School Psychology Review,
2001, Volume 30, No. 2, pp. 438-453

Long-term Stability of the Wechsler Intelligence Scale for Children–Third Edition among Students with Disabilities

Gary L. Canivez
Eastern Illinois University

Marley W. Watkins
The Pennsylvania State University

Abstract. Long-term stability of the Wechsler Intelligence Scale for Children-Third Edition (WISC-III) was investigated for children with specific learning disability (SLD), serious emotional disability (SED), and mental retardation (MR). Participants were 522 students from 33 states twice evaluated for special education eligibility over a mean test-retest interval of 2.87 years. There were no differential effects of disability groups on long-term stability coefficients. Stability coefficients for Full Scale IQ (FSIQ) scores were acceptable for all three disability groups. Of the global IQ and Index scores, only the Freedom from Distractability Index, Processing Speed Index, and Verbal IQ-Performance IQ discrepancy score stability coefficients were inadequate. Subtest stability was also inadequate. Mean changes from first testing to second testing for IQ and Index scores were not significant and the two significant subtest changes were not clinically meaningful due to small effect sizes. Individual change scores revealed that only the FSIQ was sufficiently stable for use with individual students with SLD, SED, or MR. Results extended those of Canivez and Watkins (1998, 1999) supporting long-term stability for the WISC-III.

Intelligence is a psychological construct presumed to be stable over time; thus, intelligence tests must produce similar scores from one time to another (Moffitt, Caspi, Harkness, & Silva, 1993). Correlation coefficients obtained in studies investigating temporal change are appropriately referred to as *stability coefficients* (Jensen, 1980); however, they only indicate the rank order of scores at different times. McDermott (1988) emphasized the need to examine mean changes to supplement correlational analyses in order to investigate *level* as well as *pattern* (rank order) of relationships. Additionally, individual score changes from one testing session to another also have been utilized by researchers as another indicator of stability (Canivez & Watkins, 1998, 1999; Elliott et al., 1985; Stavrou, 1990).

The Wechsler Scales are the most frequently used measures of cognitive abilities

This research was supported, in part, by a 1995-1996 Eastern Illinois University Faculty Development Grant and a 1995-1996 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 regarding this article should be addressed to Gary L. Canivez, Ph.D., 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.

among clinical and school psychologists (Goh, Teslow, & Fuller, 1981; Hutton, Dubes, & Muir, 1992; 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 typically been conducted with nondisabled youths across retest intervals of fewer than 6 months. Test-retest correlations for the Verbal IQ (VIQ), Performance IQ (PIQ), and Full Scale IQ (FSIQ) scores were generally in the .80s and .90s (Covin, 1977; Irwin, 1966; Quereshi, 1968; Throne, Schulman, & Kasper, 1962; Tuma & Appelbaum, 1980; Wechsler, 1974). Short-term stability studies usually have indicated significant increases in VIQ, PIQ, and FSIQ scores at retest, with the largest increases in PIQ. Exceptions to these general findings include Throne et al.'s (1962) finding that students with mental retardation showed no improvement in VIQ, PIQ, or FSIQ, and Covin's (1977) data indicating that students with learning difficulties showed improvement only in PIQ. WISC and WISC-R subtests were generally less stable than global IQs in most studies.

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 also been extensively examined, with evidence of moderate to high stability coefficients (*r*s generally ranging from the .50s to .90s). Practice effects were usually not observed when the retest interval exceeded 1 year. When practice effects were observed in long-term stability studies, the effect sizes were quite small and of no practical consequence. Most long-term stability studies have utilized students with disabilities (mostly students with specific learning disability and mental retardation) as participants due to the availability of data from special education triennial reevaluations.

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, and ranged from a low of .71 (FDI for ages 6–7) to a high of .95 (FSIQ for ages 14–15). Consistent with previous Wechsler scales, stability coefficients for the subtests were generally lower, ranging from .54 (Mazes for ages 14-15) to .93 (Vocabulary for ages 14-15). Significant increases in VIQ, PIQ, and FSIQ scores were again noted and attributed to practice effects or reduced novelty due to the short time interval (Kaufman, 1994; Sattler, 1992). The largest score gains were observed for the PIQ, consistent with findings from short-term stability studies of the WISC and WISC-R.

Long-term stability of the WISC-III has received attention only recently. Stavrou and Flanagan (1996) investigated the 3-year stability of the WISC-III among students with specific learning disabilities ($n = 50$) and found significant stability coefficients for VIQ ($r = .76$), PIQ ($r = .71$), and FSIQ ($r = .82$) scores. Mean VIQ, PIQ, and FSIQ test-retest differences were not significant. Finkelson and Stavrou (1999) found *r*s of .84 (VIQ), .87 (PIQ), and .88 (FSIQ), with no significant mean changes across time among 80 students with specific learning disabilities twice tested across a 3-year time span. Zhu, Woodell, and Kreiman (1997) also examined the long-term stability of the WISC-III among students with specific learning disabilities ($n = 60$) using retest intervals from 32-48 months. Stability coefficients for the VIQ ($r = .79$), PIQ ($r = .70$), and FSIQ ($r = .78$) were all significant. Zhu et

al. found significant decreases in VIQ, PIQ, and FSIQ scores across the retest interval. Smith, Smith, Bramlett, and Hicks (1999) also observed a significant decrease across time on VIQ scores but not for PIQ or FSIQ scores among 54 rural students with specific learning disabilities. Correlations for VIQ, PIQ, and FSIQ scores were .83, .78, and .87, respectively.

