Psychological Assessment
2013, Vol. 25, No. 2, 477–483

1040-3590/13/$12.00 DOI: 10.1037/a0031653

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

Marley W. Watkins
Baylor University

Lourdes G. Smith
Arizona State University

Long-term stability of the Wechsler Intelligence Scale for Children—Fourth Edition (WISC–IV; Wechsler, 2003) was investigated with a sample of 344 students from 2 school districts twice evaluated for special education eligibility at an average interval of 2.84 years. Test–retest reliability coefficients for the Verbal Comprehension Index (VCI), Perceptual Reasoning Index (PRI), Working Memory Index (WMI), Processing Speed Index (PSI), and the Full Scale IQ (FSIQ) were .72, .76, .66, .65, and .82, respectively. As predicted, the test–retest reliability coefficients for the subtests (*Mdn* = .56) were generally lower than the index scores (*Mdn* = .69) and the FSIQ (.82). On average, subtest scores did not differ by more than 1 point, and index scores did not differ by more than 2 points across the test–retest interval. However, 25% of the students earned FSIQ scores that differed by 10 or more points, and 29%, 39%, 37%, and 44% of the students earned VCI, PRI, WMI, and PSI scores, respectively, that varied by 10 or more points. Given this variability, it cannot be assumed that WISC–IV scores will be consistent across long test–retest intervals for individual students.

Keywords: WISC–IV, long-term stability, special education, reliability

Reliability, or consistency, of scores is vital for tests of intelligence because IQ scores are often used for diagnostic and intervention purposes. Given the importance of reliability, consistency across items (or internal consistency reliability) has been routinely investigated for most intelligence tests. Investigations have found that the internal consistency reliability of omnibus intelligence test scores for standardization samples tends to exceed .90, and internal consistency reliability of subtest scores from those samples tends to exceed .80. Although not frequently investigated, internal consistency reliability estimates of intelligence test scores for clinical or referral samples have generally been found to be equivalent to those found in standardization samples (Krouse & Braden, 2011; Ryan, Glass, & Bartels, 2009; Silverstein, 1969; Zhu, Tulsky, Price, & Chen, 2001).

Internal consistency reliability is a vital foundation for the validity of test scores and their subsequent interpretation (Reynolds & Miliam, 2012), but consistency of test scores across time is also important (Schmidt, Le, & Ilies, 2003). The reliability of intelligence test scores across time is typically quantified by administering the same test to the same individuals twice, then correlating the test and retest scores to produce a stability coefficient. Test–retest reliability of intelligence tests has not been investigated with regularity. However, stability coefficients for some modern intelligence tests across short-term test–retest intervals of a few days or weeks have been reported (e.g., Wechsler, 2003).

Although short-term stability of intelligence test scores is critical, long-term stability is also consequential because high-stakes decisions based on intelligence test scores result in long-term placements, thereby necessitating longitudinal predictive validity. For example, children who are placed in special education programs may periodically be recertified as eligible for services, but they are not generally readministered an intelligence test (Madaus & Shaw, 2006). If intelligence test scores are unstable across 2- to 3-year time spans, then these children may be retained in special education programs when they no longer are eligible for those services. Thus, it is essential that initial placement decisions are based on stable intelligence test scores because special education placement is not uniformly helpful for students and might be harmful to some (Reschly & Bergstrom, 2009).

Importantly, the longitudinal stability of intelligence test scores assumes that the construct measured by those scores (i.e., intelligence) is stable across time. Fortunately, intelligence is assumed to be a stable trait (Hunt, 2010; Mackintosh, 1998; Reeve & Bonaccio, 2011; Revelle, 2010; Simonton, 2011; Wright, 2011), and intelligence test scores have been found to be relatively stable from childhood through adulthood (Chen & Siegler, 2000; Johnson, Gow, Corley, Starr, & Deary, 2010) for both average and above-average samples (Reeve & Bonaccio, 2011; Simonton, 2011).

Given the relative stability of intelligence over time, strong test–retest reliability should be evidenced by individual tests of intelligence (Wright, 2011). In clinical practice, the Wechsler scales are the most frequently used individual intelligence tests with children and adolescents (Kamphaus, Petoskey, & Rowe, 2000). The internal consistency reliability of early Wechsler child scales (Wechsler Intelligence Scale for Children; WISC) were found to be good to excellent in a variety of child populations

This article was published Online First February 11, 2013.

Marley W. Watkins, Department of Educational Psychology, Baylor University; Lourdes G. Smith, Mary Lou Fulton Teachers College, Arizona State University.

