Journal of Psychoeducational Assessment
1999, *17*, 300–313

LONG-TERM STABILITY OF THE WECHSLER INTELLIGENCE SCALE FOR CHILDREN-THIRD EDITION AMONG DEMOGRAPHIC SUBGROUPS: GENDER, RACE/ETHNICITY, AND AGE

Gary L. Canivez
Eastern Illinois University

Marley W. Watkins
Pennsylvania State University

Long-term stability of the Wechsler Intelligence Scale for Children-Third Edition was investigated separately across gender, race/ethnicity, and age subgroups. Participants were 642 students from 33 states evaluated twice for special education eligibility over a mean test-retest interval of 2.83 years. Gender, race/ethnicity, and age produced few or no differential effects on long-term stability coefficients. Most of the demographic subgroup stability coefficients for VIQ, PIQ, FSIQ, VCI, and POI scores demonstrated satisfactory long-term stability. However, stability coefficients for FDI, PSI, and VIQ-PIQ discrepancy scores were not adequate. Mean differences from first testing to second testing were either not statistically significant or not clinically meaningful for all groups, except Hispanic/Latino youths. Analysis of individual change scores indicated that only the FSIQ was sufficiently stable for use with individual students. Results extended those of Canivez and Watkins (1998), supporting long-term stability for the WISC-III among most demographic subgroups studied.

Long-term stability of intelligence tests has been extensively investigated as one facet of their construct validity. Intelligence is a construct presumed to be stable over time; thus, tests measuring this construct must also produce similar scores from one time to another (Moffitt, Caspi, Harkness, & Silva, 1993). Jensen (1980) appropriately referred to correlation coefficients obtained in

Gary L. Canivez is Associate Professor of Psychology at Eastern Illinois University principally involved in the school psychology training program. Marley W. Watkins is Associate Professor and Professor-in-Charge of Graduate Programs in School Psychology at The Pennsylvania State University.

This research was supported in part by a 1995-96 Eastern Illinois University Faculty Development Grant and a 1995-96 Pennsylvania State University College of Education Alumni Society Faculty Research Initiation Grant.

The authors wish to express their gratitude to the 145 school psychologists who generously responded to our request for WISC-III data. They also thank Tim Runge, Lisa Samuels, and Daniel Heupel for assistance in data entry.

Correspondence concerning this article should be addressed to Gary L. Canivez, Department of Psychology, 600 Lincoln Avenue, Charleston, IL 61920-3099. Dr. Canivez may also be contacted via e-mail at cfglc@eiu.edu or the World Wide Web at <http://www.ux1.eiu.edu/~cfglc>.

studies investigating temporal change as stability coefficients. These stability coefficients, however, indicate only the rank order of scores. McDermott (1988) stressed the need to examine mean changes to supplement correlational analyses in order to investigate *level* as well as *pattern* of relationships. Researchers have also presented frequency distributions to reveal individual changes that occur from one testing session to another.

School and clinical psychologists have consistently ranked the Wechsler Scales as the most frequently used measures of cognitive ability (Stinnett, Havey, & Oehler-Stinnett, 1994; Watkins, Campbell, Nieberding, & Hallmark, 1995). Short-term stability research with the Wechsler Intelligence Scale for Children (WISC; Wechsler, 1949) and the Wechsler Intelligence Scale for Children-Revised (WISC-R; Wechsler, 1974) has yielded stability coefficients for Verbal IQ (VIQ), Performance IQ (PIQ), and Full Scale IQ (FSIQ) scores in the .80 to .90 range (Covin, 1977; Irwin, 1966; Quereshi, 1968; Throne, Schulman, & Kaspar, 1962; Tuma & Appelbaum, 1980; Wechsler, 1974). Significant increases in VIQ, PIQ, and FSIQ scores at retest were observed, with the largest increases found in PIQ. WISC and WISC-R subtest stability coefficients were almost always lower than global IQ stability coefficients.

Long-term stability coefficients for the WISC (Coleman, 1963; Conklin & Dockrell, 1967; Friedman, 1970; Gehman & Matyas, 1956; Reger, 1962; Rosen, Stallings, Floor, & Nowakiwska, 1968; Walker & Gross, 1970; Whatley & Plant, 1957) and WISC-R (Anderson, Cronin, & Kazmierski, 1989; Bauman, 1991; Elliott & Boeve, 1987; Elliott, Piersol, Witt, Argulewicz, Gutkin, & Galvin, 1985; Ellzey & Karnes, 1990; Haynes & Howard, 1986; Naglieri & Pfeiffer, 1983; Oakman & Wilson, 1988; Smith, 1978; Stavrou, 1990; Truscott, Narrett, & Smith, 1994; Vance, Blixt, Ellis, & Debell, 1981; Vance, Hankins, & Brown, 1987; Webster, 1988; Whorton, 1985) have been significant and moderate to high, with *r*s generally ranging from the .50s to the .90s. The practice effects seen in short-term stability studies usually disappeared when the retest interval was greater than 1 year. Even when practice effects were found in long-term stability studies, their magnitudes were usually quite small and of no clinical significance.

In contrast to the WISC and WISC-R, there have been few investigations of the stability of the Wechsler Intelligence Scale for Children-Third Edition (WISC-III; Wechsler, 1991). Wechsler reported a short-term stability study with a sample of normal children across a test-retest interval ranging from 12 to 63 days ($Mdn = 23$). Stability coefficients ranged from .71 (FDI for ages 6-7) to .95 (FSIQ for ages 14-15). As expected, test-retest reliability coefficients for the subtests were generally lower, ranging from .54 to .93. VIQ, PIQ, and FSIQ scores significantly increased over the short retest interval, probably due to practice effects (Kaufman, 1994; Sattler, 1992). As seen in short-term stability studies on the WISC and WISC-R, the largest score gains were observed for the PIQ.

