB-5 standardization sample consisted of 4,800 individuals selected to match the U.S. census. Dickens and Flynn (2006) reported white–black *d*s of 0.90 and 0.77 for the 1985 and 2001 norming samples, respectively.

Child Samples: Wechsler Intelligence Scale for Children (WISC Revised, WISC Third Edition, WISC Fourth Edition), 1972, 1989, and 2002

Each of the samples was a standardization sample for a new version of the Wechsler Intelligence Scale for Children (WISC). The fourth edition (WISC-IV) norming sample was based on 2,200 children from 11 age groups (each covering 1 year, from ages 6 to 16), with an equal number of males and

females in each group, and an ethnic, parental education, and geographic breakdown that matched the 2000 U.S. census. The standardization samples for the revised edition (WISC-R) in 1972 included both non-Hispanic white and Hispanic in the white group. Every version of the WISC is designed to be appropriate for assessing children from approximately ages 6 to 16 (Kaufman, Flanagan, Alfonso, & Mascolo, 2006); however, the specific subscales are not necessarily the same across different versions of the WISC. Dickens and Flynn (2006) reported white–black *d*s decreasing from a high of 1.15 in the 1972 norming sample to 0.78 in the 2002 norming sample.

Wide Range Achievement Test (WRAT), Pre-1970

The Wide Range Achievement Test (WRAT) sample included 7,028 students (6,049 white, 979 black) who were part of the National Health Examination Survey–Cycle II (Svanum & Bringle, 1982). This survey assessed a nationally representative sample of 6- to 11-year-olds from 1963 to 1965 on a number of physical, physiological, and psychological characteristics. These students were assessed on the reading and arithmetic subtests of the WRAT. The white–black *d* was 0.90.

Results

The findings are summarized in Table 12.3. The *d* values from the young adult norming samples for the AFQT are also reported here as Dickens and Flynn (2006) combined the AFQT data with the IQ norming samples for their analysis of time trends in white–black *d*s. As Table 12.3 shows, all samples showed a decrease in white–black *d* over time.

Differences in Preschool and Grade School Samples

Here, we turn to white–black and white–Hispanic differences in preschool and grade school children. We discussed the NAEP age 17 sample in the context of differences among high school-aged youths; here, we include findings from the age 9 and age 13 assessments. We briefly outline the additional nationally representative studies from which we extracted *d* values and then present findings.

National Longitudinal Study of Youth–Child Supplement, 1988 (Average)

The National Longitudinal Study of Youth–Child Supplement (NLSY-CS) is the supplemental assessment of the children of women in the National

TABLE 12.3

White–Black Score Gap in Norming Samples

	Pre-1970	1972	1978	1980	1985	1989	1995	1997	2001	2002
WISC (R, III, IV)	1.13	1.15				1.09				0.78
WAIS (<25, R & III)			1.00				0.89			
WAIS (All ages, R & III)			1.01				0.92			
Stanford-Binet					0.90				0.77	
WRAT	0.90									
AFQT				1.23				0.99		

Longitudinal Study of Youth, a panel study of a nationally representative sample of 14- to 21-year-olds. These data came from the 1986, 1988, and 1992 assessments of two groups of children, those who were 3–4 or 5–6 at the time of the assessments. The children assessed were not themselves a nationally representative sample because they represent children of younger women (Brooks-Gunn, Klebanov, Smith, Duncan, & Lee, 2003). The final 3- to 4-year-old sample consisted of 1,354 children, and the final 5- to 6-year-old sample consisted of 2,220 children. Children were tested on the Peabody Picture Vocabulary Test–Revised (PPVT-R), a measure of spoken word understanding (Dunn & Dunn, 1981). The same data from the NLSY-CS were also presented in the work of Phillips, Brooks-Gunn, Duncan, Klebanov, and Crane (1998).

