Psychological Assessment
1998, Vol. 10, No. 3, 285–291

1040-3590/98/$3.00

Long-Term Stability of the Wechsler Intelligence Scale for Children—Third Edition

Gary L. Canivez
Eastern Illinois University

Marley W. Watkins
Pennsylvania State University

Long-term stability of the Wechsler Intelligence Scale for Children—Third Edition (WISC–III; D. Wechsler, 1991) was investigated with a sample of 667 students from 33 states twice evaluated for special education consideration. With an average test–retest interval of 2.87 years, test–retest reliability coefficients for the Verbal IQ, Performance IQ, and Full Scale IQ were .87, .87, and .91, respectively ($p < .0001$). As expected, test–retest reliability coefficients for the subtests were generally lower than for global IQ and factor index scores. Mean differences from first testing to second testing were either not statistically significant or not clinically meaningful. Results provided the highest estimates of long-term stability for the WISC–III yet reported.

Surveys of test use by clinical and school psychologists have consistently found the Wechsler scales to be the most frequently used tests of cognitive abilities (Goh, Teslow, & Fuller, 1981; Hutton, Dubes, & Muir, 1992; Stinnett, Havey, & Oehler-Stinnett, 1994; Watkins, Campbell, Nieberding, & Hallmark, 1995). It is common practice for school psychologists to readminister comprehensive intelligence tests in triennial special education reevaluations, thereby providing opportunities for investigation of long-term stability. Stability of intelligence tests is an important characteristic as intelligence as a construct is presumed to be an enduring trait.

Research with the Wechsler Intelligence Scale for Children (WISC; Wechsler, 1949) and the Wechsler Intelligence Scale for Children—Revised (WISC–R; Wechsler, 1974) produced short-term test–retest reliability coefficients for the Verbal IQ (VIQ), Performance IQ (PIQ), and Full Scale IQ (FSIQ) scores in the .80s and .90s (Covin, 1977; Irwin, 1966; Quereshi, 1968; Throne, Schulman, & Kaspar, 1962; Tuma & Appelbaum, 1980; Wechsler, 1974). However, significant practice effects were reflected in higher IQ scores at retest, especially for the PIQ. Additionally, WISC and WISC–R subtest test–retest reliability coefficients were almost always lower than global IQ test–retest reliability coefficients.

Long-term stability of 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 et al., 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) has been thoroughly investigated. Significant and moderate to high test–retest reliability coefficients (*r*s generally ranging from the .50s to .90s) have been reported. More important, practice effects seemingly disappeared when the retest interval was greater than 1 year. When practice effects were observed in long-term stability studies, the effect sizes were usually quite small and of no practical consequence. Juliano, Haddad, and Carroll (1988) also found significant long-term stability for the WISC–R factor structure among youths with learning disability.

In contrast to the WISC and WISC–R, stability of Wechsler Intelligence Scale for Children—Third Edition (WISC–III; Wechsler, 1991) scores across time has received little attention. Short-term stability of the WISC–III with a sample of 353 normal children was reported in the WISC–III manual (Wechsler, 1991) for a test–retest interval ranging from 12–63 days (*Mdn* = 23 days). Test–retest reliability estimates for the three IQ and four factor index scores were generally excellent, ranging from .71 (FDI for ages 6–7) to .95 (FSIQ for ages 14–15). Test–retest reliability coefficients for the subtests were lower and ranged from .54 (Mazes for ages 14–15) to .93 (Vocabulary for ages 14–15). Significant increases in VIQ, PIQ, and FSIQ scores were found and attributed to practice effects or exposure to test materials (reduced novelty) due to the short-time interval (Kaufman, 1994; Sattler, 1992). The largest score gains were noted for the PIQ, results that were also found in short-term stability studies on the WISC and WISC–R.

Long-term stability of the WISC–III has only recently been investigated. Stavrou and Flanagan (1996, March) found sig-

Gary L. Canivez, Department of Psychology, Eastern Illinois University; Marley W. Watkins, Department of Educational and School Psychology and Special Education, Pennsylvania State University.