Only recently has long-term stability of the WISC-III received attention. Stavrou and Flanagan (1996) investigated the 3-year stability of the WISC-III among students with learning disabilities ($N = 50$) and found stability coefficients for VIQ, PIQ, and FSIQ scores of .76, .71, and .82, respectively. Mean VIQ, PIQ, and FSIQ test-retest differences were not significant. Zhu, Woodell, and Kreiman (1997) also examined the long-term stability of the WISC-III with a sample (N = 60) of students with learning disabilities. Using retest intervals from 32 to 48 months, stability coefficients for the VIQ, PIQ, and FSIQ were .79, .70, and .78, respectively. Zhu et al. (1997) found significant decreases in VIQ, PIQ, and FSIQ scores across the retest interval.

Using the WISC-III with students diagnosed with mild mental retardation, Bolen (1998) found significant stability coefficients over a 3-year retest interval. After correcting for restricted range at first testing, stability coefficients were .91, .81, and .92 for the VIQ, PIQ, and FSIQ, respectively. Bolen also found a significant decrease in VIQ across the retest interval that had a moderate effect strength. As expected, stability coefficients for subtests were lower than for IQs.

Cassidy (1997) found that WISC-III VIQ, PIQ, and FSIQ scores of a sample of exceptional children remained stable over a 3-year interval. Canivez and Watkins (1998) also studied the long-term stability of the WISC-III for a large sample (N = 667) of predominately disabled youths and found substantial stability for VIQ, PIQ, FSIQ, VCI, and POI scores (*r*s = .87, .87, .91, .85, and .85, respectively). Stability coefficients for FDI, PSI, and VIQ-PIQ discrepancy scores were lower, as were stability coefficients for the WISC-III subtests. Mean changes from first to second testing were either not significant or the effect strength was very low and of no practical consequence. Canivez and Watkins provided strong evidence of stability of the WISC-III; however, results were presented for the total sample and stability among demographic subgroups is as yet undetermined.

Differential stability of Wechsler scores across racial, gender, and age subgroups has rarely been assessed. One exception is Elliott et al. (1985), who used a 3-year WISC-R retest interval and found no differences among age groups. However, stability coefficients differed across race and gender categories. Anglos (Caucasians) had higher VIQ, PIQ, and FSIQ stability coefficients than Blacks and Mexican Americans (Hispanic/Latino), and females exhibited higher VIQ stability coefficients than males. Mean differences across the retest interval were not statistically compared, but frequency distributions suggested that the majority of individuals showed minimal changes in IQ scores across the 3-year retest interval.

There have been no substantial investigations of the stability of the WISC-III for students of diverse race, gender, and age. This information is vital to ensure nonbiased assessment (Rogers, 1998). Consequently, the purpose of the present study was to examine the long-term stability of the WISC-III IQ, Index, and VIQ-PIQ discrepancy scores within and between various demographic subgroups (gender, race/ethnicity, and age) obtained from a large, heterogeneous sample of predominately disabled children.

METHOD

Participants

Characteristics of participants for the total sample in the present study are presented in Table 1. The average test-retest interval in the present study was 2.83 years (SD = .55), with a range of .5 to 6.2 years. Only seven (1.1%) of the

Table 1
Demographic and Sample Characteristics at First and Second Testing

Variable	First Testing		Second Testing	
	n	%	*n*	%
Gender				
Male	433	67.4		
Female	209	32.6		
Race/Ethnicity				
Caucasian	502	78.2		
Black/African American	98	15.3		
Hispanic/Latino	42	6.5		
Age				
6	69	10.7	-	-
7	140	21.8	3	0.5
8	115	17.9	16	2.5
9	84	13.1	81	12.7
10	86	13.4	125	19.6
11	61	9.5	110	17.3
12	49	7.6	85	13.3
13	35	5.5	74	11.6
14	3	0.5	64	10.0
15	-	-	48	7.5
16	-	-	31	4.9
Disability				
LD	372	57.9	353	55.0
MIMR	60	9.3	52	8.1
ED	45	7.0	45	7.0
SLI	18	2.8	15	2.3
OHI	7	1.1	8	1.2
MOMR	4	0.6	7	1.1
Other	37	5.8	40	6.2
Not Disabled	19	3.0	40	6.2
Missing	80	12.5	82	12.8

Note.–LD = Learning Disabled, MIMR = Mild Mental Retardation, ED = Emotionally Disabled, SLI = Speech/Language Impaired, OHI = Other Health Impaired, MOMR = Moderate Mental Retardation. Other disabilities included low incidence disabilities such as Traumatic Brain Injury, Multiple Disabilities, Physical Disabilities, Autism, and Visual Impairment. Percents may not add to 100 due to rounding.

reevaluations occurred less than 1 year after the first evaluation. The mean age of students at first testing was 9.15 years (*SD* = 2.07) and ranged from 6.00 to 14.60 years. The mean age of students at second testing was 11.96 (*SD* = 2.12) and ranged from 7.50 to 16.90 years. Students were determined to be disabled (or not disabled) by multidisciplinary evaluation teams according to state and federal guidelines governing special education classification.