Correspondence concerning this article should be addressed to Marley W. Watkins, Department of Educational Psychology, Baylor University, Waco, TX 76798-7301. E-mail: marley_watkins@baylor.edu

(Mishra & Lord, 1982; Quereshi, 1968; Wechsler, 1974). The stability of WISC scores has been investigated across a variety of test–retest intervals with healthy children, gifted children, children with learning disabilities, and children diagnosed with mental retardation (Anderson, Cronin, & Kazmierski, 1989; Bauman, 1991; Canivez & Watkins, 1998, 1999, 2001; Ellzey & Karnes, 1990; Naglieri & Pfeiffer, 1983). Stability coefficients in the .80–.90 range have been found in short-term test–retest investigations of WISC IQ scores (Tuma & Applebaum, 1980; Wechsler, 1974, 1991). Long-term stability coefficients were not as consistently high (Schuerger & Witt, 1989). When longer retest intervals (e.g., 3 years or more) of WISC IQ score stability were examined, coefficients had more variability, with *r*s ranging from the .50s to .90s (Bauman, 1991; Canivez & Watkins, 1998; Oakman & Wilson, 1988; Smith, 1978; Stavrou, 1990). Additionally, long-term stability coefficients for subtest scores have generally been weaker than omnibus IQ scores, ranging from .55 to .78.

The Wechsler Intelligence Scale for Children—Fourth Edition (WISC–IV; Wechsler, 2003) is the current Wechsler scale used in clinical practice with children. Given that around 60% of the items in its core subtests are new or revised (Watkins, 2010), internal consistency reliability and test–retest reliability of the WISC–IV cannot be assumed to be equivalent to previous versions and must be reestablished for competent and ethical practice (Adams, 2000; Strauss, Spreen, & Hunter, 2000). As with prior WISC versions, evidence regarding internal consistency reliability of WISC–IV scores has been positive. Strong internal consistency reliability coefficients have also been reported for the WISC–IV among primary school students (Ryan et al., 2009) and for students with hearing impairments (Krouse & Braden, 2011). In summary, "the internal reliability of the WISC-IV is excellent" (Strauss, Sherman, & Spreen, 2006, p. 331).

The stability of WISC–IV scores across time has been investigated in only three studies. The first examination of the short-term stability of the WISC–IV involved retesting 18–27 children from each of the 11 age groups in the standardization sample after an interval of 13–63 days (N = 243; Wechsler, 2003). The stability coefficients of the omnibus IQ scores ranged from .86 for the Processing Speed Index (PSI) to .93 for the Verbal Comprehension Index (VCI) and Full Scale IQ (FSIQ). WISC–IV subtest scores were less stable, with 43 of 75 stability coefficients for the 15 subtests below .80 (see Table 1). Although the results of this study provided valuable information, they may not generalize to clinical populations or longer test–retest intervals.

In the second study the test–retest stability of WISC–IV scores across a medium term of 11 months was evaluated in a sample of 43 elementary and middle-school children and revealed subtest stability coefficients that were consistently smaller than the short-term stability coefficients reported for the normative sample (Ryan, Glass, & Bartels, 2010). Specifically, subtest stability coefficients ranged from .26 for Picture Concepts to .84 for Vocabulary. The omnibus score stability coefficients ranged from .54 for the PSI to .88 for the FSIQ. It was noted that 42% of FSIQ scores changed 5 or more points on retest, indicating that although the FSIQ had the largest stability coefficient, considerable score variation might still occur (see Table 1). Although addressing an important issue, the sample size of this study was small, thus the large confidence intervals around stability coefficients (Charter, 1999). The results also may not be generalizable because the sample was composed of primarily White students from a single private school and was not representative of ethnically diverse populations or public school populations.

The third and final investigation of the stability of WISC–IV scores involved 131 students with a learning disability enrolled in a New York suburban school district who had been tested twice with the WISC–IV approximately 3 years apart (Lander, 2010). This study was the only one in which the long-term stability of WISC–IV scores was investigated. Stability coefficients for the

Table 1
Test–Retest Coefficients and Mean Difference Scores (Retest − Test) From Three WISC–IV Stability Studies

	Test–retest interval								
	13–63 days[a]			11 months[b]			3 years[c]		
Scale	r_{12}	Δ	d	r_{12} [95% CI]	Δ	d	r_{12} [95% CI]	Δ	d
Similarities	.81	+0.6	+0.24	.63 [.41, .78]	+0.51	+0.17	.48 [.34, .60]	+0.57	+0.26
Vocabulary	.85	+0.3	+0.13	.81 [.67, .89]	−0.02	−0.01	.56 [.43, .67]	−0.49	−0.22
Comprehension	.82	+0.2	+0.08	.49 [.22, .69]	−0.23	−0.08	.55 [.42, .66]	+0.19	+0.23
Block Design	.81	+1.2	+0.41	.67 [.46, .81]	+0.56	+0.21	.62 [.50, .72]	−0.40	−0.16
Picture Concepts	.71	+0.8	+0.29	.22 [.00, .49]	+0.17	+0.07	.44 [.29, .57]	+0.38	+0.13
Matrix Reasoning	.77	+0.6	+0.23	.40 [.11, .63]	−0.53	−0.21	.48 [.34, .60]	−0.25	−0.10
Digit Span	.81	+0.5	+0.18	.74 [.56, .85]	+0.30	+0.11	.46 [.31, .59]	−0.14	−0.05
Letter-Number Sequencing	.75	+0.4	+0.16	.49 [.22, .69]	+0.31	+0.11	.31 [.15, .46]	+0.08	+0.01
Coding	.81	+1.4	+0.48	.42 [.14, .64]	+0.24	+0.08	.46 [.31, .59]	−0.90	−0.40
Symbol Search	.68	+1.1	+0.41	.41 [.13, .63]	+0.72	+0.31	.28 [.11, .43]	+0.05	+0.02
Median subtest	.81	+0.6	+0.27	.49	+0.27	+0.08	.47	−0.05	−0.03
Verbal Comprehension	.89	+2.1	+0.18	.75 [.58, .86]	+0.25	+0.02	.65 [.54, .74]	+0.64	+0.06
Perceptual Reasoning	.85	+5.2	+0.39	.58 [.34, .75]	+0.42	+0.04	.62 [.50, .72]	−0.30	−0.02
Working Memory	.85	+2.6	+0.20	.73 [.55, .85]	+2.05	+0.14	.54 [.41, .65]	−0.45	−0.04
Processing Speed	.79	+7.1	+0.51	.49 [.22, .69]	+2.32	+0.20	.52 [.38, .64]	−2.14	−0.18
Median index	.85	+5.2	+0.32	.66	+1.63	+0.10	.58	−0.45	−0.05
Full Scale IQ	.89	+5.6	+0.46	.80 [.66, .89]	+1.63	+0.15	.70 [.60, .78]	−0.55	−0.06