Using the WISC-III with students diagnosed with mental retardation, Bolen (1998) found significant stability coefficients of .68 (VIQ), .62 (PIQ), and .73 (FSIQ) over a 3-year retest interval. After correcting for restricted range at first testing, stability coefficients increased to .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 moderate effect strength. As expected, stability coefficients for subtests were generally lower than for the IQ scores.

Canivez and Watkins (1998) also studied the long-term stability of the WISC-III in the largest sample to date (n = 667). They reported high stability coefficients for VIQ, PIQ, FSIQ, Verbal Comprehension Index (VCI), and Perceptual Organization Index (POI scores) (rs = .87, .87, .91, .85, and .85, respectively) for youths who were predominately disabled. Stability coefficients for Freedom from Distractibility Index (FDI), Processing Speed Index (PSI), and VIQ-PIQ discrepancy scores were lower, as were stability coefficients for most of the WISC-III subtests. Mean changes from first to second testing were either not significant or the effect strength was very low and of little practical consequence. Canivez and Watkins (1999) also found few differential effects of WISC-III stability on the basis of gender, race/ethnicity, and age, and concluded that the FSIQ demonstrated adequate stability for use with individual students. Cassidy (1997) also found that WISC-III VIQ, PIQ, and FSIQ scores remained stable over a 3-year interval for a sample of children with disabilities.

Although there is general support for the stability of the WISC series with exceptional students, differential stability of WISC, WISC-R, and WISC-III scores among disability subgroups has not yet been adequately investigated. Kaufman (1990, 1994) pointed out in his comments on temporal stability of Wechsler Scales (WISC-R, WISC-III, WAIS-R) that research (primarily short-term stability) indicates substantial gains in PIQ due to practice effects and progressive error even when tests are administered years apart. Rubin, Goldman, and Rosenfeld (1985, 1990) indicated that individuals with moderate mental retardation showed greater than expected WISC-R to WAIS-R IQ gains than did individuals with mild mental retardation. Rubin et al. argued that these changes in IQ scores had implications for classification, educational programming, and funding. Differential stability *within* the WISC-III could also have similar implications for classification, placement, and funding when disabled students are reevaluated with the WISC-III. As IQ scores are used in the classification and eligibility decisions for students, particularly those with SLD and MR, differential IQ changes as a result of test instability might affect some groups but not others.

Substantial changes in special education eligibility and placement following triennial reevaluations have been observed (Clarizio & Halgren, 1991; Halgren & Clarizio, 1993). Specifically, students with speech and language impairment (SLI), specific learning disability (SLD), and serious emotional disability (SED) were most likely to be terminated or reclassified. An important factor in reclassification was IQ, in that those with lower IQs were more likely to be reclassified whereas those with higher IQs were more likely to be terminated from special education. Thus, changes in IQ as a result of test instability may differentially affect individuals with different disabilities.

Although several investigations of WISC-R stability have utilized various disability groups (Elliott & Boeve, 1987; Elliott et al., 1985; Stavrou, 1990; Vance et al., 1981; Vance et al., 1987), none directly examined differences in stability coefficients *between* the disability groups. However, analysis of the Stavrou (1990) stability coefficients indicates that there were no differences in stability estimates between students with SLD and students with MR for the FSIQ, but that VIQ stability coefficients were higher among the SLD group.

Furthermore, Public Law 105-17, The Individuals with Disabilities Education Act

Amendments of 1997, does not *require* additional testing during reevaluations. In reevaluations, the only requirement is that existing data be reviewed and the need for additional data be determined. If no additional data are needed, the child's disability status may be continued and the parents notified. If it is determined that additional data are needed, then such data need to be gathered. However, can school psychologists assume that IQ scores obtained at one point in time with a particular instrument will remain the same in the future? How can school psychologists determine if existing cognitive performance data are adequate? If IQ remains stable for some disability groups but not for others, then it may be necessary to reassess the intellectual status during reevaluations for disability groups not showing acceptable IQ score stability.

To date, there have been no investigations of the differential stability of the WISC-III for students comprising different disability groups. The purpose of the present study was to further examine the long-term stability of the WISC-III IQ, Index, VIQ-PIQ discrepancy, and subtest scores within *and* between the largest disability subgroups (specific learning disability, serious emotional disability, and mental retardation) obtained from a large, heterogeneous sample of predominately disabled children (Canivez & Watkins, 1998). Specific research questions were:

1. What is the WISC-III stability for individual groups of students with SLD, SED, and MR?

2. Is the long-term WISC-III stability for individual groups of students with SLD, SED, and MR similar to the short-term stability estimates found in the WISC-III manual?

3. Do groups of students with SLD, SED, and MR show significant differences between long-term WISC-III stability coefficients?

Method

Participants

Participants in the present study were a subset of the total sample ($n = 667$) employed in a long-term WISC-III stability study (Canivez & Watkins, 1998). Students included in the present study ($n = 522$) were independently classified with specific learning disability (SLD, $n = 409$), serious emotional disability (SED, $n = 66$), or mental retardation (MR, $n = 47$) by multidisciplinary evaluation teams consistent with state and federal guidelines governing special education classification. Cases were categorized according to the special education classification reported during their first WISC-III administration (Time 1).

Demographic information for the current sample is presented in Table 1. Males were disproportionately represented in both the SLD and SED groups. The mean retest interval for the SLD group was 2.87 years ($SD = .39$) with a range from .70–4.00 years. The mean retest interval among the SED group was 2.81 years ($SD = .45$) with a range of 2.00-4.00 years. Finally, the mean retest interval for the MR group was 2.90 years ($SD = .49$) with a range from 1.00-4.00 years.

