APPLIED NEUROPSYCHOLOGY: CHILD
2022, VOL. 11, NO. 3, 422–428
https://doi.org/10.1080/21622965.2021.1875827

Long-term stability of Wechsler Intelligence Scale for Children–fifth edition scores in a clinical sample

Marley W. Watkins[a], Gary L. Canivez[b], Stefan C. Dombrowski[c], Ryan J. McGill[d], Alison E. Pritchard[e], Calliope B. Holingue[f], and Lisa A. Jacobson[e]

[a]Department of Educational Psychology, Baylor University, Waco, Texas, USA; [b]Department of Psychology, Eastern Illinois University, Charleston, Illinois, USA; [c]Department of Graduate Education, Leadership and Counseling, Rider University, Lawrenceville, New Jersey, USA; [d]Department of School Psychology and Counselor Education, William & Mary, Williamsburg, Virginia, USA; [e]Department of Neuropsychology, Kennedy Krieger Institute, Baltimore, Maryland, USA; [f]Johns Hopkins University Bloomberg School of Public Health, Baltimore, Maryland, USA

ABSTRACT

This study investigated the stability of Wechsler Intelligence Scale for Children–Fifth Edition (WISC-V) scores for 225 children and adolescents from an outpatient neuropsychological clinic across, on average, a 2.6 year test-retest interval. WISC-V mean scores were relatively constant but subtest stability score coefficients were all below 0.80 ($M = 0.66$) and only the Verbal Comprehension Index (VCI), Visual Spatial Index (VSI), and omnibus Full Scale IQ (FSIQ) stability coefficients exceeded 0.80. Neither intraindividual subtest difference scores nor intraindividual composite difference scores were stable across time ($M = 0.26$ and 0.36, respectively). Rare and unusual subtest and composite score differences as well as subtest and index scatter at initial testing were unlikely to be repeated at retest (kappa = 0.03 to 0.49). It was concluded that VCI, VSI, and FSIQ scores might be sufficiently stable to support normative comparisons but that none of the intraindividual (i.e. idiographic, ipsative, or person-relative) measures were stable enough for confident clinical decision making.

KEYWORDS
Assessment; reliability; stability; WISC-V

The Wechsler Intelligence Scale for Children–Fifth Edition (WISC-V; Wechsler, 2014a) is one of the most frequently used tests in clinical practice (Benson et al., 2019; Groth-Marnat & Wright, 2016). Although it can produce a plethora of scores, clinical applications of the WISC-V often focus on its ten primary subtest scores, five primary index scores, and omnibus Full Scale score (FSIQ; Freeman & Chen, 2019). Considerable evidence regarding the reliability and validity of WISC-V scores has been provided by its publisher (Wechsler, 2014b) and independent researchers (e.g. Canivez et al., 2020; Farmer & Kim, 2020). Based on this evidence, the WISC-V has been judged to be psychometrically sound (Groth-Marnat & Wright, 2016).

Recommendations for clinical interpretation of WISC-V scores are often based on successive-level approaches designed to estimate the examinee's: (a) general intellectual ability; (b) broad intellectual abilities; and (c) cognitive strengths and weaknesses within both nomothetic and idiographic frameworks (Freeman & Chen, 2019; Groth-Marnat & Wright, 2016; Kaufman et al., 2016; Sattler et al., 2016; Wechsler, 2014b). There is some variability among these approaches, but most place considerable emphasis on estimation of general and broad intellectual abilities followed by identification of cognitive strengths and weaknesses. In current practice, the WISC-V composite scores (i.e. FSIQ and factor index scores) "are the primary level of analysis, because they are the most reliable and comprehensive representatives of the child's performance" (Kaufman et al., 2016, p. 232).

Nomothetic framework

WISC-V scores reflect how well an individual performs relative to the national standardization sample and are, therefore, "population-relative metrics" (McDermott et al., 1992, p. 505). Nomothetic interpretations are based on these norm-referenced scores (Freeman & Chen, 2019), and extremely low or high scores might have diagnostic implications (i.e. special education or gifted programs). The verity of nomothetic interpretation rests on the reliability of WISC-V scores because reliability constrains validity (Thorndike & Thorndike-Christ, 2010; Wasserman & Bracken, 2013); that is, how consistent scores are across items (internal consistency reliability), raters or examiners (interrater reliability), and test occasions (test-retest reliability or stability). Wechsler (2014b) provided considerable evidence regarding the internal consistency, interrater reliability, and short-term (i.e. <3 months) stability of WISC-V scores with the standardization sample, but did not provide any evidence about *long-term* (i.e. >12 months) stability.

CONTACT Marley W. Watkins ✉ marley_watkins@baylor.edu Department of Educational Psychology, Baylor University, Waco, TX, USA.
This article has been corrected with minor changes. These changes do not impact the academic content of the article.

Temporal stability is consequential because decisions about individuals based on intelligence test scores may have long-term effects (Watkins & Smith, 2013). This is especially pertinent for decisions regarding program eligibility because those decisions may not be empirically reevaluated for several years (Borreca et al., 2013). However, long-term stability assumes that the construct measured by test scores is sufficiently stable across time. Fortunately, intelligence is presumed to be an enduring trait and intelligence test scores have been found to be relatively stable from childhood through adulthood (Hunt, 2011; Mackintosh, 2011; Schuerger & Witt, 1989).