Note. WISC–IV = Wechsler Intelligence Scale for Children—Fourth Edition; r_{12} = uncorrected test–retest correlation; d = standardized mean difference; Δ = mean score difference; CI = confidence interval.
[a] N = 243 from Wechsler (2003). [b] N = 43 from Ryan, Glass, and Bartels (2010). [c] N = 131 from Lander (2010).

omnibus IQ scores ranged from .52 for the PSI to .65 for the VCI and .70 for the FSIQ. Subtest stability coefficients ranged from .28 for Symbol Search to .62 for Block Design (see Table 1). On average, stability coefficients were smaller than those found in short-term and medium-term investigations, but the sample size was not large enough for precise estimates of the population (Charter, 1999).

These three studies have provided valuable information regarding the stability of WISC–IV scores. Unfortunately, only one study examined the long-term stability of WISC–IV scores, and its participants were all students with learning disabilities from a single school district in New York. Consequently, these results may not generalize to other populations and situations. In the present study, we extend the investigation of the long-term stability of WISC–IV scores to include a sample of students from two different school districts in the southwest who were twice tested for special education eligibility with the WISC–IV over a 1- to 3-year time interval.

Method

Participants

The special education files of all students with English as their parent-reported primary and home language were reviewed. In total, there were 457 students who were twice administered the WISC–IV. However, only 344 students (66% male) had complete subtest and composite scores at both test and retest. These 344 students served as participants given that the mechanism for missing scores was unknown, and inclusion of missing values might bias parameter estimates (D. B. Rubin, 1976). Initial WISC–IV testing was accomplished by 124 separate examiners, and retesting was completed by 86 separate examiners. Only 66 participants were twice tested by the same examiner.

The ethnic background of participants was 79% White, 11% Hispanic, 6% Black, and 4% "other." The mean age at first testing was 8.74 years (SD = 1.57 years, range = 6.1–14.3 years), and the mean age at second testing was 11.6 years (SD = 1.69 years, range = 7.5–16.6 years) for an average test–retest interval of 2.84 years (SD = 0.75 years). Special education placement was determined by local multidisciplinary evaluation teams following state regulations. Special education diagnosis on initial evaluation included 66% learning disabled, 9% other health impairment (OHI; attention-deficit/hyperactivity disorder [ADHD]), 8% emotional disability, 5% nonhandicapped, 4% autism, 2% mental impairment, 3% OHI (non-ADHD), and 3% "other." To preserve respondents' privacy, no other information was collected on examiners or students.

School district demographic information was obtained from the State Department of Education website and from each school district's website. School district one is located in a suburban area with an enrollment of 33,500 students. It consists of 31 elementary schools, eight middle schools, and six high schools. The ethnic makeup of its student population was 67.2% White, 23.8% Hispanic, 4.0% Black, 3.9% Asian, and 1.1% Native American. School district two is located in a suburban region and serves 26,000 students. It consists of 16 elementary schools, three K–8 schools, six middle schools, five high schools, and one alternative school. The ethnic makeup of its student population was 83.1% White, 10.5% Hispanic, 2.9% Asian, 1.7% Black, 0.6% Native American, and 1.2% "other."

Instrument

The WISC–IV is an individually administered intelligence test for children of ages 6 years 0 months through 16 years 11 months. The WISC–IV was standardized on 2, 200 children selected as a representative sample of children from the United States. The standardization sample closely corresponded with the composition of the 2000 United States census data on the variables of age, gender, geographic region, ethnicity, and socioeconomic status (Wechsler, 2003).