This research was supported, in part, by 1995-1996 Eastern Illinois University Faculty Development Grant and 1995-1996 Pennsylvania State University College of Education Alumni Society Faculty Research Initiation Grant.

We thank Tim Runge, Lisa Samuels, and Daniel Heupel for assistance in data entry and preliminary analyses.

Correspondence concerning this article should be addressed to Gary L. Canivez, Department of Psychology, Eastern Illinois University, 600 Lincoln Avenue, Charleston, Illinois 61920-3099. Electronic mail may be sent to cfglc@eiu.edu or to http://www.ux1.eiu.edu/~cfglc on the World Wide Web.

nificant test–retest reliability coefficients for VIQ, PIQ, and FSIQ scores among students with learning disabilities (n = 50) retested at a 3-year interval (rs = .76, .71, and .82, respectively). No significant differences between first and second testings in VIQ, PIQ, or FSIQ scores were observed. Zhu, Woodell, and Kreiman (1997, August) also examined the long-term stability of the WISC–III with a sample (n = 60) of 6- to 12-year-old students with learning disabilities. A retest interval from 32–48 months resulted in test–retest reliability coefficients for the VIQ, PIQ, and FSIQ of .79, .70, and .78, respectively. Significant decreases in VIQ, PIQ, and FSIQ scores were found from first to second testing. Subtest test–retest reliability coefficients ranged from .34 (Arithmetic) to .69 (Information). Significant decreases from first to second testing were found for the Similarities, Vocabulary, Comprehension, and Object Assembly subtests. Investigation of VIQ–PIQ discrepancies resulted in a test–retest reliability coefficient of .67.

The purpose of the present study was to investigate the long-term stability of the WISC–III IQ, index, and subtest scores with a large, heterogeneous sample of disabled children. This study also investigated the stability of VIQ–PIQ discrepancies, an analysis lacking in most investigations of WISC stability.

Method

Participants

Demographic information and sample characteristics of participants at first and second testing are presented in Table 1. The mean age of students at first testing was 9.18 years (SD = 2.06), with a range of 5.80 to 14.60 years. The mean age of students at second testing was 11.99 (SD = 2.12), with a range of 7.50 to 16.90 years. The mean test–retest interval was 2.83 years (SD = 0.55), with a range of 0.5 to 6.2 years. Only seven (1%) of the reevaluations occurred less than 1 year following the first evaluation. Most students were classified as disabled according to state and federal guidelines governing special education classification.

Instrument

The Wechsler Intelligence Scale for Children—Third Edition (Wechsler, 1991) is an individually administered test of intelligence for children of ages 6 years through 16 years, 11 months. As with previous editions, the WISC–III comprises several subtests that measure different aspects of intelligence and yield three composite IQs (viz., VIQ, PIQ, and FSIQ), which provide estimates of the individual's verbal, perceptual–nonverbal, and general intellectual abilities. Additionally, the WISC–III yields four optional factor-based index scores (viz., Verbal Comprehension Index [VCI], Perceptual Organization Index [POI], Freedom From Distractibility Index [FDI], and Processing Speed Index [PSI]). The WISC–III was standardized on a representative sample (N = 2,200) closely approximating the 1988 U. S. Census on gender, parent education socioeconomic status (SES), race–ethnicity, and geographic region. Extensive evidence of reliability (internal consistency and short-term test–retest) and validity (criterion related and construct) is presented in the WISC–III manual (Wechsler, 1991).

Procedure

To obtain a large sample of test–retest data on the WISC–III, we randomly selected 2,000 school psychologists from the National Association of School Psychologists membership and invited them to participate by providing test scores and demographic data extracted from recent special education reevaluations. Data on 667 students were reported by 145 school psychologists in 33 states. Some scores were not routinely reported (i.e., factor index scores) so when subtest data were available, these were calculated on the basis of the reported subtest scores. In addition, certain disabilities (i.e., physical disability, deaf–hearing impaired, blind–visually impaired) prevented administration of specific subtests pertaining to the VIQ or PIQ, and thus, the FSIQ could not be calculated or reported. For these reasons, sample sizes varied by IQ, index, and subtest scores.