Early Childhood Longitudinal Study–Kindergarten Cohort Math and Reading, 1998

The Early Childhood Longitudinal Study–Kindergarten Cohort (ECLS-K) began in 1998 to assess a nationally representative sample of 13,000 kindergarteners and will continue to track and assess this cohort until the eighth grade. The data here are for students in the kindergarten cohort in kindergarten (for white–black and white–Hispanic) and the third grade (for white–black differences) on both reading and mathematics reported by Magnuson and Duncan (2006). Items for the reading and mathematics assessments were adapted from national and state standards and other similar assessments (e.g., NAEP; National Center for Education Statistics, n.d.-b).

Panel Study of Income Dynamics–Child Development Supplement Math and Reading, 1997 and 2002

The Panel Study of Income Dynamics (PSID) is a longitudinal study that began in the 1960s of representative samples of men, women, children, and

families in the United States. The Panel Study of Income Dynamics–Child Development Supplement (PSID-CDS) is a supplemental data collection effort that focuses on young children. In the first wave of data collection in 1997, information on 3,563 children between 0 and 12 years old was gathered. In 2002, follow-up data collection was conducted on 2,907 children between 5 and 18 years old. Children were assessed using the Woodcock-Johnson Psych-Educational Battery–Revised (WJ-R), an intellectual ability test designed for use on individuals between 2 and 90 years of age, either in English or Spanish. Children under 6 were assessed on two subtests, Letter-Word and Applied Problem Sets, and children over 6 were given an additional subtest, Passage Comprehension (Mainieri, 2006). The data reported here are based on calculations reported by Magnuson and Duncan (2006).

Iowa Test of Basic Skills–Science, 1993

The Iowa Test of Basic Skills–Science (ITBS-Science) was administered as part of a study examining the score gap between white and minority students on performance-based science assessments (Klein et al., 1997). Further information on the ITBS-Science can be found in Hoover, Hieronymus, Frisbie, and Dunbar (1994). Students in fifth and sixth grade were given the corresponding level of the ITBS-Science according to their grade in 1993. In total, the study assessed over 2,021 fifth- and sixth-grade students. The white–black and white–Hispanic difference scores reported here are in *z*-score units and not standardized mean differences (*d* scores) per se. However, research evidence suggested that variances were relatively equal between different racial groups (Hedges & Nowell, 1998), such that this *z*-score difference should approximate a standardized mean difference.

Prospects Math and Reading, 1991

Prospects (PROS) was also known as the congressionally mandated study of educational growth and opportunity. PROS was a 6-year longitudinal study following several cohorts of national samples of public school students (Puma, Jones, Rock, & Fernandez, 1993). PROS primary goals involved examining the impact of Chapter 1/Title I programs and the differential effects of poverty in schools or students.

Equality of Educational Opportunity Math and Reading, 1965

The EEO utilized a national sample of students in several grades between the 1st and 12th grades. One of the primary concerns of EEO was to assess the educational opportunities for children of different backgrounds and circumstances.

Results

Table 12.4 presents the white–black and white–Hispanic *d* values from these studies. For the white–black comparisons, there are 30 math *d*s (mean = 0.87, *SD* = 0.14), 29 reading *d*s (mean = 0.78, *SD* = 0.15), and 5 vocabulary *d*s (mean = 0.84, *SD* = 0.19). For the white–Hispanic comparisons, there are 19 math *d*s (mean = 0.77, *SD* = 0.12) and 21 reading *d*s (mean = 0.73, *SD* = 0.12).

Of particular interest are the NAEP findings as these involved samples over an extended period of time. For both the white–black and the white–Hispanic comparisons, at both Grade 4 and Grade 8 the pattern is for considerable fluctuation in the math *d*, making it hard to discern any time trend. However, the reading *d*s show a consistent trend of decrease over time.

Summary and Conclusions

Table 12.5 presents mean white–black and white–Hispanic *d*s across the different categories of studies discussed. When *d* values for math and verbal were available but scores on a composite of the two were not reported, we estimated *d* on a composite of the two using the formula provided by Sackett and Ellingson (1997). That formula requires the correlation between the two tests; we used $r = 0.65$ as the correlation between math and verbal tests as this is a typical value for the correlation between the two domains in large unrestricted samples. For example, the correlation between math and verbal composites in the large-scale AFQT norming sample is 0.64; the correlation between the math and verbal subsets of the SAT is about 0.70.