The WISC–IV contains 10 core subtests (Block Design, Similarities, Digit Span, Matrix Reasoning, Coding, Vocabulary, Letter-Number Sequencing, Symbol Search, Comprehension, and Picture Concepts) and five supplementary subtests (Information, Word Reasoning, Picture Completion, Arithmetic, and Cancellation) with standard score means of 10 and standard deviations of 3. The 10 core subtests combine to form four composite index scores (M = 100, SD = 15); VCI, Perceptual Reasoning Index (PRI), Working Memory Index (WMI), and PSI. The FSIQ is derived from the sum of the 10 core subtest scores.

Results and Discussion

Descriptive statistics, t tests, and stability coefficients for all WISC–IV scores are presented in Table 2. As expected, FSIQ scores were the most stable (r = .82) and subtest scores the least stable, ranging from .46 for Picture Concepts to .70 for Block Design, with a median of .56. Long-term stability of Index scores ranged from .65 for PSI to .76 for PRI (Mdn = .69). These coefficients remained relatively unchanged when corrected for variability of the normative sample (e.g., median subtest stability coefficient increased to .58, median index stability to .73, and FSIQ stability to .84). The PSI score demonstrated the lowest coefficient of internal consistency among the four index scores in the normative sample (.88) and the lowest long-term stability coefficient among the four index scores in every extant stability study (see Tables 1 and 2).

According to the Flynn effect, slight increases in test scores might be expected (Flynn & Weiss, 2007). In contrast, declines in IQ scores on retesting have been reported for samples of children enrolled in special education (Kanaya & Ceci, 2011). In this study, dependent t tests were conducted to examine performance changes from test to retest (see Table 2). Means for the subtests and composite scores were relatively consistent between test and retest, with the largest subtest difference occurring on Coding (−.90). On the composite scores, PSI had the largest difference between test and retest (−1.90). However, effect sizes for both subtest and composite scores were very small (mean d = −0.01), and only three (Coding, Similarities, and Block Design) were statistically significant ($p < .05$). Difference scores were not significantly related to age of the participants (r = .04–.10) and were not significantly different across gender and ethnic groups ($p > .05$).

Individual variations in scores across the test–retest interval are presented in cumulative frequency distributions in Table 3. These distributions reveal that FSIQ test–retest scores diverged by as much as 28 points, VCI scores diverged by up to 31 points, PRI

Table 2
Means, Standard Deviations, and Correlation Coefficients for WISC–IV Test–Retest Interval of 2.84 Years for 344 Students Tested for Special Education Eligibility

Scale	Test		Retest							
	M	*SD*	*M*	*SD*	Δ *M*	Δ *SD*	Δ *d*	r_{12} [95% CI]	r_c	r_t
Similarities	8.79	2.62	9.19	2.76	+0.40*	2.47	+0.15	.580 [.505, .646]	.632	.530
Vocabulary	8.59	2.64	8.44	2.75	−0.15	2.13	−0.06	.688 [.628, .740]	.733	.634
Comprehension	8.87	2.64	8.96	2.60	+0.09	2.67	+0.03	.481 [.394, .559]	.529	.432
Block Design	9.20	2.81	8.70	2.91	−0.50*	2.23	−0.18	.695 [.636, .746]	.718	.587
Picture Concepts	9.54	3.33	10.12	2.92	+0.58*	3.26	+0.19	.463 [.376, .542]	.426	.420
Matrix Reasoning	9.06	2.92	9.18	3.02	+0.12	2.56	+0.04	.629 [.561, .689]	.639	.592
Digit Span	7.98	2.64	7.81	2.54	−0.17	2.33	−0.07	.596 [.523, .660]	.645	.547
Letter-Number Sequencing	8.04	2.82	8.22	3.10	+0.18	3.02	+0.06	.483 [.398, .560]	.506	.458
Coding	8.44	3.16	7.54	2.89	−0.90*	2.98	−0.30	.518 [.436, .591]	.498	.475
Symbol Search	8.47	3.22	8.72	3.10	+0.25	3.05	+0.08	.535 [.455, .606]	.508	.468
Average subtest	8.70	2.88	8.69	2.86	−0.01	2.67	−0.01	.558	.581	.503
Verbal Comprehension	92.54	12.45	93.09	13.21	+0.55	9.59	+0.04	.722 [.667, .769]	.783	.690
Perceptual Reasoning	95.55	14.88	95.92	15.18	+0.37	10.51	+0.03	.756 [.707, .798]	.759	.709
Working Memory	88.27	13.12	88.10	14.04	−0.17	11.31	−0.01	.655 [.590, .712]	.704	.620
Processing Speed	91.44	15.15	89.54	14.99	−1.90	12.62	−0.13	.649 [.583, .706]	.645	.596
Average index	91.95	13.90	91.66	14.36	−0.29	11.07	−0.02	.689	.732	.655
Full Scale IQ	90.32	13.47	90.20	14.21	−0.12	8.45	−0.01	.815 [.776, .848]	.843	.732

Note. WISC–IV = Wechsler Intelligence Scale for Children—Fourth Edition; r_{12} = uncorrected test–retest correlation; CI = confidence interval; r_c = correlation corrected for the variability of the standardization sample; r_t = accounting for additive effect of time and content-sampling error as per Macmann and Barnett (1997); Average = mean for scores and median for coefficients.
* $p < .01$.

scores by up to 35 points, PSI by as much as 40 points, and WMI by up to 41 points. Around 25% of the students exhibited FSIQ scores that differed by 10 or more points, whereas 29%, 39%, 37%, and 44% of the students attained VCI, PRI, WMI, and PSI scores, respectively, that varied by 10 or more points.