Changes in disability classification similar to those reported by Clarizio and Halgren (1991) and Halgren and Clarizio (1993) were observed in the present study. Of the 409 students with SLD at the first testing, 20 (5.4%) were reclassified not disabled, 8 (2.2%) were reclassified as SED, 3 (.8%) were reclassified MR, 3 (.8%) were reclassified as SLI, and 329 (89.2%) were again classified as SLD at the second testing. Of the 47 students classified as SED at the first testing, 4 (8.7%) were reclassified not disabled, 7 (15.2%) were reclassified as SLD, and 33 (71.7%) were again classified as SED at the second testing. Finally, of the 66 students with MR at the first testing, 2 (3.1%) were reclassified not disabled, 5 (7.8%) were reclassified SLD, and 57 (89.1%) were again classified as MR at the second testing. Forty students with SLD, 1 student with SED, and 2 students with MR had missing data at the second testing, and 6 (1.6%) students with SLD and 2 (4.3%) students with SED were reclassified with some "other" low incidence disability such as traumatic brain injury, autism, and other health impairment.

Instrument

The Wechsler Intelligence Scale for Children-Third Edition (Wechsler, 1991) is an individually administered test of intelligence

Table 1
Demographic and Sample Characteristics at First and Second Testing

	First Testing		Second Testing	
Variable	*n*	%	*n*	%
Gender				
SLD				
Male	297	72.6	249	73.0
Female	112	27.4	92	27.0
SED				
Male	34	72.3	29	70.7
Female	13	27.7	12	29.3
MR				
Male	38	57.6	36	60.0
Female	28	42.4	24	40.0
Race/Ethnicity				
SLD				
Caucasian	319	78.0	268	78.6
Black/African American	49	12.0	41	12.0
Hispanic/Latino	25	6.1	19	5.6
Other	8	2.0	7	2.1
Missing	8	2.0	6	1.8
SED				
Caucasian	34	72.3	29	70.7
Black/African American	10	21.3	8	19.5
Hispanic/Latino	1	2.1	2	4.9
Other	0	0.0	0	0.0
Missing	2	4.3	2	4.9
MR				
Caucasian	40	60.6	38	63.3
Black/African American	20	30.3	17	28.3
Hispanic/Latino	5	7.6	4	6.7
Other	1	1.5	1	1.7
Missing	0	0.0	0	0.0
Grade				
K	19	3.6	-	-
1	90	17.2	-	-
2	120	23.0	6	1.1
3	77	14.8	28	5.4
4	65	12.5	85	16.3
5	72	13.8	119	22.8
6	39	7.5	74	14.2
7	25	4.8	62	11.9
8	11	2.1	66	12.6
9	2	0.4	46	8.8
10	-	-	22	4.2
11	-	-	8	1.5
Missing	2	0.4	6	0.1

(Table continues)

(Table 1 continued)

Variable	First Testing		Second Testing	
	M	*SD*	*M*	*SD*
Age				
SLD	9.08	1.95	11.93	1.97
SED	10.26	2.32	13.07	2.21
MR	10.01	2.41	12.84	2.54

Note. SLD = Specific Learning Disability; SED = Serious Emotional Disability; MR = Mental Retardation. Percents may not sum to 100 due to rounding.

for children ages 6 years through 16 years, 11 months. The WISC-III has 13 subtests that measure different aspects of intelligence and yield three composite IQs (viz., Verbal [VIQ], Performance [PIQ], and Full Scale [FSIQ]), which provide estimates of the individual's verbal, perceptual/nonverbal, and general intellectual abilities. Additionally, the WISC-III provides four optional factor-based index scores (viz., VCI, POI, FDI, and PSI).

The WISC-III was standardized on a representative sample ($N = 2,200$) closely approximating the 1988 United States Census on gender, parent education, race/ethnicity, and geographic region. Extensive evidence of reliability (internal consistency and short-term stability) and validity (criterion related and construct) is presented in the WISC-III manual (Wechsler, 1991).

Procedure

A random sample of 2,000 school psychologists drawn from the membership of the National Association of School Psychologists (NASP) was invited to participate in this study by providing test scores and demographic data extracted from their recent special education reevaluations. School psychologists were asked to report test scores and demographic information for students who were recently administered the WISC-III during a special education triennial reevaluation only if the student was also administered the WISC-III during an earlier evaluation. There was no specification of how many cases to report and additional selection criteria (i.e., disability, gender, age) were not imposed.

Data were received from 145 school psychologists from 33 states. Although a 7.25% return rate is low for survey research, this study was not intended to sample opinions or perspectives of the participating school psychologists, and the purpose of sampling 2,000 school psychologists was to produce as large a sample of students as possible. The school psychologists who participated provided an average of 4.6 cases each, with a range of 1 to 25 cases. The 33 states were grouped by geographic region specified in the WISC-III manual (Wechsler, 1991) to examine the distribution of cases produced within each region. Of the 522 cases selected in the present study, 125 (23.9%) were from the West, 157 (30.1%) were from the North Central, 45 (8.6%) were from the North East, and 195 (37.4%) were from the South. Although somewhat underrepresentative of the North East, this distribution is reasonably close to the percentages of students obtained in the WISC-III standardization sample (20.0%, 26.3%, 17.9%, and 35.8%, respectively) as presented in the WISC-III manual (Wechsler, 1991).

Analyses

Within groups. For each disability subgroup (SLD, SED, and MR), Pearson product-moment correlation coefficients between first and second testing were calculated for the WISC-III IQ, VIQ-PIQ discrepancy, Index, and subtest scores[1]. Due to limited variability observed in WISC-III performance in the MR group, stability coefficients were corrected for restricted range (Guilford & Fruchter, 1978) based on the variability observed at the first testing. In addition to testing hypotheses that stability coefficients were significantly greater than zero (H_0: $r = 0$), stability coeffi-

cients were also statistically compared to the short-term stability coefficients presented in the WISC-III manual (Wechsler, 1991) with independent *z*-tests within each disability group. Short-term stability coefficients from the WISC-III manual were selected for comparison purposes as they were obtained from the largest and most representative sample of nondisabled students. Stability of VIQ-PIQ discrepancies was examined because it is a commonly calculated index (Kaufman, 1994; Sattler, 1992).