For the Caucasian group, 67.3% were male, the mean age at first testing was 8.81 (*SD* = 2.00) years, and the mean age at second testing was 11.63 (*SD* = 2.07) years. For the Black/African American group, 68.4% were male, the mean age at first testing was 9.26 (*SD* = 2.21) years, and the mean age at second testing was 11.97 (*SD* = 2.35) years. Among the Hispanic/Latino group, 66.7% were male, the mean age at first testing was 8.52 (*SD* = 2.10) years, and the mean age at second testing was 11.39 (*SD* = 2.10) years. Among the male students, 78.1% were Caucasian, 15.5% were Black/African American, and 6.5% were Hispanic/Latino. The mean age of males at first testing was 8.91 (*SD* = 2.06) years, while their mean age at second testing was 11.70 (*SD* = 2.14) years. Among female students, 78.5% were Caucasian, 14.8% were Black/African American, and 6.7% were Hispanic/Latino. The mean age for females at first testing was 8.76 (*SD* = 2.03), and their mean age at second testing was 11.60 (*SD* = 2.08) years.

Instrument

The Wechsler Intelligence Scale for Children-Third Edition (Wechsler, 1991) is an individually administered test of intelligence for children aged 6 years through 16 years 11 months. The WISC-III is comprised of 13 subtests that measure different dimensions of intelligence and yields three composite IQs—Verbal (VIQ), Performance (PIQ), and Full Scale (FSIQ)—that provide estimates of the individual's verbal, perceptual/nonverbal, and general intellectual abilities. The WISC-III also yields four optional factor-based index scores—Verbal Comprehension (VCI), Perceptual Organization (POI), Freedom from Distractibility (FDI), and Processing Speed (PSI)—based on exploratory and confirmatory factor analytic procedures. The WISC-III was standardized on a nationally representative sample ($N \doteq 2{,}200$) closely approximating the 1988 United States Census on gender, parent education (SES), race/ethnicity, and geographic region. Extensive evidence of reliability and validity is presented in the WISC-III manual (Wechsler, 1991).

Procedure

In order to obtain long-term stability data on the WISC-III with a sufficiently large and diverse sample, 2,000 school psychologists were randomly selected from the National Association of School Psychologists membership and invited to participate by anonymously providing test scores and demographic data obtained from recent special education reevaluations. Data were reported by 145 school psychologists from 33 states. Participating school psychologists selectively administered WISC-III subtests based upon the clinical demands of each case. As a consequence, sample sizes varied by IQ, Index, and VIQ-PIQ discrepancy scores.

RESULTS

For each demographic subgroup (gender, race, and age), Pearson product-moment correlation coefficients between first and second testing were calculated for WISC-III IQ, Index, and VIQ-PIQ discrepancy scores. Stability of VIQ-PIQ discrepancies was examined because it is a commonly calculated index (Kaufman, 1994; Sattler, 1992). Dependent *t* tests were conducted to investigate performance changes across the retest interval for each demographic subgroup. Due to the impact of the large sample sizes on statistical significance of the *t* tests, effect sizes for performance changes across the retest interval were calculated using Cohen's d statistic (Cohen, 1988). Stability coefficients between demographic subgroups were compared using independent z tests for differences between correlation coefficients using Fisher z transformations (Guilford & Fruchter, 1978). Frequency distributions were used to explore individual variations in scores across the retest interval.

Gender

Stability coefficients, descriptive statistics, t tests, and retest interval effect sizes (d) for the WISC-III IQ scores, Index scores, and VIQ-PIQ discrepancies by gender are presented in Table 2. Pearson product-moment correlation coefficients were all significant ($p < .0001$) for both male and female students. Additionally, dependent t tests for differences between means from first testing to second testing indicated significant decreases in the VIQ, PIQ, FSIQ, VCI, and POI for females and a significant increase in POI for males. Effect sizes for these differences were small, ranging from .08 to .12, indicating that the differences were not clinically meaningful. Comparisons of stability coefficients between males and females resulted in only one significant difference: The FDI stability coefficient for females ($r = .82$) was significantly higher than for males ($r = .71$), $z = 2.70$, $p < .007$.

Table 2
Test-Retest Correlations, Descriptive Statistics, t tests, and Retest Interval Effect Strengths by Gender

				First Testing		Second Testing				
	n	*r*	*p*	*M*	*SD*	*M*	*SD*	*t*	*p*	*d*
Female										
VIQ	209	.88	.0001	85.65	15.93	84.18	15.41	2.70	.007	.09
PIQ	207	.86	.0001	88.38	16.73	86.39	17.07	3.17	.002	.12
FSIQ	205	.92	.0001	85.67	16.37	83.85	16.42	3.86	.001	.11
VCI	201	.86	.0001	87.22	15.81	85.96	15.40	2.16	.032	.08
POI	195	.87	.0001	88.02	16.77	86.58	17.93	2.26	.025	.08
FDI	155	.82	.0001	83.31	15.28	82.78	13.91	0.75	.455	.04
PSI	58	.59	.0001	89.59	17.49	88.47	15.21	0.57	.573	.07
VIQ-PIQ	207	.55	.0001	−2.75	12.16	−2.27	11.68	0.62	.438	.04
Male										
VIQ	426	.85	.0001	90.65	15.56	90.40	15.59	0.61	.542	.02
PIQ	428	.87	.0001	91.93	16.70	92.59	17.66	1.53	.127	.04
FSIQ	424	.91	.0001	90.33	15.77	90.50	16.66	0.50	.620	.01
VCI	394	.84	.0001	92.27	15.64	92.15	15.60	0.27	.788	.01
POI	386	.86	.0001	93.31	16.72	95.13	17.85	3.90	.001	.11
FDI	295	.71	.0001	87.13	14.21	87.16	13.52	0.05	.961	.00
PSI	118	.65	.0001	93.91	15.03	92.42	14.33	1.31	.192	.10
VIQ-PIQ	425	.65	.0001	−1.32	14.11	−2.22	12.88	1.62	.107	.07