Results

Pearson product–moment correlation coefficients between first and second testing were calculated for the WISC–III IQ,

Table 1
Demographic and Sample Characteristics at First and Second Testings

	First testing		Second testing	
Variable	n	%	n	%
Gender				
Boys	452	67.8		
Girls	215	32.2		
Race–Ethnicity				
Caucasian	508	76.2		
Hispanic–Latino	42	6.3		
Black–African American	98	14.7		
Native American–American Indian	4	0.6		
Asian American	1	0.1		
Other–Missing	14	2.1		
Grade				
Kindergarten	25	3.7	—	—
1	120	18.0	1	0.2
2	158	23.7	13	1.9
3	100	15.0	45	6.7
4	83	12.4	111	16.6
5	81	12.1	153	22.9
6	47	7.0	93	13.9
7	30	4.5	78	11.7
8	12	1.8	75	11.2
9	2	0.3	54	8.1
10	—	—	26	3.9
11	—	—	9	1.3
Missing	9	1.3	9	1.3
Disability				
Not disabled	19	2.8	40	6.0
LD	391	58.6	368	55.2
ED	47	7.0	48	7.2
MIMR	62	9.3	54	8.1
SLI	19	2.8	16	2.4
OHI	7	1.0	8	1.0
MOMR	4	0.6	7	1.0
Other	37	5.5	40	6.0
Missing	81	12.1	86	12.9

Note. LD = learning disabled; ED = emotionally disabled; MIMR = mild mental retardation; 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. Percentages may not add to 100 because of rounding.

Table 2
Descriptive Statistics, t Tests, Effect Strengths, and Test–Retest Reliability Coefficients

		First testing		Second testing				
Scale	*n*	*M*	*SD*	*M*	*SD*	*t*	η^2	*r*
IQ scores								
VIQ	660	88.99	15.83	88.35	15.79	2.00*	.01	.87
PIQ	660	91.00	16.86	90.73	17.83	0.79	.00	.87
FSIQ	654	88.92	16.13	88.41	16.94	1.86	.01	.91
Index scores								
VCI	618	90.62	15.84	90.09	15.79	1.53	.00	.85
POI	604	91.85	17.03	92.58	18.44	1.95	.01	.87
FDI	464	85.65	14.69	85.59	13.76	0.11	.00	.75
PSI	182	92.72	16.07	90.84	14.67	1.88	.02	.62
Subtests								
PC	615	8.7	3.3	9.0	3.4	3.33***	.02	.66
I	619	7.7	3.1	8.0	3.2	2.61**	.01	.73
CD	611	8.3	3.4	7.7	3.2	5.60***	.05	.63
S	621	8.3	3.4	8.4	3.2	1.78	.01	.68
PA	618	8.4	3.6	8.6	3.9	1.29	.00	.68
A	618	7.3	3.1	7.2	3.0	0.68	.00	.67
BD	617	8.4	3.7	8.3	4.0	1.08	.00	.78
V	618	8.0	3.2	7.5	3.1	5.91***	.05	.75
OA	599	8.4	3.4	8.5	3.6	0.64	.00	.68
C	609	8.6	3.7	8.4	3.5	2.10*	.01	.68
SS	181	8.5	3.8	8.8	3.4	0.94	.00	.55
DS	458	7.3	2.7	7.4	2.8	0.55	.00	.65

Note. PC = Picture Completion; I = Information; CD = Coding; S = Similarities; PA = Picture Arrangement; A = Arithmetic; BD = Block Design; V = Vocabulary; OA = Object Assembly; C = Comprehension; SS = Symbol Search; DS = Digit Span; 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. All correlations were significant at $p < .0001$.
$^{*}p < .05$. $^{**}p < .01$. $^{***}p < .001$.

index, and subtest scores, as well as for VIQ–PIQ discrepancies.[1] Dependent *t* tests were conducted to investigate performance changes from test to retest. Because of the impact of the large sample size on statistical significance of the *t* tests, effect strengths of performance changes across the retest interval were estimated using η^2, an index of the proportion of variability explained by the effect across the retest interval (Kiess, 1996). Individual variation in scores across the test–retest interval was explored with use of cumulative frequency distributions.