There is presently no evidence regarding the long-term stability of WISC-V scores among clinical examinees. As noted by Thorndike and Thorndike-Christ (2010), reliability estimates obtained from standardization samples likely approximate the maximum because they were collected under strictly controlled conditions. In contrast, when a test is used in clinical practice, examiners may not be so specially trained, test conditions as closely controlled, and scoring errors as limited (McDermott et al., 2014; Styck & Walsh, 2016).

Wasserman and Bracken (2013) suggested that the validity of high-stakes decisions about individuals require coefficients of internal consistency and stability ≥0.90. However, the length of the test-retest interval influences stability coefficients with longer intervals negatively impacting the stability of scores (Bandalos, 2018). For example, a meta-analysis of test-retest stability coefficients of intelligence test scores found that coefficients were, on average, 0.89 for intervals of 0–10 months and decreased to 0.80 for longer intervals (Schuerger & Witt, 1989). Given these empirical results, 0.80 may be a more reasonable goal for long-term stability.

Idiographic framework

Following their nomothetic interpretation, idiographic comparisons among WISC-V scores are often employed by practitioners to identify a profile of intraindividual cognitive strengths and weaknesses (Freeman & Chen, 2019; Groth-Marnat & Wright, 2016; Kaufman et al., 2016; Miller et al., 2016; Wechsler, 2014b). Concretely, each score is subtracted from the examinee's average or FSIQ to create a profile of difference scores wherein a negative value is thought to represent an idiographic weakness and a positive value is thought to represent an idiographic strength (Kaufman et al., 2016). Within such a framework, score profiles are seen as more useful for interpretation than the scores themselves because they focus on within-person performance in contrast to the between-person performance emphasized by the nomothetic approach (Styck et al., 2019). Idiographic scores were also called ipsative scores by McDermott et al. (1992) who described them as "person-relative metrics" (p. 505). Historically, subtest scores were used for these comparisons but "contemporary approaches have minimized emphasis of comparisons between subtests" (Farmer & Kim, 2020, p. 2) due to "lack of evidence supporting subtest analysis" (McGill et al., 2018, p. 110). Nevertheless, these idiographic interpretational approaches have achieved wide-spread clinical application and remain popular among practitioners and trainers (Benson et al., 2020; Miller et al., 2016).

The validity of idiographic interpretations depends on the reliability of the difference scores upon which those interpretations are based (American Educational Research Association (AERA), American Psychological Association (APA), & National Council on Measurement in Education (NCME), 2014; Freeman & Chen, 2019; Wasserman & Bracken, 2013). Statistically, "the reliability of differences between two scores can be lower than the reliability of the individual scores" (Bandalos, 2018, p. 202). In essence, the true score components in the two test scores overlap whereas the error accumulates. A recent study investigated WISC-V difference scores with its standardization sample and found that the median subtest difference score reliability was 0.70 and the median composite difference score reliability was 0.81 (Farmer & Kim, 2020). However, the reliability of WISC-V difference scores among clinical samples has yet to be investigated so it is presently unknown whether these estimates will replicate in more focal populations (Thorndike & Thorndike-Christ, 2010).

The identification of cognitive strengths and weaknesses with WISC-V difference scores underpins idiographic recommendations for remedial strategies, classroom modifications, instructional accommodations, curricular modifications, targeted interventions, and program placements (Courville et al., 2016; Groth-Marnat & Wright, 2016; Kaufman et al., 2016; Miller et al., 2016; Sattler et al., 2016; Wechsler, 2014b), which are likely to have long lasting effects on examinees. For example, "any long-term recommendations as to a strategy for teaching a student would need to be based on aptitudes that are likely to remain stable for months, if not years" (Cronbach & Snow, 1977, p. 161). To the extent that WISC-V difference scores are not consistent across time, "their potential for accurate prediction of criteria, for beneficial examinee diagnosis, and for wise decision making is limited" (American Educational Research Association (AERA), American Psychological Association (APA), & National Council on Measurement in Education (NCME), 2014, p. 35) and will "lead to poor-quality clinical inferences" (Bowden & Finch, 2017, p. 103).

Current study

In summary, WISC-V scores are commonly interpreted in clinical practice based on: (a) nomothetic reference to the normative sample (i.e. population-relative or between-person metrics) and (b) idiographic reference to differences among scores assumed to reflect an individual's cognitive strengths and weaknesses (i.e. ipsative, person-relative, or within-person metrics). However, there is no extant evidence regarding the long-term temporal stability of WISC-V scores for a clinical sample. The current study addresses that evidential lacuna.

Method

Participants & procedure

Participants were 225 children and adolescents (160 male and 65 female) who were twice administered all ten of the

WISC-V primary subtests as part of assessments conducted in a large outpatient neuropsychological clinic in the mid-Atlantic region of the United States between October 2014 and March 2020. Participants' average age at initial testing was 9.1 ($SD = 2.1$, range of 6.1–14.8) years and at retest was 11.7 ($SD = 2.2$, range of 7.4–16.8) years for an average test-retest interval of 2.6 ($SD = 0.9$, range of 0.2–5.1) years. Participants' ethnic background was 51.6% White, 28.0% Black, 8.0% Multi-Racial, 6.7% Hispanic, 3.1% Asian, and 3.1% other or missing. Although individual socioeconomic data was not available, private insurance was used by 58.7% of the participants and public insurance by 41.3% of the participants. Billing codes indicated that approximately 40% of the sample was referred for medical concerns (40 with encephalopathy [a code used for multiple neurodevelopmental disorders], 22 with cancer, 8 with a genetic condition, 6 with congenital malformations, 4 with epilepsy, etc.) and 60% for mental health concerns (110 with ADHD, 9 with anxiety, 6 with adjustment disorder, 5 with conduct disorder, 2 with depression, etc.). Among the participants with medical concerns, 64 experienced neurological problems (encephalopathy, epilepsy, nervous system neoplasms).