Dependent *t*-tests for differences between means were conducted to investigate performance changes across the retest interval for each disability subgroup. Due to the impact of sample size on statistical significance of the *t*-tests, effect sizes (*d*) were calculated to estimate the importance of performance changes across the retest interval (Cohen, 1988). Bonferroni correction for family-wide error rates was used within each disability group for all statistical tests. Individual variation in scores across the retest interval was examined by summarizing percentages of individuals with changes within standard error of measurement groupings.

Between groups. Stability coefficients were compared *between* the three disability subgroups using independent *z*-tests for differences between correlation coefficients using Fisher *z* transformations (Guilford & Fruchter, 1978).

Results

Within Disability Group Analyses

Specific learning disability. Long-term stability coefficients, descriptive statistics, *t*-tests, and retest interval effect sizes (*d*) for the WISC-III IQ scores, VIQ-PIQ discrepancy scores, Index, and subtest scores for students with SLD are presented in Table 2. All long-term stability coefficients for IQ, VIQ-PIQ discrepancies, Index, and subtest scores were significantly different from zero ($p < .05$) with Bonferroni correction ($\alpha = .0025$).

Bonferroni correction for the independent *z*-tests was applied to control for the family-wide error rate and produced an adjusted $\alpha = .0026$ for the long-term versus short-term stability coefficient comparisons. Long-term stability coefficients within the SLD group were significantly lower than short-term stability coefficients for the VIQ, FSIQ, VCI, FDI, and PSI composite scores and all subtest scores except Picture Arrangement, Block Design, and Object Assembly.

Bonferroni correction for dependent *t*-tests for differences between means from first testing to second testing produced an adjusted $\alpha = .0025$. No significant changes were observed for IQ, Index, or VIQ-PIQ discrepancy scores. At the subtest level, Coding and Vocabulary showed significant decreases from Time 1 to Time 2 but effect sizes were small ($d = .22$ and .18, respectively).

Serious emotional disability. Stability coefficients, descriptive statistics, *t*-tests, and retest interval effect sizes (*d*) for the WISC-III IQ, VIQ-PIQ discrepancy, Index, and subtest scores for students with SED are presented in Table 2. All long-term stability coefficients for IQ, VIQ-PIQ discrepancies, Index, and subtest scores were significantly different from zero ($p < .05$) with Bonferroni correction ($\alpha = .0028$).

Bonferroni correction for the independent *z*-tests was applied to control for the family-wide error rate and produced an adjusted $\alpha = .0029$ for the long-term versus short-term stability coefficient comparisons. Long-term stability coefficients for the SED group were significantly lower than short-term stability coefficients (Wechsler, 1991) for the FSIQ and VCI composites and for the Picture Completion and Digit Span subtests.

Bonferroni correction for dependent *t*-tests for differences between means from first testing to second testing produced an adjusted $\alpha = .0026$. No significant changes were observed across the retest interval for IQ, VIQ-PIQ discrepancy, Index, or subtest scores.

Mental retardation. Stability coefficients, descriptive statistics, *t*-tests, and retest interval effect sizes (*d*) for the WISC-III IQ, VIQ-PIQ discrepancy, and Index scores for students with MR are also presented in Table 2. All long-term stability coefficients for IQ, VIQ-PIQ discrepancies, Index, and subtest scores

Table 2
Stability Coefficients, Descriptive Statistics, t-tests, and Retest Interval Effect Sizes for Students with Specific Learning Disability, Serious Emotional Disability, and Mental Retardation

			First Testing		Second Testing			
	n	*r*	*M*	*SD*	*M*	*SD*	*t*	*d*
IQ Scores								
VIQ								
SLD	406	.82*	92.67	13.31	91.54	13.70	2.77	.08
SED	47	.86	92.55	14.00	93.83	13.97	1.17	.09
MR[a]	66	.85*	65.89	9.11	64.89	10.92	1.02	.10
PIQ								
SLD	406	.82	95.85	14.21	95.44	15.48	0.92	.03
SED	47	.81	91.45	16.77	93.49	15.79	1.39	.13
MR[a]	66	.90	65.77	11.20	65.11	11.75	0.83	.06
FSIQ								
SLD	403	.87*	93.54	12.92	92.69	14.34	2.40	.06
SED	47	.88	91.30	15.61	92.94	14.86	1.51	.11
MR[a]	66	.93	63.00	9.88	62.03	11.56	1.35	.09
VIQ-PIQ								
SLD	406	.64	-3.18	13.86	-3.90	12.69	1.28	.05
SED	47	.56	1.11	10.91	0.34	12.05	0.49	.07
MR[a]	66	.60	0.12	10.02	-0.21	8.20	0.28	.04
Index Scores								
VCI								
SLD	387	.81*	94.14	13.64	93.28	13.88	1.98	.06
ED	43	.82*	92.98	13.58	95.28	14.00	1.81	.17
MR[a]	58	.84*	67.93	8.96	66.83	10.74	1.04	.11
POI								
SLD	380	.81	96.89	14.07	97.52	15.61	1.33	.04
SED	42	.80	93.05	17.06	95.64	16.88	1.57	.15
MR[a]	57	.87	65.61	11.63	65.39	12.35	0.23	.02
FDI								
SLD	295	.66*	88.42	12.41	87.57	11.31	1.50	.07
SED	33	.75	88.64	16.38	89.58	16.13	0.47	.06
MR[a]	40	.81	65.78	10.90	68.23	12.40	1.74	.21
PSI								
SLD	118	.58*	95.99	15.58	93.97	13.49	1.64	.14
SED	8	-	87.38	10.58	90.00	8.60	0.81	.27
MR[a]	16	-	74.81	16.40	78.81	19.03	1.07	.23
Subtest Scores								
PC								
SLD	389	.58*	9.41	3.02	9.70	2.88	2.16	.10
SED	44	.47*	9.27	3.11	10.30	3.05	2.15	.33
MR[a]	59	.59*	4.44	2.69	4.41	2.93	0.10	.01