Note.—VIQ = Verbal IQ, PIQ = Performance IQ, FSIQ = Full Scale IQ, VCI = Verbal Comprehension Index, POI = Perceptual Organization Index, FDI = Freedom from Distractibility Index, PSI = Processing Speed Index, VIQ-PIQ = Verbal IQ-Performance IQ discrepancy.

Race/Ethnicity

Table 3 presents the stability coefficients, descriptive statistics, *t* tests, and retest interval effect sizes (*d*) for the WISC-III IQ scores, Index scores, and VIQ-PIQ discrepancies by race/ethnicity. All stability coefficients were significant ($p < .0001$). There were no significant changes in mean IQ scores, Index scores, or VIQ-PIQ discrepancies among Caucasian or Black/African American youths. Significant decreases in VIQ, FSIQ, and VCI were observed for Hispanic/Latino youths, but these differences represented small effect sizes. Stability coefficients between Caucasian, Hispanic/Latino, and Black/African American youths did not differ.

Table 3
Test-Retest Correlations, Descriptive Statistics, t tests, and Retest Interval Effect Strengths by Race

				First Testing		Second Testing				
	n	*r*	*p*	*M*	*SD*	*M*	*SD*	*t*	*p*	*d*
Caucasian										
VIQ	500	.86	.0001	90.56	15.88	90.09	15.73	1.27	.205	.03
PIQ	497	.86	.0001	92.57	16.97	92.54	17.76	0.07	.945	.00
FSIQ	495	.91	.0001	90.59	16.19	90.33	16.89	0.84	.401	.02
VCI	470	.85	.0001	92.00	15.89	91.70	15.81	0.77	.439	.02
POI	459	.86	.0001	93.41	16.96	94.21	18.35	1.82	.069	.05
FDI	357	.76	.0001	86.31	15.16	85.99	14.23	0.59	.554	.02
PSI	148	.68	.0001	91.82	16.33	91.28	15.54	0.52	.605	.03
VIQ-PIQ	497	.61	.0001	-2.03	13.73	-2.50	12.54	0.89	.372	.04
Black/African American										
VIQ	98	.83	.0001	83.27	14.13	82.67	13.83	0.72	.475	.04
PIQ	98	.87	.0001	83.11	14.92	82.37	15.89	0.92	.361	.05
FSIQ	97	.89	.0001	81.74	14.18	80.99	14.75	1.09	.280	.05
VCI	90	.83	.0001	85.49	14.93	85.02	14.08	0.51	.609	.03
POI	88	.87	.0001	83.42	15.35	84.24	16.37	0.99	.327	.05
FDI	67	.66	.0001	83.81	11.97	84.24	11.31	0.37	.713	.04
VIQ-PIQ	98	.65	.0001	0.15	12.35	0.31	11.75	0.15	.882	.01
Hispanic/Latino										
VIQ	37	.86	.0001	83.11	15.13	79.86	15.68	2.37	.023	.21
PIQ	40	.76	.0001	87.25	12.59	86.18	14.63	0.71	.481	.08
FSIQ	37	.87	.0001	83.46	13.43	80.81	14.33	2.22	.033	.19
VCI	35	.81	.0001	84.31	13.92	81.03	14.21	2.26	.030	.24
POI	34	.81	.0001	87.18	13.00	86.65	16.10	0.32	.749	.04
VIQ-PIQ	37	.62	.0001	-3.70	13.26	-5.43	12.97	0.92	.364	.13

Note.—VIQ = Verbal IQ, PIQ = Performance IQ, FSIQ = Full Scale IQ, VCI = Verbal Comprehension Index, POI = Perceptual Organization Index, FDI = Freedom from Distractibility Index, PSI = Processing Speed Index, VIQ-PIQ = Verbal IQ-Performance IQ discrepancy. Statistics not presented when the sample was less than 30.

Age

Table 4 presents the stability coefficients, descriptive statistics, *t* tests, and retest interval effect sizes (*d*) for the WISC-III IQ scores, Index scores, and VIQ-PIQ discrepancies by age at initial testing (6–13). All stability coefficients were significant, with the lowest stability observed among FDI, PSI, and VIQ-PIQ discrepancy scores. Significant differences across the retest interval were observed for VIQ (age 9), FSIQ (ages 9 and 13), POI (ages 11 and 13), FDI (ages 6, 9, and 10) and VIQ-PIQ (age 9). Effect sizes of these changes were generally quite small and, given their isolated nature, were not considered meaningful. As illustrated in Table 4, most of the correlations were quite similar in magnitude across the age dimension. Only 12 of the 207 stability coefficient comparisons between the eight age groups for IQ, Index, and VIQ-PIQ discrepancy scores were significant.