In total, 39 separate psychologists appropriately credentialed in this jurisdiction (5 PsyD and 34 PhD, 17 neuropsychology and 22 clinical specialty, 10 board certified) assessed these participants. The number of children seen by each psychologist at each test occasion ranged from 1 to 27 and each psychologist evaluated, on average, 3% of the sample. All of these providers completed clinical predoctoral internships as well as supervised post-doctoral fellowship training. De-identified data were extracted from a database maintained by the clinic following approval by the hospital's institutional review board.

Instrument

The WISC-V is an individually administered test of cognitive ability for children ages 6–16 years. The FSIQ is composed of seven primary subtests: Similarities (SI), Vocabulary (VO), Block Design (BD), Matrix Reasoning (MR), Figure Weights (FW), Digit Span (DS), and Coding (CD). The Visual Puzzles (VP), Picture Span (PS), and Symbol Search (SS) subtests can be added to the battery to compute five primary index scores, each composed of two subtests: SI and VO for the Verbal Comprehension Index (VCI); BD and VP for the Visual Spatial Index (VSI); MR and FW for the Fluid Reasoning Index (FRI); DS and PS for the Working Memory Index (WMI); and CD and SS for the Processing Speed Index (PSI). Subtest scaled scores have means of 10 and standard deviations of 3, whereas standard index scores have means of 100 and standard deviations of 15. Detailed descriptions of WISC-V measures are available in the *Technical and Interpretive Manual* (Wechsler, 2014b) and prominent interpretive resources (e.g. Kaufman et al., 2016; Sattler et al., 2016).

Table 1. Nomothetic comparisons of Wechsler Intelligence Scale for Children and Adolescent–Fifth Edition Scores for 225 children in a clinical sample twice tested across, on average, a 2.6 year interval.

	Test		Retest		Retest-test difference		
Score	Mean	SD	Mean	SD	Mean*	|*d*|	*r*[a]
Subtest							
BD	8.81	3.09	8.60	3.33	−0.22	0.07	0.68 [0.60, 0.75]
SI	8.81	3.29	9.04	2.80	0.24	0.07	0.68 [0.61, 0.75]
MR	8.77	3.26	9.04	3.05	0.27	0.09	0.59 [0.49, 0.67]
DS	7.62	3.05	7.90	3.08	0.28	0.09	0.76 [0.70, 0.81]
CD	7.49	3.35	7.31	3.22	−0.18	0.05	0.66 [0.58, 0.73]
VO	8.76	3.58	8.83	3.50	0.07	0.02	0.79 [0.74, **0.84**]
FW	9.54	2.89	9.20	3.15	−0.34	0.11	0.53 [0.43, 0.62]
VP	9.59	3.26	9.60	3.15	0.01	0.00	0.75 [0.69, 0.80]
PS	8.52	3.09	8.45	3.13	−0.07	0.02	0.50 [0.40, 0.60]
SS	7.68	3.44	7.80	3.17	0.12	0.04	0.62 [0.54, 0.70]
Composite							
VCI	93.32	17.46	94.00	16.38	0.68	0.04	**0.84** [0.79, 0.87]
VSI	95.68	15.96	95.26	16.92	−0.42	0.04	**0.82** [0.77, 0.86]
FRI	95.39	15.38	94.84	16.39	−0.55	0.02	0.69 [0.61, 0.75]
WMI	88.38	15.10	89.15	15.31	0.77	0.05	0.74 [0.67, 0.79]
PSI	86.47	17.52	86.19	17.02	−0.28	0.03	0.77 [0.71, 0.82]
FSIQ	89.97	16.03	89.98	16.42	0.01	0.00	**0.86** [0.82, 0.89]

Note. BD: Block Design; SI: Similarities; MR: Matrix Reasoning; DS: Digit Span; CD: Coding; VO: Vocabulary; FW: Figure Weights; VP: Visual Puzzles; PS: Picture Span; SS: Symbol Search; VCI: Verbal Comprehension Index; VSI: Visual Spatial Index; FRI: Fluid Reasoning Index; WMI: Working Memory Index; PSI: Processing Speed Index; FSIQ: Full Scale IQ; *SD*: standard deviation; *d*: standardized mean difference; and *r*: test-retest correlation.

[a] *r* and 95% confidence limits for total sample. Coefficients ≥0.80 in bold.

*No mean WISC-V score differences were statistically significant with the experiment-wise error rate held at 0.05 (Holm, 1979).

Results

Descriptive statistics for WISC-V test and retest scores were computed with Stata version 16.1 and are presented in Table 1. Overall, mean subtest and composite scores at both test and retest were slightly below average, but within one standard deviation of population means, as is common in clinical samples. All subtests and composite scores showed univariate normal distributions with no appreciable skewness or kurtosis (maximum skew of 0.34 and maximum kurtosis of 0.58).