Descriptive statistics, *t* tests, retest interval effect strengths (η^2), and test–retest reliability coefficients for the WISC–III IQ scores, index scores, and subtest scores are presented in Table 2. Pearson product–moment correlation coefficients for the VIQ ($r = .87$), PIQ ($r = .87$), and FSIQ ($r = .91$) were all significant ($p < .0001$) and indicated substantial long-term stability. Additionally, dependent *t* tests for differences between means from first testing to second testing were not significant for the FSIQ or PIQ and effect strengths were negligible. Although the decrease of less than 1 IQ point in VIQ from first testing ($M = 88.99$) to second testing ($M = 88.35$) was statistically significant, $t(659) = 2.00$, $p = .046$, the effect strength ($\eta^2 = .01$) indicated that this difference was not clinically meaningful.

WISC–III Factor Index scores (VCI, POI, FDI, PSI) also possessed substantial long-term stability with significant correlations of .85, .87, .75, and .62, respectively ($p < .0001$). Mean performance on these index scores from first testing to second testing did not differ, and effect strengths were negligible.

As expected, test–retest reliability coefficients for the WISC–III subtests were generally lower than the IQ and Factor Index scores, ranging from .55 (Symbol Search) to .78 (Block Design) and resulting in a median $r = .68$. As with the IQ and index score correlations, all subtest stability coefficients were statistically significant, $p < .0001$ (see Table 2). Dependent *t* tests revealed statistically significant increases from first to second testing on the Picture Completion and Information subtests and significant decreases from first to second testing on the Coding, Vocabulary, and Comprehension subtests. However, statistical significance was likely due to the large sample size as all effect strengths were small and differences were judged not clinically meaningful. Figure 1 presents the mean WISC–III subtest profiles at first and second testing to better illustrate mean subtest variation across time.

An additional analysis investigated the stability of VIQ–PIQ discrepancies, a commonly calculated index (Kaufman, 1994; Sattler, 1992). The test–retest reliability coefficient ($r = .62$) was statistically significant, $p < .0001$, but lower than stability

[1] Some data were not reported by participating school psychologists or were not available because of selective administration of subtests related to specific disabilities, therefore, pairwise elimination was used to allow for the maximum sample size in analyses.

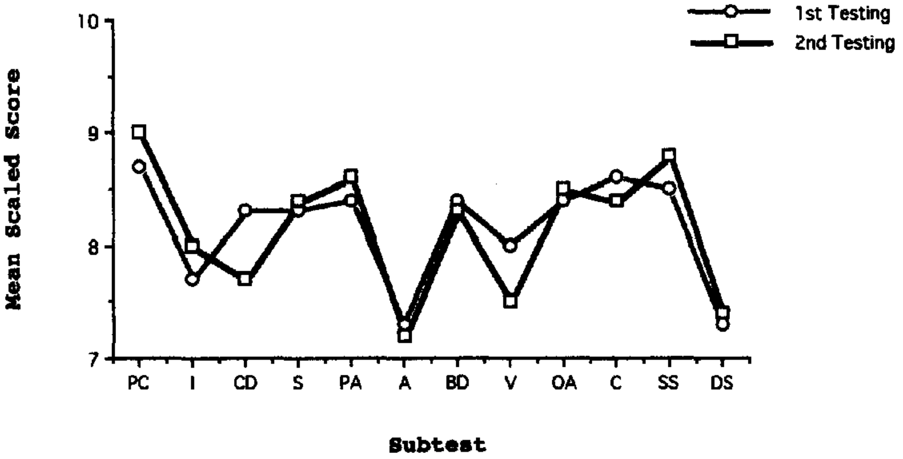

Figure 1. Wechsler Intelligence Scale for Children—Third Edition mean subtest score profiles for first and second testings. (See Table 2 for subtest names.)

coefficients for IQ and index scores. There was no significant difference between VIQ–PIQ discrepancy scores between the first and second testings, and the effect strength was negligible ($\eta^2 = .00$).