Table 4
Test-Retest Correlations, Descriptive Statistics, t tests, and Retest Interval Effect Strengths by Age at First Testing

				First Testing		Second Testing				
	n	*r*	*p*	*M*	*SD*	*M*	*SD*	*t*	*p*	*d*
Age 6										
VIQ	69	.80	.0001	88.67	15.79	88.84	17.40	0.14	.893	.01
PIQ	69	.87	.0001	92.38	17.85	90.45	18.38	1.74	.087	.11
FSIQ	69	.89	.0001	89.58	16.39	88.74	18.12	0.83	.411	.05
VCI	65	.81	.0001	91.68	15.44	90.11	17.73	1.21	.232	.09
POI	64	.88	.0001	92.06	17.36	90.42	19.45	1.39	.169	.09
FDI	53	.68	.0001	80.17	15.70	85.64	13.05	3.38	.001	.38
VIQ-PIQ	69	.58	.0001	-3.71	14.87	-1.61	12.80	1.37	.177	.15
Age 7										
VIQ	137	.86	.0001	91.33	15.77	90.65	15.32	0.98	.330	.04
PIQ	138	.81	.0001	92.91	15.70	92.24	16.60	0.79	.429	.04
FSIQ	136	.89	.0001	91.18	14.85	90.55	15.84	1.03	.303	.04
VCI	129	.84	.0001	92.73	15.79	91.86	15.70	1.12	.264	.06
POI	128	.80	.0001	92.32	15.18	92.91	16.99	0.64	.523	.04
FDI	96	.79	.0001	87.70	14.60	89.04	13.04	1.43	.155	.10
PSI	35	.55	.0010	97.03	16.76	96.29	13.06	0.30	.763	.05
VIQ-PIQ	137	.64	.0001	-1.51	15.44	-1.47	13.30	0.03	.973	.00
Age 8										
VIQ	115	.85	.0001	94.02	15.45	93.07	14.18	1.25	.213	.06
PIQ	114	.82	.0001	96.51	15.25	96.77	16.90	0.29	.775	.02
FSIQ	114	.87	.0001	94.54	15.22	94.10	15.39	0.60	.548	.03
VCI	109	.85	.0001	94.83	15.91	94.17	14.64	0.82	.414	.04
POI	105	.85	.0001	96.91	14.86	97.83	17.65	1.00	.320	.06
FDI	81	.72	.0001	88.96	14.62	87.65	12.76	1.14	.257	.10
PSI	38	.50	.0020	95.79	13.60	94.95	14.94	0.36	.719	.06
VIQ-PIQ	114	.63	.0001	-2.61	12.67	-3.87	12.55	1.25	.214	.10

(Table continues)

Table 4 (Continued)

				First Testing		Second Testing				
	n	*r*	*p*	*M*	*SD*	*M*	*SD*	*t*	*p*	*d*
Age 9										
VIQ	84	.88	.0001	89.60	16.85	87.26	16.18	2.58	.012	.14
PIQ	83	.91	.0001	90.47	17.98	90.51	18.77	0.04	.966	.00
FSIQ	81	.94	.0001	89.25	17.58	87.74	18.09	2.19	.032	.08
VCI	83	.87	.0001	90.59	16.97	89.02	16.17	1.69	.094	.09
POI	80	.91	.0001	93.00	18.18	92.95	19.33	0.06	.956	.00
FDI	59	.75	.0001	88.08	14.16	86.24	15.28	2.10	.040	.19
PSI	31	.71	.0001	89.23	13.34	86.84	15.19	1.21	.236	.17
VIQ-PIQ	83	.62	.0001	-0.58	13.14	-3.05	11.67	2.06	.042	.20
Age 10										
VIQ	86	.82	.0001	89.71	12.52	88.63	13.33	1.29	.200	.08
PIQ	85	.86	.0001	89.75	15.20	89.14	16.68	0.65	.516	.04
FSIQ	85	.90	.0001	88.51	13.25	87.51	14.85	1.41	.161	.07
VCI	82	.75	.0001	90.91	12.03	90.73	12.71	0.19	.851	.01
POI	78	.86	.0001	90.85	15.85	91.65	17.16	0.79	.430	.05
FDI	66	.75	.0001	86.08	12.28	82.55	13.20	3.16	.002	.28
VIQ-PIQ	85	.62	.0001	-0.29	12.70	-0.78	12.50	0.41	.686	.04
Age 11										
VIQ	61	.89	.0001	84.87	16.52	85.62	16.95	0.76	.448	.05
PIQ	61	.89	.0001	86.02	16.65	87.20	18.22	1.10	.278	.07
FSIQ	61	.94	.0001	84.05	16.80	85.13	17.93	1.33	.190	.06
VCI	54	.90	.0001	86.54	17.18	87.41	17.48	0.81	.422	.05
POI	54	.91	.0001	87.30	18.18	89.59	19.38	2.10	.041	.12
FDI	40	.79	.0001	83.20	13.22	83.15	13.12	0.04	.971	.00
VIQ-PIQ	61	.56	.0001	-1.15	11.97	-1.57	11.56	0.30	.764	.04
Age 12										
VIQ	47	.87	.0001	79.55	12.76	78.68	14.01	0.86	.395	.07
PIQ	49	.89	.0001	82.76	16.08	81.96	16.55	0.75	.458	.05
FSIQ	47	.92	.0001	79.28	14.22	78.40	14.84	1.04	.303	.06
VCI	41	.86	.0001	80.54	12.78	80.44	13.89	0.09	.930	.01
POI	41	.91	.0001	82.68	17.31	83.90	17.88	1.06	.297	.07
VIQ-PIQ	47	.74	.0001	-3.00	12.71	-2.87	12.50	0.10	.924	.01
Age 13										
VIQ	33	.93	.0001	81.21	16.92	82.79	17.77	1.36	.185	.09
PIQ	33	.87	.0001	83.94	18.20	86.24	17.35	1.47	.152	.13
FSIQ	33	.95	.0001	80.85	17.76	82.85	17.90	2.06	.047	.11
VCI	30	.89	.0001	84.00	16.70	86.20	16.53	1.58	.126	.13
VIQ-PIQ	33	.56	.0010	-2.73	12.47	-3.45	12.37	0.36	.723	.06