Nomothetic comparisons

As detailed in Table 1, the differences in WISC-V subtest scores and primary index scores across time were small (mean $d = 0.02$ for subtests and 0.03 for composite scores). None of these differences were statistically significant when holding the experiment-wise error rate at 0.05 using Holm's (1979) sequential Bonferroni method. Likewise, there were no statistically significant differences when smaller subsamples based on age, insurance type, Black versus other ethnic groups, sex, medical versus psychological concerns, etc. were tested. On average, the test-retest FSIQ scores differed by less than 1 standard score point but 8.7% of the FSIQ scores, 10.5% of the VCI scores, 10.0% of the VSI scores, 16.1% of the FRI scores, 13.7% of the WMI scores, and 14.2% of the PSI scores changed by more than 15 points from test to retest.

Subtest stability coefficients ranged from 0.50 (PS) to 0.79 (VO) with *M* of 0.66. Primary index score stability coefficients ranged from 0.69 (FRI) to 0.84 (VCI) with a *M* of

0.77. VCI and VSI scores exceeded the minimum reliability standard of 0.80 but the stability of the FRI, WMI, and PSI scores were all below 0.80. The most stable WISC-V score was the FSIQ ($r = 0.86$). Consequently, it appears that only the VCI, VSI, and FSIQ scores are sufficiently reliable in the long-term to support nomothetic clinical decisions. These results are generally consistent with the long-term stability of prior versions of the WISC among both clinical and non-referred samples (e.g. Canivez & Watkins, 1998; Kieng et al., 2015). For example, the FSIQ has always been the most stable WISC score and the composite scores the next most stable but often lower than 0.80 (Bartoi et al., 2015; Watkins & Smith, 2013).

The length of the test-retest interval and age at first testing had a small effect on the stability of WISC-V scores. The correlations between retest interval and composite difference scores averaged −0.07, suggesting that score stability may have decreased as the retest interval increased. In contrast, the correlations between age at first testing and composite scores were positive ($M = 0.08$), indicating that score stability tended to increase with age of the participant. However, all of these correlation estimates included zero in their 95% CI, demonstrating a lack of statistical significance. Additionally, the test-retest interval accounted for less than 1% of the variance in composite score differences and the age at first testing accounted for less than 2% of the variance in composite score differences. Thus, neither the test-retest interval nor age of participants seemed to have a substantial effect on score stability.

Table 2. Idiographic comparisons of Wechsler Intelligence Scale for Children and Adolescents–Fifth Edition scores in a clinical sample of 225 children retested, on average, after 2.6 years.

	Test			Retest			Stability	
Score	Mean	SD	Rare[a]	Mean	SD	Rare[a]	r[b]	Kappa[c]
Subtest								
BD	1.69	1.32	8	1.66	1.24	5	0.32 [0.20, 0.43]	0.29
SI	1.78	1.41	8	1.52	1.14	3	0.20 [0.07, 0.32]	−0.02
MR	1.81	1.48	7	1.65	1.34	5	0.06 [-0.07, 0.20]	−0.03
DS	1.83	1.31	4	1.85	1.34	4	0.35 [0.23, 0.46]	0.49
CD	2.14	1.71	15	2.14	1.80	18	0.42 [0.31, 0.53]	0.25
VO	1.99	1.44	10	1.87	1.29	2	0.38 [0.26, 0.49]	0.32
FW	1.93	1.40	6	1.64	1.34	5	0.09 [-0.04, 0.22]	0.16
VP	1.96	1.49	10	1.78	1.41	7	0.37 [0.25, 0.47]	0.33
PS	2.01	1.52	10	1.99	1.45	11	0.19 [0.07, 0.32]	0.15
SS	2.18	1.57	10	1.97	1.43	10	0.17 [0.04, 0.29]	0.16
Scatter	7.59	2.27	16	7.11	2.28	11	0.34 [0.22, 0.45]	0.33
Composite								
VCI	7.59	6.58	31	7.56	6.09	29	0.40 [0.28, 0.51]	0.33
VSI	10.14	7.34	53	9.36	7.26	48	0.43 [0.31, 0.53]	0.28
FRI	8.62	6.47	34	8.11	6.50	35	0.32 [0.19, 0.43]	0.21
WMI	9.01	6.99	13	9.40	6.95	17	0.23 [0.10, 0.35]	0.22
PSI	10.87	8.45	32	10.79	8.55	31	0.43 [0.31, 0.53]	0.41
Scatter	26.80	11.94	21	26.44	11.18	19	0.34 [0.22, 0.45]	0.35

Note. BD: Block Design; SI: Similarities; MR: Matrix Reasoning; DS: Digit Span; CD: Coding; VO: Vocabulary; FW: Figure Weights; VP: Visual Puzzles; PS: Picture Span; SS: Symbol Search; VCI: Verbal Comprehension Index; VSI: Visual Spatial Index; FRI: Fluid Reasoning Index; WMI: Working Memory Index; PSI: Processing Speed Index; FSIQ: Full Scale IQ; *SD*: standard deviation; *d*: standardized mean difference; *r*: test-retest correlation.

[a] Participants with rare and unusual score differences of 5 points between the mean of the 10 primary subtest scores and each primary subtest score; differences of 15 points between the FSIQ score and VCI, VSI, and PRI scores; and 21 points between the FSIQ and WMI and PSI scores (Kaufman et al., 2016) or subtest and index scores with intraindividual variability (or scatter) of ≥12 and ≥44 points, respectively (Courville et al., 2016).

[b] *r* and 95% confidence limits for mean differences between test and retest.

[c] Standard error of kappa ranged from 0.065 to 0.067.

Idiographic comparisons

Idiographic comparisons were based on the 'intelligent rules of thumb' provided by Kaufman et al. (2016) and the 'clinically meaningful' levels of score variability or scatter reported by Courville et al. (2016).