Individual variations in scores across the test–retest interval are presented in cumulative frequency distributions in Table 3. Only 13% of the students earned FSIQ scores that differed by more than ±10 points, and fewer than 3% of the students earned FSIQ scores that differed by more than ±15 points. However, 19%, 26%, 28%, and 42% earned VCI, POI, FDI, and PSI scores, respectively, which varied by ±10 or more points. FSIQ test–retest scores diverged by as much as 24 points, VIQ scores diverged by 31 points, PIQ scores diverged by 29 points, VCI and POI scores diverged by 30 points, FDI scores diverged by 36 points, and PSI scores diverged by as much as 43 points. Variation in VIQ–PIQ discrepancies was also observed, with 35% obtaining changes of ±10 points or more and changing as much as 45 points across the test–retest interval. Descriptive statistics presented in Table 4 indicate that the changes in IQ, index, and VIQ–PIQ discrepancies across the retest interval appear to be normally distributed.

Discussion

The long-term WISC–III test–retest reliability coefficients in this sample of predominately disabled children ranged from .55 to .78 for subtests and from .62 to .91 for IQ and index scores. One implication of these findings is that WISC–III scores appeared to be more stable over a 2–3-year time span for disabled students than was the WISC (Coleman, 1963; Conklin & Dockrell, 1967; Friedman, 1970; Gehman & Matyas, 1956; Reger, 1962; Rosen et al., 1968; Walker & Gross, 1970; Whatley & Plant, 1957). The test–retest reliability coefficients were among the highest obtained with the WISC–R (Bauman, 1991; Haynes & Howard, 1986; Webster, 1988) and higher than most obtained with the WISC–R (Anderson et al., 1989; Elliott et al., 1985; Ellzey & Karnes, 1990; Naglieri & Pfeiffer, 1983; Oakman & Wilson, 1988; Stavrou, 1990; Truscott et al., 1994; Vance et al., 1981; Vance et al., 1987; Whorton, 1985) scores. The long-term test–retest reliability coefficients found in the present study are more similar to those obtained in previous studies of short-term stability (Covin, 1977; Irwin, 1966; Quereshi, 1968; Throne, Schulman, & Kaspar, 1962; Tuma & Appelbaum, 1980; Wechsler, 1974, 1991).

The results of this study are consistent with those of Stavrou and Flanagan (1996, March) and Zhu et al. (1997, August) except that they reported somewhat lower test–retest reliability coefficients and that Zhu et al. (1997, August) found significant decreases in VIQ, PIQ, FSIQ, and specific subtests among their students with learning disabilities. Decreases in VIQ, PIQ, or FSIQ were also reported in several WISC–R stability studies involving students with learning disabilities (Bauman, 1991; Elliott & Boeve, 1987; Elliott et al., 1985; Stavrou, 1990; Vance et al., 1981). Follow-up analyses with the 298 students in the present study who maintained a learning disability diagnosis across both test administrations resulted in lower stability coefficients (i.e., FSIQ dropped from .91 to .86, VIQ dropped from .87 to .81, and PIQ dropped from .87 to .80) but not of the magnitude reported by Zhu et al.; nor were significant or meaningful changes in mean levels across the retest interval observed in this sample. These discrepant results might be attributable to sample variation, but further investigation is required.

Long-term stability of the WISC–III's VIQ, PIQ, VCI, POI, and FSIQ scores appear to be adequate for most diagnostic purposes, approaching or exceeding the .90 criterion recommended by Salvia and Ysseldyke (1991). Stability coefficients of the FDI, PSI, VIQ–PIQ discrepancy, and subtest scores were not of sufficient magnitude for confident use with individuals.