Note.—VIQ = Verbal IQ, PIQ = Performance IQ, FSIQ = Full Scale IQ, VCI = Verbal Comprehension Index, POI = Perceptual Organization Index, FDI = Freedom from Distractibility Index, PSI = Processing Speed Index, VIQ-PIQ = Verbal IQ-Performance IQ discrepancy. Statistics not presented when the sample was less than 30.

Individual variations in FSIQ scores across the retest interval for gender and race/ethnicity are presented in Table 5. FSIQ scores that differed by more than ±10 points were observed in 14.0% of females and 12.4% of males. Only 4.7%

of females and 3.1% of males had FSIQ scores that varied by more than ±15 points. Similar results were obtained for race/ethnicity, where 12.4% of Caucasians, 12.3% of Black/African Americans, and 16.2% of Hispanic/Latinos had FSIQ scores that varied by more than ±10 points. Only 3.6% of Caucasians, 4.1% of Black/African Americans, and 5.4% of Hispanic/Latinos had FSIQ differences greater than ±15 points.

As with gender and race/ethnicity, individual variations in FSIQ scores across the retest interval by age seemed reasonably stable. For youths aged 6 to 13 at the time of first testing, 2.1% to 20.0% showed FSIQ differences greater than ±10 points while 0% to 7.1% showed FSIQ changes greater than ±15 points. The greatest individual variations appeared to be present among the youngest ages (6-8 at initial testing).

DISCUSSION

The present study is the first to separately investigate long-term stability of the WISC-III among demographic subgroups. In contrast to the Elliott et al. (1985) study of differential long-term stability of the WISC-R for race, gender, and age, the present study did not find significant differences between stability coefficients for gender or race/ethnicity on VIQ, PIQ, FSIQ, VCI, POI, or VIQ-PIQ discrepancies. Few (12 of 207) stability coefficient comparisons between the eight age groups for IQ, Index, and VIQ-PIQ discrepancy scores were significant. Thus, it appeared that gender, race/ethnicity, and age had little differential effect on long-term stability coefficients for the WISC-III.

Long-term stability of the WISC-III's FSIQ appeared to be adequate for most diagnostic purposes for all demographic subgroups, because stability coefficients met the .85 to .90 criterion recommended by measurement experts (Hills, 1981; Salvia & Ysseldyke, 1991). Most of the demographic subgroup stability coefficients for VIQ, PIQ, VCI, and POI scores also demonstrated satisfactory reliability. However, demographic subgroup stability coefficients for FDI, PSI, and VIQ-PIQ discrepancy scores were not adequate for confident use with individuals. This result supplements the previously reported conclusions of Canivez and Watkins (1998) with the total sample.

To further explore how individual scores varied across the retest interval, frequency distributions of FSIQ changes were produced for each demographic subgroup. Only the FSIQ was examined for the demographic subgroups because Canivez and Watkins (1998) found that for all other IQ and Index scores, large percentages of individuals showed differences greater than ±15 points and only the FSIQ showed relatively stable change scores for individual students. This idiographic comparison showed that the WISC-III FSIQ was quite stable for the majority of individual students, with 80.0% to 97.9% of individuals showing changes of less than ±10 points and 92.9% to 100% of individuals showing changes of less than ±15 points, depending on the demographic subgroup (see Table 5). These results are similar to those reported by Elliott et al. (1985) and Stavrou (1990) in investigating the long-term stability of the WISC-R among students with disabilities, although greater percentages of their students showed significant FSIQ changes.

Table 5

Frequency Distributions (Percent) of WISC-III FSIQ Test-Retest Changes for Gender and Race

Δ	Female	Male	Caucasian	Black/African American	Hispanic/ Latino
-24		0.2	0.2		
-23		-	-		
-22	0.5	0.2	0.4		
-21	0.5	-	-		2.7
-20	0.5	0.2	-	2.1	-
-19	-	-	-	-	-
-18	-	-	-	-	-
-17	1.0	0.5	0.4	1.0	2.7
-16	-	0.2	0.2	-	-
-15	0.5	0.2	0.4	-	-
-14	1.0	1.4	1.0	2.1	2.7
-13	1.5	1.2	1.2	1.0	2.7
-12	2.4	1.4	1.8	1.0	2.7
-11	2.4	1.2	1.6	1.0	2.7
-10	2.0	2.1	2.0	3.1	-
-9	3.4	1.2	2.2	-	2.7
-8	3.9	2.6	3.0	2.1	5.4
-7	4.9	4.5	4.4	4.1	8.1
-6	3.4	3.5	4.0	1.0	2.7
-5	5.9	4.5	5.3	5.2	-
-4	4.9	4.5	4.4	6.2	2.7
-3	5.4	4.2	4.4	4.1	8.1
-2	4.4	5.9	4.8	8.2	5.4
-1	8.3	4.5	6.1	5.2	2.7
0	5.4	6.1	5.1	9.3	8.1
1	6.8	5.2	5.9	6.2	2.7
2	5.9	5.4	4.2	11.3	8.1
3	6.8	6.4	6.9	3.1	10.8
4	3.9	4.0	4.8	1.0	-
5	3.4	7.1	5.7	7.2	5.4
6	2.0	3.3	3.0	2.1	2.7
7	1.5	5.0	4.4	1.0	2.7
8	1.0	2.8	2.2	2.1	2.7
9	2.9	3.1	3.2	3.1	-
10	1.0	1.7	1.2	2.1	2.7
11	0.5	1.7	1.4	1.0	
12	0.5	0.7	0.6	1.0	
13	-	0.5	0.4	-	
14	-	0.9	0.6	1.0	
15	1.5	0.2	0.6	1.0	
16	-	0.2	0.2		
17	0.5	0.5	0.6		
18	-	0.5	0.4		
19	-	0.2	0.2		
20	-	-	-		
21	-	-	-		
22	-	-	-		
23	0.2	0.2	0.2		