Intraindividual subtest and index score differences

On average, intraindividual subtest score differences from their respective mean were 1.93 points at initial testing and 1.81 points at retest while intraindividual index score differences from the FSIQ were 9.25 points at initial testing and 9.04 points at retest (Table 2). Nevertheless, the stability across time of those score differences was poor, with correlations ranging from 0.06 for MR to 0.43 for VSI and PSI. Consequently, none of the WISC-V score differences were sufficiently reliable in the long term to support clinical decision making.

This poor long-term stability is not surprising given that the median subtest difference score reliability was 0.70 and the median composite difference score reliability was 0.81 for the WISC-V standardization sample (Farmer & Kim, 2020). When repeated across time to assess their stability, the reliability of these difference scores would be expected to deteriorate (Bandalos, 2018). Poor test-retest stability coefficients (e.g. 0.05–0.45) were also reported for score discrepancies across an 11 month test-retest interval on a previous version of the WISC (Ryan et al., 2010). Likewise, the reliability of subtest and composite profile scores on an earlier iteration of the WISC was estimated to be 0.37 and 0.53, respectively (Styck et al., 2019).

Differences ≥5 points between the mean of the 10 subtests and each subtest were defined as "significant and unusual" (Kaufman et al., 2016, p. 244) and differences ≥15 points between the FSIQ score and VCI, VSI, and PRI scores and ≥21 points between the FSIQ and WMI and PSI scores were considered to be "rare and unusual" (p. 242). In total, one to four rare and unusual subtest score differences were exhibited by 28% of the participants at initial testing and 24% at retest. In contrast, one or more rare and unusual index score differences were displayed by 53% of the participants at both initial testing and retest. However, these rates were not consistent across time. For example, 40% of participants with no rare index score difference at initial testing displayed one or more rare difference at retest, while 64% of participants with one or more rate index score difference at initial testing displayed one or more rare difference at retest.

The number of rare and unusual score differences for each subtest and index at both test and retest are reported in Table 2. Although relatively consistent (e.g. 8 vs. 5 for BD, 31 vs. 29 for VCI at test and retest, respectively), rare and unusual differences were not stable across time. That is, a rare and unusual difference for a subtest or index difference score at initial testing was unlikely to be repeated at retest or vice versa. This tendency was quantified by kappa (Cohen, 1960), which expresses the proportion of agreement beyond what would be expected by chance. Kappa coefficients ranged from −0.03 to 0.49 for rare subtest score differences and from 0.19 to 0.39 for rare composite score differences. These kappa values indicate poor agreement on rare score discrepancies across the test-retest interval (Wasserman & Bracken, 2013). Agreement on rare score discrepancies across time was also examined for sub-groups (i.e. gender, ethnicity, type of insurance, type of disorder, etc.) with similar results, but there were too few participants for stable estimates.

Overall, idiographic score comparisons were too unstable over time for confident clinical decision making. Similar near chance results were obtained when idiographic scores on a prior version of the WISC were analyzed longitudinally (Kieng et al., 2015; Watkins & Canivez, 2004). Theoretically, these results were foreshadowed by McDermott et al. (1992) who explored the reliability and validity of person-relative scores and found them to be inferior to population-relative scores.

Intraindividual subtest and index score scatter

It has been proposed that unusual intraindividual subtest and index score variability or scatter has "clinically meaningful implications" for WISC-V score interpretation (Courville et al., 2016, p. 225) and signifies "that a child has unique strengths and weaknesses and may benefit from specialized instruction" (Sattler et al., 2016, p. 176). Accordingly, intraindividual variability among subtest and index scores of ≥12 and ≥44 points, respectively, were considered rare and unusual at the 5% level (Courville et al., 2016).

On average, the normative sample exhibited subtest score scatter of 7.0 (*SD* = 2.2) points and index score scatter of 25.1 (*SD* = 10.2) points (Kaufman et al., 2016). Results from this clinical sample were relatively equivalent, with mean subtest scatter of 7.4 (*SD* = 2.3) points and mean index score scatter of 26.6 (*SD* = 17.5) points. As with intraindividual score differences, rare and unusual scatter was not stable across time: kappa coefficients for the presence of rare and unusual scatter were 0.33 and 0.35 for subtest and index scatter, respectively. As with rare and unusual score differences, there were too few participants for stable estimates with sub-groups. Overall, rare and unusual intraindividual variability at initial testing was unlikely to be repeated at retest and vice versa. These results are consistent with research that found IQ score scatter to exhibit poor validity (McGill, 2018; Watkins, 2005; Watkins & Glutting, 2000) given that poor reliability likely constrains psychometric validity (Bandalos, 2018).

Summary & conclusions

Psychologists often interpret WISC-V scores by nomothetic reference to the normative sample and by idiographic reference to within-person score differences to identify intraindividual cognitive strengths and weaknesses. This study investigated the temporal stability of WISC-V scores for a clinical sample twice assessed across an average 2.6 year test-retest interval in an outpatient neuropsychological clinic. From a nomothetic perspective, only the VCI, VSI, and FSIQ scores were sufficiently reliable (≥0.80) in the long-term to support clinical decision making. Although many of the participants demonstrated rare and unusual intratest score differences, those differences replicated across test-retest occasions at near chance levels. That is, a cognitive strength or weakness identified by WISC-V difference scores would likely not be repeated in a later administration of the WISC-V. Likewise, unusual intraindividual subtest and index score scatter did not replicate across time.