This table gives a clear answer to the question, Is there something specific about the employment context that causes or contributes to subgroup differences? The answer is, No: Differences in the employment context are very similar to differences found in young adult and adult samples in other contexts. Roth et al. (2001) reported mean white–black *d* values in job applicant samples of 1.00 for overall *g* measures, with smaller values for specific ability measures (*d* = 0.83 for verbal and 0.74 for math). The value of 1.00 for *g* measures is very close to the values obtained for composites of verbal and math for the national norming of the AFQT (*d* = 1.11), for college admissions test composites (*d* =1.06 for SAT and 1.00 for ACT), for representative samples of high school students (*d* = 0.98), and for norming samples for IQ tests (*d* = 0.90).

Moving to white–Hispanic comparisons, Roth et al. (2001) reported mean white–Hispanic *d* values of 0.84 for *g* measures in the employment

setting. This is very similar to the values obtained here for composites for college admissions (d = 0.75 for SAT and 0.65 for ACT) and for representative samples of high school students (d = 0.88).

Note that these data include a mix of tests taken in high-stakes settings (e.g., employment and college admissions) and tests taken in low-stakes research settings (e.g., AFQT norming, IQ norming, and studies of high school students). Thus, the pressures of a high-stakes setting do not appear to affect minority student performance differentially as d values are similar in high-stakes and low-stakes settings.

A second question of interest is whether subgroup differences vary by age. Table 12.5 also contains data on white–black differences on g measures for preschool (d = 0.92) and elementary school (d = 0.90) samples. As an alternate approach to this question, we estimated regression models for math and verbal tests with examinee age and study year as predictors. There were 75 math and 105 verbal effect sizes available for this analysis. Note that these analyses excluded d values obtained from samples varying in age (e.g., IQ norming samples). Table 12.6 presents the results. These analyses produced statistically significant coefficients of 0.014 and 0.020 for age for math and verbal, respectively, net of the effects of study year. Thus, d is estimated to increase by 0.014 for math and 0.020 per year from age 4 to age 18.

For the white–Hispanic comparison, Table 12.5 shows white–Hispanic differences on g measures for elementary school samples (d = 0.81), very similar to the d value of 0.84 obtained in the employment setting by Roth et al. (2001). Table 12.6 shows regression analyses using age and study year to predict d values for math (k = 55) and verbal (k = 78) tests. Unlike the white–black analyses, for which age was related to d, for the white–Hispanic data there was no systematic relationship between age and d.

These data showed that subgroup differences measured in early childhood were similar to those obtained in young adulthood. While age was related to d in the white–black comparisons, it is nonetheless the case that d values in early childhood were nearly as large as values obtained in young adulthood. These findings do not identify the causes of group differences, but the fact that differences are observed in early childhood does make clear that it is not something about the employment context or about the transition from adolescence to young adulthood that is a primary determinant of these differences

A third question of interest is whether subgroup differences are changing over time. Our sense is that the preponderance of evidence is that there is some narrowing of the subgroup differences. Dickens and Flynn (2006) concluded that the IQ norming sample data reported here supported a narrowing of the white–black gap; Hedges and Nowell (1995) reached a similar conclusion about the set of nationally representative studies of high school students that they examined and that we also report here. Our Table 12.6 regression analysis also supported this conclusion regarding