(Table 2 continues)

(Table 2 continued)

				First Testing		Second Testing			
		n	*r*	*M*	*SD*	*M*	*SD*	*t*	*d*
I									
	SLD	389	.66*	8.30	2.91	8.50	2.94	1.86	.07
	SED	44	.71	8.07	3.17	9.14	2.82	3.09	.36
	MR[a]	59	.69	4.05	2.02	3.86	2.08	0.73	.09
CD									
	SLD	385	.56*	8.85	3.12	8.17	3.05	4.63*	.22
	SED	44	.70	8.05	3.51	7.57	3.50	1.16	.14
	MR[a]	58	.61	5.28	3.33	4.78	3.23	1.38	.15
S									
	SLD	390	.58*	8.94	3.05	8.97	2.88	0.21	.01
	SED	44	.71	8.98	3.46	9.80	3.15	2.12	.25
	MR[a]	60	.48*	3.95	2.30	4.03	2.44	0.25	.03
PA									
	SLD	389	.58	9.27	3.23	9.31	3.65	0.24	.01
	SED	44	.68	9.23	3.33	9.57	3.48	0.82	.10
	MR[a]	60	.65	3.75	2.44	3.78	2.59	0.11	.01
A									
	SLD	389	.56*	7.86	2.67	7.60	2.65	2.07	.10
	SED	44	.72	8.14	3.60	7.98	3.37	0.40	.05
	MR[a]	59	.60	3.00	1.99	3.47	2.13	1.68	.23
BD									
	SLD	389	.73	9.26	3.32	9.28	3.65	0.16	.01
	SED	44	.73	8.80	3.89	8.39	4.03	0.93	.10
	MR[a]	59	.74	3.51	2.42	2.97	2.48	2.08	.22
V									
	SLD	388	.70*	8.62	2.81	8.11	2.85	4.62*	.18
	SED	44	.79	8.89	3.37	8.70	3.00	0.58	.06
	MR[a]	59	.57*	3.98	2.04	3.44	2.16	1.85	.26
OA									
	SLD	376	.60	9.29	2.97	9.39	3.26	0.70	.03
	SED	42	.62	7.79	3.71	8.71	3.45	1.93	.26
	MR[a]	58	.59	4.50	3.16	4.52	2.80	0.05	.01
C									
	SLD	384	.59*	9.49	3.28	9.17	3.18	2.19	.10
	SED	43	.50	8.86	3.47	8.74	3.03	0.23	.04
	MR[a]	58	.66	4.17	2.32	4.07	2.43	0.35	.04
SS									
	SLD	117	.54*	9.28	3.72	9.52	3.16	0.77	.07
	SED	8	-	8.38	2.50	8.50	2.67	0.15	.05
	MR[a]	16	-	4.25	2.82	6.75	4.17	2.60	.72
DS									
	SLD	290	.53*	7.72	2.52	7.65	2.32	0.52	.03
	SED	32	.60*	7.72	2.74	7.97	3.14	0.53	.09
	MR[a]	41	.82	4.71	2.36	4.95	2.69	0.86	.10

(Table 2 continues)

(Table 2 continued)

Note. SLD = Specific Learning Disability; SED = Serious Emotional Disability; MR = Mental Retardation; VIQ = Verbal IQ; PIQ = Performance IQ; FSIQ = Full Scale IQ; VIQ-PIQ = Verbal IQ-Performance IQ discrepancy; VCI = Verbal Comprehension Index; POI = Perceptual Organization Index; FDI = Freedom from Distractibility Index; PSI = Processing Speed Index; 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. All correlations significant (H_0: $r = 0$) $p < .05$ (with Bonferroni correction for family wide error rates within each disability group). Correlations with asterisks indicate significant difference from WISC-III short-term stability coefficients obtained in the WISC-III manual (Wechsler, 1991). Correlations not presented when the sample size was less than 30.

[a] Correlations for the MR group were corrected for restricted variability of WISC-III scores observed at the first testing (Guilford & Fruchter, 1978).

*$p < .05$ (with Bonferroni correction for family-wide error rates).

were significantly different from zero ($p < .05$) with Bonferroni correction ($\alpha = .0028$).

Bonferroni correction for the independent *z*-tests was applied to control for the family-wide error rate and produced an adjusted $\alpha = .0029$ for the long-term versus short-term stability coefficient comparisons. Long-term stability coefficients for students with MR were significantly lower than short-term stability coefficients (Wechsler, 1991) for the VIQ and VCI composites and the Picture Completion, Similarities, and Vocabulary subtests.

Bonferroni correction for dependent *t*-tests for differences between means from first testing to second testing produced an adjusted $\alpha = .0025$. No significant changes across the retest interval were observed for IQ, VIQ-PIQ discrepancy, Index, or subtest scores.