The results of this study revealed long-term stability coefficients somewhat lower than those found with the WISC–III (Canivez & Watkins, 1998), but higher than those found in a prior study of the long-term stability of the WISC–IV (Lander, 2010). Although FSIQ was the most stable score in the present study with an *r* of .82, even that might be an optimistic estimate of overall reliability (Schmidt et al., 2003) because stability coefficients are insensible to some sources of measurement error (Viswanathan, 2005). This is also true for internal consistency coefficients. As a result, "reported reliability coefficients tend to overestimate the trustworthiness of educational measures" (Feldt & Brennan, 1993, p. 108). The internal consistency coefficients provided by Wechsler (2003) and the long-term stability coefficients found with the present sample were used to better estimate the combined effects of content and time sampling (Macmann & Barnett, 1997). Taking both internal consistency and stability into account, the estimated reliability coefficients for VCI, PRI, WMI, PSI, and FSIQ were .69, .71, .62, .60, and .73, respectively (see Table 2).

There is no gold standard for reliability coefficients. Instead, measurement experts suggest that reliability coefficients should be considered in the context of how the test scores were obtained and how they are to be used (Thorndike & Thorndike-Christ, 2010). Given this general principle, reliabilities in the .60–.70 range may be acceptable for group decisions, whereas reliabilities greater than .80, .90, or even .95 may be necessary for individual diagnostic decisions (Bracken, 1987; Nunnally & Bernstein, 1994; Salvia & Ysseldyke, 2004).

The group versus individual guidelines provided by Salvia and Ysseldyke (2004) and the conservative approach recommended by Nunnally and Bernstein (1994) permit several general conclusions. First, there was considerable stability of the FSIQ score across the 2.84 year test–retest interval for this heterogeneous group of students. This suggests that FSIQ scores from the WISC–IV can be confidently used for longitudinal screening decisions and group research (Salvia & Ysseldyke, 2004). Using those same standards, the VCI and PRI scores might only be useful for longitudinal group research, but the WMI and PSI scores lack sufficient reliability for long-term group decisions. Unstable subtest scores "suggests that subtest profiles are typically unstable and should not be used for diagnostic and/or decision-making purposes" (Ryan et al., 2010, p. 71). Additionally, around 25% of the participants exhibited FSIQ scores that differed by 10 or more points, and around 6% of the participants earned FSIQ scores that differed by more than 15 points. Thus, even the most reliable WISC–IV score, the FSIQ, may not be sufficiently stable for longitudinal individual decisions. This instability will be magnified if test score patterns (i.e., difference scores) are considered (Thorndike & Thorndike-Christ, 2010).

Of course, these conclusions must be tempered by the limitations of this study. First, archival data were used and were not the product of random selection and assignment. A second limitation is that using cases involving reassessment for special education eligibility means that those students who were no longer eligible for special education placement were not evaluated and therefore were not included in the sample. These concerns are partially dispelled by the results from a study of 50 nonclinical children who were twice administered the French WISC–IV across an interval of 2.64 years that found only VCI and FSIQ scores to exhibit stability coefficients above .80 (Kieng et al., 2012). A third

Table 3
Cumulative Frequency Distributions (in Percentages) of WISC–IV IQ and Index Scores for Test–Retest Interval of 2.84 Years for 344 Students Tested for Special Education Eligibility