Although group subtest profiles (see Figure 1) and mean IQ, index, and subtest levels (see Table 2) are similar, these provide a nomothetic rather than an idiographic perspective. That is, an individual's scores might deviate even though group averages and profiles are stable. This supposition was supported by the

Table 3

Cumulative Frequency Distributions (in Percentages) of Wechsler Intelligence Scale for Children—Third Edition IQ, Index Score, and Verbal IQ–Performance IQ (VIQ–PIQ) Test–Retest Changes

Δ	FSIQ	VIQ	PIQ	VCI	POI	FDI	PSI	VIQ–PIQ
0	5.7	6.8	5.5	6.1	7.5	15.3	9.3	4.6
1	16.8	18.5	14.4	14.7	12.4	15.3	9.3	12.8
2	27.7	29.2	21.8	22.8	19.7	17.2	15.9	20.7
3	38.8	37.7	30.2	33.7	27.6	36.0	23.1	26.6
4	47.6	45.6	39.1	46.1	35.1	36.0	23.1	32.7
5	58.4	52.0	47.6	51.6	42.4	39.9	34.6	39.0
6	65.0	59.7	52.9	57.6	51.5	55.2	39.6	46.7
7	73.7	66.2	59.8	65.0	59.1	55.2	41.2	51.6
8	79.1	72.0	65.9	71.4	64.4	61.2	52.2	58.3
9	84.1	78.5	70.3	76.4	70.2	72.0	52.2	61.8
10	87.5	82.9	75.2	81.1	74.5	72.2	57.7	65.4
11	90.2	85.5	78.6	84.3	77.5	77.2	65.4	70.2
12	92.5	87.7	83.5	87.7	81.3	81.7	65.9	74.1
13	94.2	90.0	87.3	89.6	84.8	82.3	73.6	76.9
14	96.2	91.4	89.2	91.3	87.7	85.8	74.2	81.0
15	97.1	93.2	91.4	92.4	89.9	90.1	76.4	84.3
16	97.4	94.1	92.4	93.9	92.9	90.7	81.3	86.5
17	98.5	96.1	94.7	94.5	95.0	93.3	81.3	88.9
18	98.8	97.3	96.1	95.6	96.2	94.4	83.0	90.4
19	98.9	97.7	97.0	96.1	96.7	94.6	83.0	91.6
20	99.2	98.6	97.6	97.2	97.8	95.5	84.1	92.7
21	99.4	98.8	98.3	97.7	98.2	95.9	85.7	94.4
22	99.7	98.8	98.6	98.2	98.3	95.9	87.9	95.0
23	99.8	98.9	98.8	98.5	98.8	97.0	89.6	96.5
24	100.0	99.2	99.1	99.0	99.2	97.2	91.8	97.3
25		99.4	99.2	99.0	99.5	97.8	92.3	97.7
26		99.4	99.5	99.0	99.8	98.3	94.5	97.9
27		99.7	99.5	99.4	99.8	98.3	95.6	97.9
28		99.7	99.7	99.5	99.8	98.3	95.6	98.0
29		99.7	100.0	99.8	99.8	98.9	96.7	98.2
30		99.8		100.0	100.0	98.9	96.7	98.3
31		100.0				99.6	97.8	98.9
32						99.8	98.4	99.1
33						99.8	98.9	99.1
34						99.8	98.9	99.5
35						99.8	98.9	99.7
36						100.0	99.5	99.7
37							99.5	99.7
38							99.5	99.7
39							99.5	99.7
40							99.5	99.7
41							99.5	99.7
42							99.5	99.7
43							100.0	99.7
44								99.7
45								100.0

Note. Column entries represent cumulative percentages of students' change in performance across the retest interval (±). Change in scores was determined by subtracting the most recent score from the initial obtained score. Frequency distributions showing both increases and decreases in FSIQ, VIQ, PIQ, VCI, POI, FDI, PSI, and VIQ–PIQ scores across the retest interval may be obtained by contacting Gary L. Canivez (see Author Note). Δ = absolute score change; FSIQ = Full Scale IQ; VIQ = Verbal IQ; PIQ = Performance IQ; VCI = Verbal Comprehension Index; POI = Perceptual Organization Index; FDI = Freedom From Distractibility Index; PSI = Processing Speed Index; VIQ–PIQ = VIQ–PIQ Discrepancy.