Individual changes. Changes in IQ and Index scores across the retest interval are presented within standard error of measurement ranges in Table 3 for the three disability groups. These results indicated that 65.1% of the SLD, 51.2% of the SED, and 72.5% of the MR groups in this study had FSIQ changes within ±2 standard errors of measurement. Of particular interest is that 15% of the SLD, 19.2% of the SED, and 10.5% of the MR groups had FSIQ changes exceeding ±3 standard errors of measurement. Similar percentages were observed for the VIQ, PIQ, and Index scores.

Between Disabilities Analyses

Stability coefficients for IQ, VIQ-PIQ discrepancy, Index, and subtest scores for SLD, MR, and SED groups were compared to determine if significant differences existed *between* the three disability groups. Bonferroni correction was applied to control the family wide error rate and produced an adjusted $\alpha = .0009$. Independent *z*-tests for differences between correlation coefficients using Fisher *z* transformations (Guilford & Fruchter, 1978) produced no significant differences between disability groups' stability coefficients for IQ, VIQ-PIQ discrepancy, Index, or subtest scores.

Discussion

When compared to previous studies with small, local samples of students with SLD (Finkelson & Stavrou, 1999; Smith et al., 1999; Stavrou & Flanagan, 1996; Zhu et al., 1997), the present results produced equivalent WISC-III stability coefficients (VIQ, PIQ, and FSIQ coefficients of .81, .80, and .86, respectively, for the present SLD sample versus mean coefficients of .81, .78, and .84 from previous samples). There were no significant changes in mean IQ, Index, and VIQ-PIQ discrepancy scores across the retest interval. Subtest stability coefficients were consistently lower than the global IQ and Index scores and the two subtests showing significant changes across the retest interval demonstrated small effect sizes that were clinically unimportant.

WISC-III stability for the students with SED in the present study was evidenced by high and significant test-retest correlations for the VIQ, PIQ, FSIQ, VCI, and POI scores. There were no significant changes in global IQ or Index scores across the retest interval. There are no WISC-III long-term stability studies using students with SED as a sample with

Table 3
Percent of Students Showing WISC-III IQ and Index Score Changes Within Standard Error of Measurement Ranges and Descriptive Statistics for Score Changes by Disability Group

	Standard Error of Measurement Range								
	<-3	-3 to -2	-2 to -1	-1 to +1	+1 to +2	+2 to +3	>+3	*M*	*SD*
VIQ									
SLD	11.2	9.6	16.8	38.2	10.4	4.6	7.1	-1.13	8.23
SED	8.5	4.2	14.9	34.0	14.9	15.0	8.4	1.28	7.50
MR	7.5	12.1	22.7	30.3	21.1	7.5	4.5	-1.00	7.94
PIQ									
SLD	5.1	11.8	16.2	37.4	15.3	7.4	6.2	-0.41	8.99
SED	8.6	4.3	8.6	46.8	6.3	12.8	12.7	2.04	10.06
MR	1.5	7.5	16.6	52.9	15.0	4.5	1.5	-0.67	6.49
FSIQ									
SLD	10.9	9.6	14.6	36.1	14.4	9.7	4.1	-0.84	7.04
SED	6.4	12.7	10.6	27.8	12.8	17.0	12.8	1.64	7.44
MR	6.0	13.6	13.6	46.9	12.0	3.0	4.5	-0.97	5.82
VCI									
SLD	9.9	11.1	15.5	31.9	14.7	12.1	5.4	-0.87	8.59
SED	4.6	7.0	11.7	37.1	16.4	9.3	13.9	2.30	8.33
MR	5.1	15.5	25.9	27.4	12.1	5.1	8.6	-1.10	8.06
POI									
SLD	6.3	7.5	16.9	33.4	19.2	10.8	6.2	0.63	9.25
SED	9.6	2.4	11.9	38.1	12.0	16.7	9.6	2.60	10.73
MR	5.4	5.3	17.7	43.9	21.1	5.3	1.8	-0.23	7.54
FDI									
SLD	3.9	11.5	18.2	39.6	14.3	7.1	4.9	-0.85	9.81
SED	6.0	6.0	12.2	42.5	18.2	9.1	6.0	0.94	11.41
MR	2.5	2.5	12.5	50.0	17.5	7.5	7.5	2.45	8.92
PSI									
SLD	9.7	10.9	15.2	37.3	12.7	7.5	5.7	-2.03	13.40
SED	0.0	0.0	25.0	50.0	0.0	25.0	0.0	2.63	9.12
MR	12.6	0.0	12.5	25.2	25.1	6.3	18.9	4.00	14.95

Note. Change in scores was determined by subtracting the initial obtained score from the most recent score. SLD = Specific Learning Disability; SED = Serious Emotional Disability; MR = Mental Retardation; 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. Average Standard Errors of Measurement for the IQ and Index scores were utilized to categorize deviations and were obtained from Table 5.2 of the WISC-III manual (Wechsler, 1991, p. 168). Tables showing individual change scores across the retest interval are available from the first author.

which to compare the present results, but they are similar to those found by Haynes and Howard (1986) in their study of the long-term stability of the WISC-R among neglected students from a juvenile court sample.

Long-term stability of the WISC-III among students with MR in the present study was evidenced by somewhat higher test-retest correlations for the FSIQ and PIQ than those found by Bolen (1998). The present stability

coefficients for the VIQ, PIQ, and FSIQ were also higher than those found with the WISC-R (Webster, 1988; Whorton, 1985).

There were no differences in stability coefficients between the three disability groups at the IQ, Index, or subtest score levels. When individual changes in WISC-III scores across time were analyzed, the FSIQ was the most stable. However, 15% of SLD, 19.2% of SED, and 10.5% of MR students exhibited FSIQ score changes exceeding ±3 standard errors of measurement (i.e., ±9 points).