As with all research, these results must be considered within the limits of its design and sample. Reliability is sample dependent, so results may differ in other clinical samples (Bandalos, 2018). A host of influences within school, psychosocial, and family environments might affect the stability of WISC-V scores (Bronfenbrenner & Morris, 2006). In particular, the selection of participants for re-administration of the WISC-V may have introduced bias. Additionally, the assumption of trait stability may have been untenable given that some medical conditions and pharmacological interventions might have influenced cognitive development following the initial assessment. However, research with a prior version of the WISC demonstrated that medication did not significantly impact IQ scores (Schwean & McCrimmon, 2008).

The magnitude of this threat was also mitigated by a review of the stability coefficients for those participants with medical concerns versus those with mental health concerns: none of the correlations were statistically different ($p<.01$) between these groups. A comparison of test-retest difference scores produced similar results: most differed by one point or less from the values reported in Table 1 with the exception of the FRI that was almost three points lower for the participants with medical concerns. When participants with ADHD were compared to participants without ADHD, stability coefficients and mean differences were not statistically significant at $p < .01$. Additionally, the current results are consistent with prior research on several versions of the WISC (Bartoi et al., 2015; Canivez & Watkins, 1998; Farmer & Kim, 2020; Kieng et al., 2015; Lander, 2010; Ryan et al., 2010, 2013; Watkins & Canivez, 2004; Watkins & Smith, 2013), with theory regarding the reliability of person-relative scores (McDermott et al., 1992; McGill, 2018; McGill et al., 2018; Styck et al., 2019), with studies on the treatment or intervention validity of cognitive test scores (Braden & Shaw, 2009; Burns et al., 2016; Elliott & Resing, 2015; Floyd & Kranzler, 2019; Owen et al., 2010; Stuebing et al., 2015; Watkins & Glutting, 2000), and with the results of structural validity studies (Canivez et al., 2020; Canivez & Watkins, 2016; Dombrowski et al., 2018, 2019).

Given that the WISC-V was developed for individual administration and is used to make high-stakes decisions about individuals, its internal consistency and short-term test-retest reliability should exceed 0.90 (Wasserman & Bracken, 2013). Among the 15 possible WISC-V scores, this dual standard was met by only the VCI and FSIQ scores within the normative sample (Wechsler, 2014b). The current study found that only the VCI, VSI, and FSIQ scores exhibited long-term stability coefficients ≥0.80 and none of the idiographic scores were stable across time. Thus, only the VCI and FSIQ scores appear to possess sufficient reliability for clinical use. Validity studies have reported that the WISC-V factor index scores are conceptually complex and are not well-defined indicators of their underlying constructs (Watkins & Canivez, in press). Further, these factor index scores seem to add little value beyond the FSIQ score for interpretation or prediction of meaningful outcomes (Canivez et al., 2014, 2020; Canivez & Watkins, 2016; Dombrowski et al., 2018, 2019; Freeman & Chen, 2019; McDermott et al., 1992; Watkins & Canivez, in press; Watkins & Styck, 2017). Given this evidence, clinicians should be careful not to overinterpret WISC-V scores for both ethical (Weiner, 1989) and legal (Reynolds & Milam, 2012) reasons.

ORCID

Marley W. Watkins http://orcid.org/0000-0001-6352-7174
Gary L. Canivez http://orcid.org/0000-0002-5347-6534
Stefan C. Dombrowski http://orcid.org/0000-0002-8057-3751
Ryan J. McGill http://orcid.org/0000-0002-5138-0694
Calliope B. Holingue http://orcid.org/0000-0002-8190-2385
Lisa A. Jacobson http://orcid.org/0000-0002-6992-029X

References

American Educational Research Association (AERA), American Psychological Association (APA), & National Council on Measurement in Education (NCME). (2014). *Standards for educational and psychological testing*. AERA.

Bandalos, D. L. (2018). *Measurement theory and applications in the social sciences*. Guilford

Bartoi, M. G., Issner, J. B., Hetterscheidt, L., January, A. M., Kuentzel, J. G., & Barnett, D. (2015). Attention problems and stability of WISC-IV scores among clinically referred children. *Applied Neuropsychology-Child*, *4*(3), 133–140. https://doi.org/10.1080/21622965.2013.811075

Benson, N. F., Floyd, R. G., Kranzler, J. H., Eckert, T. L., Fefer, S. A., & Morgan, G. B. (2019). Test use and assessment practices of school psychologists in the United States: Findings from the 2017 national survey. *Journal of School Psychology*, *72*, 29–48. https://doi.org/10.1016/j.jsp.2018.12.004

Benson, N. F., Maki, K. E., Floyd, R. G., Eckert, T. L., Kranzler, J. H., & Fefer, S. A. (2020). A national survey of school psychologists' practices in identifying specific learning disabilities. *School Psychology*, *35*(2), 146–157. https://doi.org/10.1037/spq0000344

Borreca, C. P., Cheramie, G. M., & Borreca, E. A. (2013). Legal issues in educational testing. In K. F. Geisinger (Ed.), *APA handbook of testing and assessment in psychology: Testing and assessment in school psychology and education* (Vol. 3, pp. 517–542). American Psychological Association.

Bowden, S. C., & Finch, S. (2017). When is a test reliable enough and why does it matter? In S. C. Bowden (Ed.), *Neuropsychological assessment in the age of evidence-based practice: Diagnostic and treatment evaluations* (pp. 95–119). Oxford University Press.