Note.—Δ = Score Change. Column entries represent percentages of students' change in performance across the retest interval. Change in scores was determined by subtracting the initial obtained score from the most recent score. Columns may not sum to 100 due to rounding. Frequency distributions showing both increases and decreases in FSIQ, VIQ, PIQ, VCI, POI, FDI, PSI, and VIQ-PIQ scores across the retest interval for gender, race, and age may be obtained by writing the first author.

Changes across the retest interval for gender, race/ethnicity, and age were generally not statistically significant or resulted in effect sizes that were quite small. It was interesting that Hispanic/Latino students in the present study displayed VIQ, VCI, and FSIQ scores that decreased across the retest interval (by 3.25, 3.38, and 2.65 points, respectively). This is an interesting finding given the results in the Elliott et al. (1985) study, which found a mean WISC-R decrease of only 0.4 points in VIQ and a mean increase of 1.1 points in FSIQ among their Mexican American students. However, given the small sample sizes of Hispanic/Latino students in both studies, speculation as to the importance or causes of these changes should not be indulged. Further exploration of WISC-III stability is needed for larger samples of Hispanic/Latino students as well as for Asian American and Native American youths, which were not examined in the present study due to their small number.

These conclusions and recommendations must, however, be tempered by several limitations to the present study. First, generalization of these results is in part limited because these data were not the product of random selection and assignment. School psychologists chose to participate in response to a written request. They reported data from reevaluation cases that they personally selected. The large number of school psychologists ($n = 145$) from 33 different states who participated should, to some extent, mitigate this threat since it is unlikely that any one type of student would be preferentially or systematically selected.

A second limitation is that the use of reevaluation cases produced a situation where those students who were no longer enrolled in special education or those students who did not require reevaluation were not included in the sample. Generalization of these results to such students is not appropriate.

A third limitation is that the present sample consisted primarily of students with disabilities, particularly learning disabilities. Little is known about the long-term stability of the WISC-III among students without disabilities or differential effects for various disabilities. Future investigations should examine the stability of the WISC-III with normal youths as well as differential effects of disability type on long-term stability.

REFERENCES

Anderson, P. L., Cronin, M. E., & Kazmierski, S. (1989). WISC-R stability and re-evaluation of learning-disabled students. *Journal of Clinical Psychology, 45,* 941–944.

Bauman, E. (1991). Stability of WISC-R scores in children with learning difficulties. *Psychology in the Schools, 28,* 95–99.

Bolen, L. M. (1998). WISC-III score changes for EMH students. *Psychology in the Schools, 35,* 327–332.

Canivez, G. L., & Watkins, M. W. (1998). Long term stability of the WISC-III. *Psychological Assessment, 10,* 285–291.

Cassidy, L. C. (1997). *The stability of WISC-III scores: For whom are triennial re-evaluations necessary?* Unpublished doctoral dissertation, University of Rhode Island.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Erlbaum.

Coleman, J. C. (1963). Stability of intelligence test scores in learning disorders. *Journal of Clinical Psychology, 19,* 295–298.

Conklin, R. C., & Dockrell, W. B. (1967). The predictive validity and stability of WISC scores over a four-year period. *Psychology in the Schools, 4,* 263–266.

Covin, T. M. (1977). Stability of the WISC-R for 9-year-olds with learning difficulties. *Psychological Reports, 40,* 1297–1298.

Elliott, S. N., & Boeve, K. (1987). Stability of WISC-R IQs: An investigation of ethnic differences over time. *Educational and Psychological Measurement, 47,* 461–465.

Elliott, S. N., Piersol, W. C., Witt, J. C., Argulewicz, E. N., Gutkin, T. B., & Galvin, G. A. (1985). Three-year stability of WISC-R IQs for handicapped children from three racial/ethnic groups. *Journal of Psychoeducational Assessment, 3,* 233–244.

Ellzey, J. T., & Karnes, F. A. (1990). Test-retest stability of WISC-R IQs among young gifted students. *Psychological Reports, 66,* 1023–1026.

Friedman, R. (1970). The reliability of the Wechsler Intelligence Scale for Children in a group of mentally retarded children. *Journal of Clinical Psychology, 26,* 181–182.

Gehman, I. H., & Matyas, R. P. (1956). Stability of the WISC and Binet tests. *Journal of Consulting Psychology, 20,* 150–152.

Guilford, J. P., & Fruchter, B. (1978). *Fundamental statistics in psychology and education.* New York: McGraw-Hill.