TABLE 12.4

White–Black and White–Hispanic Test Score Gap in Children

	Pre-1970	1975	1978	1980	1982	1983	1984	1986	1987	1988	1990	1991	1992	1993	1994	1996	1997	1998	1999	2002	2004
White–black differences																					
Intelligence Tests																					
NLSY-CS-PPVT-R, Preschool										1.25											
NLSY-CS-PPVT-R, Kindergarten										10.09											
Math																					
PSID-CDS Math, Preschool																	0.79				
ECLS-K Math, Kindergarten																		0.65			
PROS Math, Grade 1												0.87									
EEO Math, Grade 3		0.86																			
ECLS-K Math, Grade 3																			0.89		
PSID-CDS Math, Grade 3 & 4																			0.99		
PROS Math, Grade 4												0.67									
NAEP Math, Grade 4			0.93		0.88			0.78			0.86		0.79	0.87	0.78	0.80			0.87		0.72
EEO Math, Grade 6	1.10																				
LSAY Math, Grade 7									0.74												
PROS Math, Grade 8												0.62									
NAEP Math, Grade 8			1.18		1.10			0.83			0.94		1.00		0.97	1.01			1.07		0.88
NELS Math, Grade 8										0.78											
Verbal/Reading																					
PSID-CDS Verbal, Preschool																	0.43				
ECLS-K Reading, Kindergarten																		0.40			

PROS Reading, Grade 1									0.74							
EEO Reading, Grade 3		0.87														
ECLS-K Reading, Grade 3														0.78		
PSID-CDS Verbal, Grades 3 & 4														0.68		
PROS Reading, Grade 4									0.67							
NAEP Reading, Grade 4		0.98	0.90	0.83		0.74		0.82		0.88		0.86	0.78		0.98	.75
EEO Reading, Grade 6	0.90															
PROS Reading, Grade 8									0.56							
NAEP Reading, Grade 8		1.09		0.97	0.78		0.55	0.60		0.77		0.82	0.88		0.78	0.61
NELS Reading, Grade 8							0.70									
Vocabulary																
PROS Vocab, Grade 1									0.70							
EEO Vocab, Grade 3	0.88															
PROS Vocab, Grade 4									0.70							
EEO Vocab, Grade 6	1.15															
PROS Vocab, Grade 8									0.77							
White–Hispanic differences																
Math																
ECLS-K Math, kindergarten														0.75		
NAEP Math, Grade 4		0.62		0.61	0.66	0.69	0.75	0.86		0.70		0.82				0.56
NAEP Math, Grade 8		0.94		0.71	0.66	0.75	0.69	0.83		0.88		0.79				0.75
Verbal/Reading																
ECLS-K Reading, kindergarten														0.41		
NAEP Reading, Grade 4		0.94		0.88	0.79		0.61	0.65		0.69		0.85	0.68	0.78		0.62
NAEP Reading, Grade 8		0.90		0.83	0.68		0.62	0.71		0.74		0.80	0.75	0.63		0.67

TABLE 12.5

Mean *d* Values and Standard Variances for White–Black and White–Hispanic Differences

Type of sample	No. of samples (*k*)	Average *d* value	*SD*
White–Black differences			
Job applicants (from Roth et al., 2001)		1.00	
Nationally representative sample of 18- to 22-year-olds (AFQT norming)	2	1.11	0.17
SAT Math	20	0.93	0.03
SAT Verbal	20	1.00	0.04
SAT Math + Verbal (estimated by formula)		1.06	
SAT Writing	10	0.87	0.04
ACT Math	10	0.90	0.02
ACT Verbal (English & Reading)	20	0.88	0.03
ACT Science	10	0.97	0.03
ACT Composite	10	1.00	0.02
High school math samples	16	0.94	0.14
High school reading samples	16	0.84	0.21
High school math + reading (estimated by formula)		0.98	
High school vocab samples	4	1.05	0.20
Adult norming	4	0.90	0.10
Child norming	5	1.01	0.16
Elementary school math samples	30	0.87	0.14
Elementary school verbal/reading samples	29	0.78	0.15
Elementary school math + verbal (estimated by formula)		0.91	
Elementary school vocabulary samples	5	0.84	0.19
Pre-elementary samples	3	0.92	0.43
White–Hispanic differences			
Industrial samples (from Roth et al., 2001)		0.83	
SAT Math	15	0.70	0.03
SAT Verbal	15	0.66	0.05
SAT Math + Verbal (estimated by formula)		0.75	
SAT Writing	10	0.95	0.09
ACT Math	10	0.51	0.04
ACT Verbal (Reading & Verbal)	20	0.60	0.06
ACT Science	10	0.61	0.03
ACT Composite	10	0.65	0.04
High school math samples	9	0.82	0.08
High school reading samples	10	0.72	0.13