Δ	FSIQ	VCI	PRI	WMI	PSI
≥−35			0.3	0.3	0.6
−34			0.4	0.3	0.9
−33			0.6	0.4	1.2
−32			0.7	0.4	1.3
−31		0.3	0.8	0.5	1.3
−30		0.4	0.9	0.5	1.4
−29		0.5	1.0	0.6	1.5
−28	0.3	0.6	1.1	0.7	2.3
−27	0.4	0.7	1.2	0.8	2.9
−26	0.6	0.8	1.4	0.9	3.5
−25	0.7	0.9	1.6	1.2	4.1
−24	0.9	1.0	1.8	1.5	5.2
−23	0.9	1.2	2.0	1.7	5.8
−22	1.0	1.5	2.9	3.5	6.4
−21	1.1	1.7	3.8	3.8	7.8
−20	1.2	2.0	4.1	4.1	8.1
−19	1.7	2.3	4.7	5.2	8.4
−18	2.0	3.2	4.9	5.8	11.3
−17	2.3	4.4	5.8	7.8	12.8
−16	2.9	4.7	6.7	9.0	13.4
−15	3.8	6.1	8.1	11.3	18.0
−14	4.9	7.8	9.0	13.7	18.9
−13	6.7	8.4	10.2	13.7	20.1
−12	8.1	11.0	13.4	17.2	23.5
−11	11.3	12.5	14.2	18.9	24.4
−10	12.5	15.4	18.6	19.2	26.5
−9	15.7	16.0	19.8	23.5	29.7
−8	20.1	17.7	23.0	26.7	32.3
−7	23.0	19.5	24.1	27.6	32.6
−6	24.7	25.3	28.2	31.1	38.1
−5	27.9	26.5	28.5	35.5	42.2
−4	32.3	31.7	34.0	38.1	45.4
−3	37.2	34.0	37.7	40.7	48.5
−2	42.7	39.2	41.3	42.7	49.7
−1	47.4	40.1	47.4	49.3	55.5
0	57.0	52.9	53.5	55.8	61.3
+1	61.0	54.1	56.7	57.0	62.0
+2	64.5	61.0	59.9	58.1	62.8
+3	68.6	64.5	60.5	63.4	67.4
+4	70.3	69.5	66.0	66.0	69.6
+5	74.7	70.6	69.1	68.6	71.8
+6	77.9	76.7	72.1	72.1	77.3
+7	80.8	79.1	73.0	73.0	77.6
+8	84.3	84.0	78.2	79.9	79.4
+9	87.2	86.0	79.7	82.0	82.6
+10	88.7	87.2	83.1	83.4	85.2
+11	90.7	87.8	86.0	86.0	85.8
+12	92.4	89.5	89.0	87.2	88.1
+13	94.2	90.1	89.5	89.0	88.7
+14	96.2	91.3	91.9	91.3	90.4
+15	97.1	92.4	92.7	92.4	92.4
+16	98.0	92.7	95.1	94.2	93.3
+17	98.3	94.8	96.5	95.9	94.2
+18	98.5	96.8	96.8	96.0	95.1
+19	99.1	97.5	97.1	96.1	95.3
+20	99.4	98.3	97.5	96.2	95.9
+21	99.5	98.8	98.0	97.4	96.5
+22	99.7	98.9	98.5	98.3	97.1
+23	99.9	99.1	99.4	98.4	97.4
+24	100	99.2	99.5	98.5	98.3
+25		99.4	99.7	98.6	98.4
+26		99.5	99.9	98.7	98.5
+27		99.7	100	98.8	99.4
+28		99.7		99.1	99.5
+29		99.8		99.4	99.7
+30		99.9		99.7	99.7
+31		100		99.7	99.8
+32				99.8	99.8
+33				99.8	99.9
+34				99.9	99.9
≥ 35				100	100

Note. Column entries represent cumulative percentages of students' change in performance across the retest interval. Change in scores was determined by subtracting the retest score from the initial score. WISC–IV = Wechsler Intelligence Scale for Children—Fourth Edition; FSIQ = Full Scale IQ; VCI = Verbal Comprehension Index; PRI = Perceptual Reasoning Index; WMI = Working Memory Index; and PSI = Processing Speed Index.

limitation is missing data. There were 113 students who were missing at least one WISC–IV score. Their absence from the analyses may have biased the results in some way. This concern is mitigated by a comparison of the total sample of 457 students to the results reported in Table 2: Stability coefficients and mean scores hardly differed. For example, the average subtest stability coefficient differed by only .003, the average subtest score differed by 0.05 points, FSIQ stability coefficients differed by .005, and mean FSIQ scores differed by 0.62 points. A fourth limitation is that there was no information about the examiners who administered the WISC–IV, nor the accuracy of their administration and scoring. However, being twice tested by the same examiner versus different examiners did not significantly affect WISC–IV difference scores.

Finally, it is possible that the assumption of trait stability was violated. Change across time can be the result of (a) change in magnitude of a trait, (b) change in the measurement instrument, or (c) change in the fundamental structure of the trait (Golembiewski, Billingsley, & Yeager, 1976). This sample received special education interventions explicitly designed to improve their school functioning, which might have changed the level or structure of the trait (Flanagan & Kaufman, 2009). Alternatively, there might have been a developmental change in the trait, or the test might have been more sensitive at some ages than others. Unfortunately, it is not possible to determine which of these three sources of change was responsible for the observed instability of WISC–IV scores. Additional studies using a variety of different participant samples and a variety of test–retest time intervals are needed to determine the generalizability of the present findings.

Regardless of these limitations, evidence-based practice calls for use of the best available research evidence (A. Rubin, 2013), and the current unstable WISC–IV results are consistent with those reported by other researchers (Lander, 2010; Ryan et al., 2010). Thus, these results provide an important starting point for future research and sound a critical warning for clinical practice. Specifically, clinicians should not assume that WISC–IV scores will be consistent across long test–retest intervals for individual students and should question recertification of eligibility for special education on the basis of historical WISC–IV scores.