Braden, J. P., & Shaw, S. R. (2009). Intervention validity of cognitive assessment: Knowns, unknowables, and unknowns. *Assessment for Effective Intervention*, *34*(2), 106–115. https://doi.org/10.1177/1534508407313013

Bronfenbrenner, U., & Morris, P. A. (2006). The bioecological model of human development. In R. M. Lerner (Ed.), *Handbook of child psychology: Theoretical models of human development* (6th ed., Vol. 1, pp. 793–828). Wiley.

Burns, M. K., Petersen-Brown, S., Haegele, K., Rodriguez, M., Schmitt, B., Cooper, M., Clayton, K., Hutcheson, S., Conner, C., Hosp, J., & VanDerHeyden, A. M. (2016). Meta-analysis of academic interventions derived from neuropsychological data. *School Psychology Quarterly: The Official Journal of the Division of School Psychology, American Psychological Association*, *31*(1), 28–42. https://doi.org/10.1037/spq0000117

Canivez, G. L., McGill, R. J., Dombrowski, S. C., Watkins, M. W., Pritchard, A. E., & Jacobson, L. A. (2020). Construct validity of the WISC-V in clinical cases: Exploratory and confirmatory factor analyses of the 10 primary subtests. *Assessment*, *27*(2), 274–296. https://doi.org/10.1177/1073191118811609

Canivez, G. L., & Watkins, M. W. (1998). Long-term stability of the Wechsler Intelligence Scale for Children—Third Edition. *Psychological Assessment*, *10*(3), 285–291. https://doi.org/10.1037/1040-3590.10.3.285

Canivez, G. L., & Watkins, M. W. (2016). Review of the Wechsler Intelligence Scale for Children–Fifth Edition: Critique, commentary, and independent analyses. In A. S. Kaufman, S. E. Railford, & D. L. Coalson (Eds.), *Intelligent testing with the WISC-V* (pp. 683–702). Wiley.

Canivez, G. L., Watkins, M. W., James, T., Good, R., & James, K. (2014). Incremental validity of WISC-IV(UK) factor index scores with a referred Irish sample: Predicting performance on the WIAT-II(UK.). *The British Journal of Educational Psychology*, *84*(Pt 4), 667–684. https://doi.org/10.1111/bjep.12056

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, *20*(1), 37–46. https://doi.org/10.1177/001316446002000104

Courville, T., Coalson, D. L., Kaufman, A. S., & Raiford, S. E. (2016). Does WISC-V scatter matter? In A. S. Kaufman, S. E. Raiford, & D. L. Coalson (Eds.), *Intelligent testing with the WISC-V* (pp. 209–225). Wiley.

Cronbach, L. J., & Snow, R. E. (1977). *Aptitudes and instructional methods: A handbook for research on interactions*. Irvington Publishers

Dombrowski, S. C., Canivez, G. L., & Watkins, M. W. (2018). Factor structure of the 10 WISC-V primary subtests across four standardization age groups. *Contemporary School Psychology*, *22*(1), 90–104. https://doi.org/10.1007/s40688-017-0125-2

Dombrowski, S. C., McGill, R. J., & Morgan, G. B. (2019). Monte Carlo modeling of contemporary intelligence test (IQ) factor structure: Implications for IQ assessment, interpretation, and theory. *Assessment*. https://doi.org/10.1177/1073191119869828

Elliott, J. G., & Resing, C. M. (2015). Can intelligence testing inform educational intervention for children with reading disability? *Journal of Intelligence*, *3*(4), 137–157. https://doi.org/10.3390/jintelligence3040137

Farmer, R. L., & Kim, S. Y. (2020). Difference score reliabilities with the RIAS-2 and WISC-V. *Psychology in the Schools*, *57*(8), 1273–1288. https://doi.org/10.1002/pits.22369

Floyd, R. G., & Kranzler, J. H. (2019). Remediating student learning problems: Aptitude-by-treatment interaction versus skill-by-treatment interaction. In M. K. Burns (Ed.), *Introduction to school psychology: Controversies and current practice*. Oxford University Press.

Freeman, A. J., & Chen, Y.-L. (2019). Interpreting pediatric intelligence tests: A framework from evidence-based medicine. In G. Goldstein, D. N. Allen, & J. DeLuca (Eds.), *Handbook of psychological assessment* (4th ed., pp. 65–101). Academic Press.

Groth-Marnat, G., & Wright, A. J. (2016). *Handbook of psychological assessment* (6th ed.). Wiley.

Holm, S. (1979). A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, *6*(2), 65–70.

Hunt, E. (2011). *Human intelligence*. Cambridge University Press.

Kaufman, A. S., Raiford, S. E., & Coalson, D. L. (2016). *Intelligent testing with the WISC-V*. Wiley.

Kieng, S., Rossier, J., Favez, N., Geistlich, S., & Lecerf, T. (2015). Long-term stability of the French WISC-IV index scores: Personal strengths and personal weaknesses. *Pratiques psychologiques*, *21*(2), 137–154. https://doi.org/10.1016/j.prps.2015.03.002

Lander, J. (2010). Fairleigh Dickinson University. *Long-term stability of scores on the Wechsler Intelligence Scale for Children–Fourth Edition in children with learning disabilities* [Unpublished doctoral dissertation]. Fairleigh Dickinson University.

Mackintosh, N. J. (2011). *IQ and human intelligence* (2nd ed.). Oxford University Press.