Haynes, J. P., & Howard, R. C. (1986). Stability of WISC-R scores in a juvenile forensic sample. *Journal of Clinical Psychology, 42,* 534–537.

Hills, J. R. (1981). *Measurement and evaluation in the classroom* (2nd ed.). Columbus, OH: Merrill.

Irwin, D. O. (1966). Reliability of the WISC. *Journal of Educational Measurement, 3,* 287–292.

Jensen, A. R. (1980). *Bias in mental testing.* New York: Free Press.

Kaufman, A. S. (1994). *Intelligent testing with the WISC-III.* New York: Wiley.

McDermott, P. A. (1988). Agreement among diagnosticians or observers: Its importance and determination. *Professional School Psychology, 3,* 225–240.

Moffitt, T. E., Caspi, A., Harkness, A. R., & Silva, P. A. (1993). The natural history of change in intellectual performance: Who changes? How much? Is it meaningful? *Journal of Child Psychology and Psychiatry, 34,* 455–506.

Naglieri, J. A., & Pfeiffer, S. I. (1983). Reliability and stability of the WISC-R for children with below average IQs. *Educational and Psychological Research, 3,* 203–208.

Oakman, S., & Wilson, B. (1988). Stability of WISC-R intelligence scores: Implications for 3-year reevaluations of learning disabled students. *Psychology in the Schools, 25,* 118–120.

Quereshi, M. J. (1968). Practice effects of the WISC subtest scores and IQ estimates. *Journal of Clinical Psychology, 24,* 79–85.

Reger, R. (1962). Repeated measurement with the WISC. *Psychological Reports, 11,* 418.

Rogers, M. R. (1998). Psychoeducational assessment of culturally and linguistically diverse children and youth. In H. B. Vance (Ed.), *Psychological assessment of children: Best practices for school and clinical settings* (2nd ed.; pp. 355–384). New York: Wiley.

Rosen, M., Stallings, L., Floor, L., & Nowakiwska, M. (1968). Reliability and stability of Wechsler IQ scores for institutionalized mental subnormals. *American Journal of Mental Deficiency, 73,* 218–225.

Salvia, J., & Ysseldyke, J. E. (1991). *Assessment* (5th ed.). Boston: Houghton Mifflin.

Sattler, J. (1992). *Assessment of children* (Rev. 3rd ed.). San Diego, CA: Author.

Smith, M. D. (1978). Stability of WISC-R subtest profiles for learning-disabled children. *Psychology in the Schools, 15,* 4–7.

Stavrou, E. (1990). The long-term stability of WISC-R scores in mildly retarded and learning-disabled children. *Psychology in the Schools, 27,* 101–110.

Stavrou, E., & Flanagan, R. (1996, March). *The stability of WISC-III scores in learning disabled children.* Paper presented at the Annual Convention of the National Association of School Psychologists, Atlanta, GA.

Stinnett, T. A., Havey, J. M., & Oehler-Stinnett, J. (1994). Current test usage by practicing school psychologists: A national survey. *Journal of Psychoeducational Assessment, 12,* 331–350.

Throne, F. M., Schulman, J. L., & Kaspar, J. C. (1962). Reliability and stability of the WISC for a group of mentally retarded boys. *American Journal of Mental Deficiency, 67,* 455–457.

Truscott, S. D., Narrett, C. M., & Smith, S. E. (1994). WISC-R subtest reliability over time: Implications for practice and research. *Psychological Reports, 74,* 147–156.

Tuma, J. M., & Appelbaum, A. S. (1980). Reliability and practice effects of WISC-R IQ estimates in a normal population. *Educational and Psychological Measurement, 40,* 671–678.

Vance, H. B., Blixt, S., Ellis, R., & Debell, S. (1981). Stability of the WISC-R for a sample of exceptional children. *Journal of Clinical Psychology, 37,* 397–399.

Vance, H. B., Hankins, N., & Brown, W. (1987). A longitudinal study of the Wechsler Intelligence Scale for Children-Revised over a six-year period. *Psychology in the Schools, 24,* 229–233.

Walker, K. P., & Gross, F. L. (1970). IQ stability among educable mentally retarded children. *Training School Bulletin, 66,* 181–187.

Watkins, C. E., Jr., Campbell, V. L., Nieberding, R., & Hallmark, R. (1995). Contemporary practice of psychological assessment by clinical psychologists. *Professional Psychology: Research and Practice, 26,* 54–60.

Webster, R. E. (1988). Statistical and individual temporal stability of the WISC-R for cognitively disabled adolescents. *Psychology in the Schools, 25,* 365–372.

Wechsler, D. (1949). *Manual for the Wechsler Intelligence Scale for Children.* New York: The Psychological Corporation.

Wechsler, D. (1974). *Manual for the Wechsler Intelligence Scale for Children-Revised.* New York: The Psychological Corporation.

Wechsler, D. (1991). *Manual for the Wechsler Intelligence Scale for Children-Third Edition.* San Antonio, TX: The Psychological Corporation.

Whatley, R. G., & Plant, W. T. (1957). The stability of the WISC IQs for selected children. *Journal of Psychology, 44,* 165–167.

Whorton, J. E. (1985). Test-retest Wechsler Intelligence Scale for Children-Revised scores for 310 educable mentally retarded and specific learning disabled students. *Psychological Reports, 56,* 857–858.

Zhu, J., Woodell, N. M., & Kreiman, C. L. (1997, August). *Three-year re-evaluation stability of the WISC-III: A learning disabled sample.* Paper presented at the Annual Convention of the American Psychological Association, Chicago.