Practice Implications

From a nomothetic perspective, long-term stability of the WISC-III FSIQ appeared to be adequate for most individual diagnostic purposes for all three disability subgroups, as stability coefficients met the .85–.90 criterion recommended by measurement experts (Hills, 1981; Salvia & Ysseldyke, 1991). Most of the disability subgroups' stability coefficients for VIQ, PIQ, VCI, and POI scores also demonstrated satisfactory stability. However, disability subgroup stability coefficients for FDI and PSI (where calculated) and VIQ-PIQ discrepancy scores were inadequate for confident use with individuals. Long-term stability coefficients for the WISC-III subtests among disability subgroups were also generally lower than those found for the global IQ and Index scores. These results supplement and extend the previously reported conclusion of Canivez and Watkins (1998) that WISC-III subtest scores are too unstable for making decisions about individual students.

Changes in IQ scores presented in Table 3 illustrate how IQ scores varied for individuals across the retest interval. This idiographic comparison demonstrated that the WISC-III FSIQ was stable for the majority of individual students. However, 10.5–19.2% of individuals, depending on the disability subgroup, showed changes greater than ±3 standard errors of measurement (i.e., ±9 standard score points). 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 slightly greater percentages of their students showed significant VIQ, PIQ, or FSIQ changes.

Nomothetic *and* idiographic perspectives on long-term stability of the WISC-III with disabled students suggest that the FSIQ is stable across time for most students. However, a sizable minority of students with disabilities exhibited FSIQ changes greater than ±3 SE_m (i.e., 14.8% of this sample). Thus, it is not appropriate to assume that estimates of ability have remained stable across time for *all* students. Unfortunately, there is no way to identify which individual students will exhibit significantly discrepant ability estimates upon reevaluation unless the WISC-III is included in the reevaluation.

However, federal special education regulations do not mandate routine readministration of tests (i.e., WISC-III) in triennial reevaluations (National Association of School Psychologists, 1999). Regulations specify that "a reevaluation of each child, in accordance with §§300.532-300.535, [be] conducted if conditions warrant a reevaluation, or if the child's parent or teacher requests a reevaluation, but at least once every three years" (Department of Education, 1999, §300.536). For reevaluation purposes, the only requirement is that existing data be reviewed and any additional data needed to determine whether the child continues to have a disability be identified. If no additional data are needed, the school may continue the child's disability status and notify the parents. If additional data are needed, the school is obligated to obtain them. Given that 14.8% of this sample of students with disabilities exhibited FSIQ changes greater than ±3 SE_m on retesting and that 11.5% were reclassified upon reevaluation, previous WISC-III evaluation data cannot routinely be considered sufficient to determine that a child continues to have a disability.

A flexible approach to reevaluations might be to include a brief test of cognitive abilities such as the Kaufman Brief Intelligence Test (K-BIT; Kaufman & Kaufman, 1990), which has been shown to be strongly related to and indicative of performance on the WISC-III (Canivez, 1995, 1996; Prewett, 1995). Newer brief measures of intelligence such as

the Wide Range Intelligence Test (WRIT; Glutting, Adams, & Sheslow, 1999) or the Wechsler Abbreviated Scale of Intelligence (WASI; Wechsler, 1999) also may be utilized to estimate (recheck) the cognitive abilities of students during reevaluations. In their evaluation of the long-term stability of the WISC-R among children with disabilities, Elliott et al. (1985) also recommended the use of a brief measure of intelligence in reevaluation of "Anglo" students.

When significant changes in the current estimate of intellectual abilities arise (i.e., K-BIT IQ Composite and previously administered WISC-III FSIQ differ by ±10 standard score points, WASI-2 subtest FSIQ and WISC-III FSIQ differ by ±10 standard score points, or WASI-4 subtest FSIQ and WISC-III FSIQ differ by ±9 standard score points)[2], readministration of the WISC-III or use of another comprehensive measure of intelligence might then be warranted. The conservation of approximately 1 hour of assessment, scoring, and interpretation time for each reevaluation not requiring readministration of the WISC-III or another comprehensive intellectual measure could permit time to be devoted to assessment of the efficacy of the student's individual educational program (Ross-Reynolds, 1990) or other professional activities such as consultation, direct intervention, and research. This procedure would allow school psychologists to provide "flexible and meaningful approaches" to reevaluations and "assist in coordinating a review of the student's progress that considers the efficacy and appropriateness of the student's current program" (NASP, 1999).

Limitations

Of course, these conclusions and recommendations must be considered in light of several limitations to the present study. First, generalization of these results is in part limited as these data were not obtained by random selection. School psychologists (145 of 2,000) chose to participate in response to a written request. They then reported data from reevaluation cases that they personally selected. The large number of school psychologists (from 33 different states) who participated should, to some extent, reduce this threat because it is unlikely that any one type of student would be systematically or preferentially selected. Second, there was no way to validate the accuracy of WISC-III test scores. Thus, administration, scoring, or reporting errors could have influenced results. Third, results for students with SED and MR are based on small samples and require replication with larger samples. A final limitation is that the use of reevaluation cases means that those students who were no longer enrolled in special education were not reevaluated and thus not included in the sample. Generalization of these results to such students is therefore not supported.

Future Research

The stability of intellectual and disability indicators among random cohorts of students with disabilities should be investigated to better understand the utility and efficacy of special education placements. Because the results of this study for students with SED and MR are based on small samples, replication with larger samples is also critical. The present study as well as those of Canivez and Watkins (1998, 1999) have demonstrated the stability of the global IQ scores; however, other aspects of stability should also be examined. For example, Juliano, Haddad, and Carroll (1988) found the factor structure of the WISC-R to be stable across a long time interval. Is the factor structure of the WISC-III stable over a 3-year interval? The stability of cognitive profiles and ipsative interpretation methods, which are often used to make decisions about students' educational placements and interventions (Alfonso, Oakland, LaRocca, & Spanakos, 2000; Kaufman, 1994), should also be investigated. Additionally, longitudinal relationships between IQ scores and other student characteristics (Austin, Hofer, Deary, & Eber, 2000) should be examined in greater detail.