References

Adams, K. M. (2000). Practical and ethical issues pertaining to test revisions. *Psychological Assessment, 12,* 281–286. doi:10.1037/1040-3590.12.3.281

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. doi:10.1002/1097-4679(198911)45:6<941::AID-JCLP2270450619>3.0.CO;2-P

Bauman, E. (1991). Determinants of WISC-R subtest stability in children with learning difficulties. *Journal of Clinical Psychology, 47,* 430–435. doi:10.1002/1097-4679(199105)47:3<430::AID-JCLP2270470317>3.0.CO;2-N

Bracken, B. A. (1987). Limitations of preschool instruments and standards for minimal levels of technical adequacy. *Journal of Psychoeducational Assessment, 5,* 313–326. doi:10.1177/073428298700500402

Canivez, G. L., & Watkins, M. W. (1998). Long-term stability of the Wechsler Intelligence Scale for Children—Third Edition. *Psychological Assessment, 10,* 285–291. doi:10.1037/1040-3590.10.3.285

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 Psychological Assessment, 17,* 300–313. doi:10.1177/073428299901700401

Canivez, G. L., & Watkins, M. W. (2001). Long-term stability of the Wechsler Intelligence Scale for Children—Third Edition among students with disabilities. *School Psychology Review, 30,* 438–453.

Charter, R. A. (1999). Sample size requirements for precise estimates of reliability, generalizability, and validity coefficients. *Journal of Clinical and Experimental Neuropsychology, 21,* 559–566. doi:10.1076/jcen.21.4.559.889

Chen, Z., & Siegler, R. S. (2000). Intellectual development in childhood. In R. J. Sternberg (Ed.), *Handbook of intelligence* (pp. 92–116). New York, NY: Cambridge University Press. doi:10.1017/CBO9780511807947.006

Ellzey, J. T., & Karnes, F. A. (1990). Test-retest stability of WISC-R IQs among gifted students. *Psychological Reports, 66,* 1023–1026. doi:10.2466/pr0.1990.66.3.1023

Feldt, L. S., & Brennan, R. L. (1993). Reliability. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 105–146). Phoenix, AZ: Oryx Press.

Flanagan, D. P., & Kaufman, A. S. (2009). *Essential of WISC-IV assessment* (2nd ed.). Hoboken, NJ: Wiley.

Flynn, J. R., & Weiss, L. G. (2007). American IQ gains from 1932 to 2002: The WISC subtests and educational progress. *International Journal of Testing, 7,* 209–224. doi:10.1080/15305050701193587

Golembiewski, R. T., Billingsley, K., & Yeager, S. (1976). Measuring change and persistence in human affairs: Types of change generated by OD designs. *Journal of Applied Behavioral Science, 12,* 133–157. doi:10.1177/002188637601200201

Hunt, E. (2010). *Human intelligence*. New York, NY: Cambridge University Press.

Johnson, W., Gow, A. J., Corley, S., Starr, J. M., & Deary, I. J. (2010). Location in cognitive and residential space at age 70 reflects a lifelong trait over parental and environmental circumstances: The Lothian birth cohort 1936. *Intelligence, 38,* 402–411. doi:10.1016/j.intell.2010.04.001

Kamphaus, R. W., Petoskey, M. D., & Rowe, E. W. (2000). Current trends in psychological testing of children. *Professional Psychology: Research and Practice, 31,* 155–164. doi:10.1037/0735-7028.31.2.155

Kanaya, T., & Ceci, S. J. (2011). The Flynn effect in the WISC subtests among school children tested for special education services. *Journal of Psychoeducational Assessment, 29,* 125–136. doi:10.1177/0734282910370139

Kieng, S., Reverte, I., Scherrer, N., Favez, N., Rossier, J., & Lecerf, T. (2012, July). *Long-term stability of the French WISC-IV: An exploratory study*. Poster presented at the meeting of the International Test Commission, Amsterdam, the Netherlands.

Krouse, H. E., & Braden, J. P. (2011). The reliability and validity of WISC-IV scores with deaf and hard-of-hearing children. *Journal of Psychoeducational Assessment, 29,* 238–248. doi:10.1177/0734282910383646

Lander, J. (2010). Long-term stability of scores on the Wechsler Intelligence Scale for Children—Fourth Edition in children with learning disabilities. *Dissertation Abstracts International: Section A. Humanities and Social Sciences*.

Mackintosh, N. J. (1998). *IQ and human intelligence*. New York, NY: Oxford University Press.

Macmann, G. M., & Barnett, D. W. (1997). Myth of the master detective: Reliability of interpretations for Kaufman's "Intelligent Testing" approach to the WISC-III. *School Psychology Quarterly, 12,* 197–234. doi:10.1037/h0088959

Madaus, J. W., & Shaw, S. F. (2006). The impact of the IDEA 2004 on transition to college for students with learning disabilities. *Learning Disabilities Research & Practice, 21,* 273–281. doi:10.1111/j.1540-5826.2006.00223.x

Mishra, S. P., & Lord, J. (1982). Reliability and predictive validity of the WISC-R with native-American Navajos. *Journal of School Psychology, 20,* 150–154. doi:10.1016/0022-4405(82)90008-5

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.

Nunnally, J. C., & Bernstein, I. H. (1994). *Psychometric theory* (3rd ed.). New York, NY: McGraw-Hill.