frequency distributions in Table 3. Global IQ and index scores differed by as much as 24 to 43 points across the retest interval. Large percentages of students earned FDI (28%) and PSI (42%) scores, which differed by more than ±10 points. Only the FSIQ produced relatively stable test–retest scores for individual students; only 13% of the students' test–retest FSIQ scores differed by more than ±10 points and only 3% varied by more than ±15 points. These results are similar to those found by Stavrou (1990) in investigating the stability of the WISC–R among students with learning disability or mild mental retardation, al-

Table 4
Descriptive Statistics for Wechsler Intelligence Scale for Children—Third Edition IQ, Index Score, and Verbal IQ–Performance IQ (VIQ–PIQ) Test–Retest Changes

Scale	*M*	*SD*	sk	SE_{sk}	Minimum	Maximum
FSIQ	0.51	6.99	.12	.10	−23	24
VIQ	0.64	8.16	.10	.10	−31	30
PIQ	0.28	9.06	.00	.10	−29	29
VCI	0.53	8.57	.16	.10	−29	30
POI	−0.73	9.24	.06	.10	−30	25
FDI	0.05	10.03	−.09	.11	−32	36
PSI	1.88	13.50	.04	.18	−32	43
VIQ–PIQ	0.36	11.33	−.08	.10	−45	45

Note. Change in scores was determined by subtracting the most recent score from the initial obtained score. sk = skewness; FSIQ = Full Scale IQ; VIQ = Verbal IQ; PIQ = Performance IQ; VCI = Verbal Comprehension Index; POI = Perceptual Organization Index; FDI = Freedom From Distractibility Index; PSI = Processing Speed Index.

though greater numbers of their students showed significant VIQ, PIQ, and FSIQ changes.

Limitations of this study must, however, temper conclusions and recommendations. First, generalization of these results is in part limited on the basis of the low response rate (7%) as only 145 of the 2,000 school psychologists randomly sampled provided data for analyses. Thus, WISC–III data obtained in this research were not the product of random selection and assignment. Rather, school psychologists chose to participate in response to the request and then reported data from specific reevaluation cases they selected. The large number of school psychologists who participated should, to some extent, ameliorate this threat because it is unlikely that any one type of student would be preferentially or systematically selected by more than 100 professionals. A second limitation is that the use of reevaluation cases created a situation where certain students were ineligible for participation; that is, those students who were no longer enrolled in special education and unavailable for reevaluation or those students who did not require reevaluation were not included in the sample. Consequently, generalization of these results to such students is not advisable. Further investigation of the long-term stability of the WISC–III is necessary; however, the present results provide a valuable starting point.

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.

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 IQ's 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.

Goh, D. S., Teslow, C. J., & Fuller, G. B. (1981). The practice of psychological assessment among school psychologists. *Professional Psychology, 12,* 696–706.

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

Hutton, J. B., Dubes, R., & Muir, S. (1992). Assessment practices of school psychologists: Ten years later. *School Psychology Review, 21,* 271–284.

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

Juliano, J. M., Haddad, F. A., & Carroll, J. L. (1988). Three-year stability of WISC–R factor scores for Black and White, female and male children classified as learning-disabled. *Journal of School Psychology, 26,* 317–325.

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

Kiess, H. O. (1996). *Statistical concepts for the behavioral sciences* (2nd ed.). Needham Heights, MA: Allyn & Bacon.

Naglieri, J. A., & Pfeiffer, S. I. (1983). Reliability and stability of the WISC–R for children with below average IQ's. *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.

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* (updated 3rd ed., rev.). 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 Wechsler Intelligence Scale for Children 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: Psychological Corporation.

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

Wechsler, D. (1991). *Manual for the Wechsler Intelligence Scale for Children—Third Edition.* San Antonio, TX: 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 reevaluation stability of the WISC–III: A learning disabled sample.* Paper presented at the 105th Annual Convention of the American Psychological Association, Chicago, IL.

Received December 18, 1997
Revision received May 3, 1998
Accepted May 26, 1998 ■