References

Alfonso, V. C., Oakland, T. D., LaRocca, R., & Spanakos, A. (2000). The course on individual cognitive assessment. *School Psychology Review, 29,* 52-64.

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.

Austin, E. J., Hofer, S. M., Deary, I. J., & Eber, H. W. (2000). Interactions between intelligence and personality: Results from two large samples. *Personality and Individual Differences, 29,* 405-427.

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. (1995). Validity of the Kaufman Brief Intelligence Test: Comparisons with the Wechsler Intelligence Scale for Children-Third Edition. *Assessment, 2,* 101-111.

Canivez, G. L. (1996). Validity and diagnostic efficiency of the Kaufman Brief Intelligence Test in reevaluating students with learning disability. *Journal of Psychoeducational Assessment, 14,* 4-19.

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

Canivez, G. L., & Watkins, M. W. (1999). Long term stability of the Wechsler Intelligence Scale for Children-Third Edition among demographic subgroups: Gender, race/ethnicity, and age. *Journal of Psychoeducational Assessment, 17,* 300-313.

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

Clarizio, H. F., & Halgren, D. W. (1991). Continuity in special education placements: Are reevaluations really necessary? *Psychology in the Schools, 28,* 317-324.

Cohen, J. (1988). *Statistical power analyses for the behavioral sciences.* Hillsdale, NJ: Lawrence 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.

Department of Education. Assistance to States for the Education of Children with Disabilities and the Early Intervention Program for Infants and Toddlers with Disabilities; Final Regulations. 34 CFR parts 300 and 303 (1999).

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., Piersel, 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.

Finkelson, L., & Stavrou, E. (1999, April). *The stability of IQ in learning disabled students.* Paper presented at the Annual Convention of the National Association of School Psychologists, Las Vegas, NV.

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.

Glutting, J., Adams, W., & Sheslow, D. (1999). *Wide Range Intelligence Test.* Wilmington, DE: Wide Range.

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

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

Halgren, D. W., & Clarizio, H. F. (1993). Categorical and programming changes in special education services. *Exceptional Children, 59,* 547-555.

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.

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

Public Law (P.L.) 105-17. Individuals with Disabilities Education Act Amendments of 1997. (20 U.S.C. 1400 et seq.).

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.

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. (1990). *Assessing adolescent and adult intelligence.* Boston: Allyn and Bacon.

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

Kaufman, A. S., & Kaufman, N. L. (1990). Kaufman Brief Intelligence Test. Circle Pines, MN: American Guidance Service.

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.

National Association of School Psychologists. (1999). *Position statement on three-year reevaluations for students with disabilities.* Bethesda, MD: Author.

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.

Prewett, P. N. (1995). A comparison of two screening tests (the Matrix Analogies Test-Short Form and the Kaufman Brief Intelligence Test) with the WISC-III. *Psychological Assessment, 7,* 69-72.

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.

Ross-Reynolds, J. (1990). Best practices in conducting reevaluations. In A. Thomas & J. Grimes (Eds.), *Best practices in school psychology-II* (pp. 195-206). Washington, DC: National Association of School Psychologists.

Rubin, H. H., Goldman, J. J., & Rosenfeld, J. G. (1985). A comparison of WISC-R and WAIS-R IQs in a mentally retarded residential population. *Psychology in the Schools. 22,* 392-397.

Rubin, H., Goldman, J. J., & Rosenfeld, J. G. (1990). A follow-up comparison of WISC-R and WAIS-R IQs in a residential mentally retarded population. *Psychology in the Schools, 27,* 309-310.

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

Sattler, J. (1992). *Assessment of children* (rev. and updated 3rd ed.). San Diego, CA: Jerome M. Sattler, Publisher.

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

Smith, T., Smith, B. L., Bramlett, R. K., & Hicks, N. (1999, April). *WISC-III stability over a three-year period in students with learning disabilities.* Paper presented at the Annual Convention of the National Association of School Psychologists, Las Vegas, NV.

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.

Wechsler, D. (1999). *Wechsler Abbreviated Scale of Intelligence.* 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.

Footnotes

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

[2]The critical value for significant ($\alpha = .05$, $z = 1.96$) differences between the K-BIT IQ Composite, WASI-2 FSIQ, and WASI-4 FSIQ, and the WISC-III FSIQ was obtained using the standard error of difference:

$$SE_{diff} = SD\sqrt{2 - r_{11} - r_{22}}$$

where r_{11} was the mean internal consistency coefficient for the K-BIT IQ Composite calculated using Fisher's z transformation for ages 6-16

from the K-BIT manual (Kaufman & Kaufman, 1990), mean internal consistency coefficient for the *WASI-2* FSIQ or WASI-4 FSIQ for ages 6-16 (Wechsler, 1999); and r_{22} was the mean internal consistency coefficient for the WISC-III FSIQ obtained from the WISC-III manual (Wechsler, 1991).

Gary L. Canivez, Ph.D., is Associate Professor of Psychology at Eastern Illinois University principally involved in the school psychologist training program. His research interests include psychometric investigations of measures of intelligence, achievement, and psychopathology.

Marley W. Watkins, Ph.D., received his doctorate in school psychology from the University of Nebraska-Lincoln and is currently Professor-in-Charge of Graduate Programs in School Psychology at The Pennsylvania State University. He is interested in encouraging scientific school psychological practice.