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. doi:10.1002/1520-6807(198804)25:2<118::AID-PITS2310250204>3.0.CO;2-T

Quereshi, M. Y. (1968). The internal consistency of the WISC scores for ages 5 to 16. *Journal of Clinical Psychology, 24,* 192–195. doi:10.1002/1097-4679(196804)24:2<192::AID-JCLP2270240216>3.0.CO;2-H

Reeve, C. L., & Bonaccio, S. (2011). On the myth and the reality of the temporal validity degradation of general mental ability test scores. *Intelligence, 39,* 255–272. doi:10.1016/j.intell.2011.06.009

Reschly, D. J., & Bergstrom, M. K. (2009). Response to intervention. In T. B. Gutkin & C. R. Reynolds (Eds.), *The handbook of school psychology* (4th ed., pp. 434–460). Hoboken, NJ: Wiley.

Revelle, W. (2010). *An introduction to psychometric theory with applications in R*. Retrieved from http://www.personality-project.org/r/book/

Reynolds, C. R., & Milam, D. A. (2012). Challenging intellectual testing results. In D. Faust (Ed.), *Coping with psychiatric and psychological testimony* (6th ed., pp. 311–334). New York, NY: Oxford University Press.

Rubin, A. (2013). *Statistics for evidence-based practice and evaluation* (3rd ed.). Belmont, CA: Brooks/Cole.

Rubin, D. B. (1976). Inference and missing data. *Biometrika, 63,* 581–592. doi:10.1093/biomet/63.3.581

Ryan, J. J., Glass, L. A., & Bartels, J. M. (2009). Internal consistency reliability of the WISC-IV among primary school students. *Psychological Reports, 104,* 874–878. doi:10.2466/pr0.104.3.874-878

Ryan, J. J., Glass, L. A., & Bartels, J. M. (2010). Stability of the WISC-IV in a sample of elementary and middle school children. *Applied Neuropsychology, 17,* 68–72. doi:10.1080/09084280903297933

Salvia, J., & Ysseldyke, J. E. (2004). *Assessment in special and inclusive education* (9th ed.). Boston, MA: Houghton Mifflin Company.

Schmidt, F. L., Le, H., & Ilies, R. (2003). Beyond alpha: An empirical examination of the effects of different sources of measurement error on reliability estimates for measures of individual differences constructs. *Psychological Methods, 8,* 206–224. doi:10.1037/1082-989X.8.2.206

Schuerger, J. M., & Witt, A. C. (1989). The temporal stability of individually tested intelligence. *Journal of Clinical Psychology, 45,* 294–302. doi:10.1002/1097-4679(198903)45:2<294::AID-JCLP2270450218>3.0.CO;2-N

Silverstein, A. B. (1969). The internal consistency of the Stanford-Binet. *American Journal of Mental Deficiency, 73,* 753–754.

Simonton, D. K. (2011). Exceptional talent and genius. In T. Chamorro-Premuzic, S. von Stumm, & A. Furnham (Eds.), *Wiley-Blackwell handbook of individual differences* (pp. 635–655). Malden, MA: Blackwell.

Smith, M. D. (1978). Stability of WISC-R subtest profiles for learning disabled children. *Psychology in the Schools, 15,* 4–7. doi:10.1002/1520-6807(197801)15:1<4::AID-PITS2310150102>3.0.CO;2-S

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. doi:10.1002/1520-6807(199004)27:2<101::AID-PITS2310270202>3.0.CO;2-D

Strauss, E., Sherman, E. M. S., & Spreen, O. (2006). *A compendium of neuropsychological tests: Administration, norms, and commentary*. New York, NY: Oxford University Press.

Strauss, E., Spreen, O., & Hunter, M. (2000). Implications of test revisions for research. *Psychological Assessment, 12,* 237–244. doi:10.1037/1040-3590.12.3.237

Thorndike, R. M., & Thorndike-Christ, T. (2010). *Measurement and evaluation in psychology and education* (8th ed.). New York, NY: Pearson.

Tuma, J. M., & Applebaum, A. S. (1980). Reliability and practice effects of WISC-R IQ estimates in a normal population. *Educational and Psychological Measurement, 40,* 671–678. doi:10.1177/0013164480040003 10

Viswanathan, M. (2005). *Measurement error and research design*. Thousand Oaks, CA: Sage.

Watkins, M. W. (2010). Structure of the Wechsler Intelligence Scale for Children—Fourth Edition among a national sample of referred students. *Psychological Assessment, 22,* 782–787.

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

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

Wechsler, D. (2003). *Wechsler Intelligence Scale for Children—Fourth Edition (WISC–IV) technical and interpretive manual*. San Antonio, TX: Psychological Corporation.

Wright, A. J. (2011). *Conducting psychological assessment: A guide for practitioners*. Hoboken, NJ: Wiley.

Zhu, J., Tulsky, D. S., Price, L., & Chen, H.-Y. (2001). WAIS-III reliability data for clinical groups. *Journal of the International Neuropsychological Society, 7,* 862–866.

Received September 30, 2012
Revision received December 14, 2012
Accepted December 17, 2012 ■