McDermott, P. A., Fantuzzo, J. W., Glutting, J. J., Watkins, M. W., & Baggaley, A. R. (1992). Illusions of meaning in the ipsative assessment of children's ability. *The Journal of Special Education*, *25*(4), 504–526. https://doi.org/10.1177/002246699202500407

McDermott, P. A., Watkins, M. W., & Rhoad, A. M. (2014). Whose IQ is it?-Assessor bias variance in high-stakes psychological assessment. *Psychological Assessment*, *26*(1), 207–214. https://doi.org/10.1037/a0034832

McGill, R. J. (2018). Confronting the base rate problem: More ups and downs for cognitive scatter analysis. *Contemporary School Psychology*, *22*(3), 384–392. https://doi.org/10.1007/s40688-017-0168-4

McGill, R. J., Dombrowski, S. C., & Canivez, G. L. (2018). Cognitive profile analysis in school psychology: History, issues, and continued concerns. *Journal of School Psychology*, *71*, 108–121. https://doi.org/10.1016/j.jsp.2018.10.007

Miller, J. L., Saklofske, D. H., Weiss, L. G., Drozdick, L., Llorente, A. M., Holdnack, J. A., & Prifitera, A. (2016). Issues related to the WISC-V assessment of cognitive functioning in clinical and special groups. In L. G. Weiss, D. H. Saklofske, J. A. Holdnack, & A. Prifitera (Eds.), *WISC-V assessment and interpretation: Scientist-practitioner perspectives* (pp. 287–343). Academic Press.

Owen, A. M., Hampshire, A., Grahn, J. A., Stenton, R., Dajani, S., Burns, A. S., Howard, R. J., & Ballard, C. G. (2010). Putting brain training to the test. *Nature*, *465*(7299), 775–779. https://doi.org/10.1038/nature09042

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). Oxford University Press.

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*(1), 68–72. https://doi:10.1080/09084280903297933.

Ryan, J. J., Umfleet, L. G., & Kane, A. (2013). Stability of WISC-IV process scores. *Applied Neuropsychology. Child*, *2*(1), 43–46. https://doi.org/10.1080/21622965.2012.670554

Sattler, J. M., Dumont, R., & Coalson, D. L. (2016). *Assessment of children: WISC-V and WPPSI-IV*. Jerome M. Sattler Publisher.

Schuerger, J. M., & Witt, A. C. (1989). The temporal stability of individually tested intelligence. *Journal of Clinical Psychology*, *45*(2), 294–302. https://doi.org/10.1002/1097-4679(198903)45:2%3C294::AID-JCLP2270450218%3E3.0.CO;2-N

Schwean, V. L., & McCrimmon, A. (2008). Cultural issues in clinical use of the WISC-IV. In A. Prifitera, D. H. Saklofske, & L. G. Weiss (Eds.), *WISC-IV clinical assessment and intervention* (2nd ed., pp. 193–215). Academic Press.

Stuebing, K. K., Barth, A. E., Trahan, L. H., Reddy, R. R., Miciak, J., & Fletcher, J. M. (2015). Are child cognitive characteristics strong predictors of responses to intervention? A meta-analysis. *Review of Educational Research*, *85*(3), 395–429. https://doi.org/10.3102/0034654314555996

Styck, K. M., Beaujean, A. A., & Watkins, M. W. (2019). Profile reliability of cognitive ability subscores in a referred sample. *Archives of Scientific Psychology*, *7*(1), 119–128. https://doi.org/10.1037/arc0000064

Styck, K. M., & Walsh, S. M. (2016). Evaluating the prevalence and impact of examiner errors on the Wechsler Scales of Intelligence: A meta-analysis. *Psychological Assessment*, *28*(1), 3–17. https://doi.org/10.1037/pas0000157

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

Wasserman, J. D., & Bracken, B. A. (2013). Fundamental psychometric considerations in assessment. In J. R. Graham & J. A. Naglieri (Eds.), *Handbook of psychology: Assessment psychology* (Vol. 10, pp. 50–81). Wiley.

Watkins, M. W. (2005). Diagnostic validity of Wechsler subtest scatter. *Learning Disabilities: A Contemporary Journal*, *3*(2), 20–29.

Watkins, M. W., & Canivez, G. L. (2004). Temporal stability of WISC-III subtest composite: Strengths and weaknesses. *Psychological Assessment*, *16*(2), 133–138. https://doi.org/10.1037/1040-3590.16.2.133

Watkins, M. W., & Canivez, G. L. (in press). Assessing the psychometric utility of IQ scores: A tutorial using the Wechsler Intelligence Scale for Children–Fifth Edition. *School Psychology Review*. https://doi.org/10.1080/2372966X.2020.1816804

Watkins, M. W., & Glutting, J. J. (2000). Incremental validity of WISC-III profile elevation, scatter, and shape information for predicting reading and math achievement. *Psychological Assessment*, *12*(4), 402–408. https://doi.org/10.1037/1040-3590.12.4.402

Watkins, M. W., & Smith, L. G. (2013). Long-term stability of the Wechsler Intelligence Scale for Children-Fourth Edition. *Psychological Assessment*, *25*(2), 477–483. https://doi.org/10.1037/a0031653

Watkins, M. W., & Styck, K. M. (2017). A cross-lagged panel analysis of psychometric intelligence and achievement in reading and math. *Journal of Intelligence*, *5*(3), 31–11. https://doi.org/10.3390/jintelligence5030031

Wechsler, D. (2014a). *Wechsler Intelligence Scale for Children* (5th ed.). NCS Pearson.

Wechsler, D. (2014b). *Wechsler Intelligence Scale for Children-Fifth Edition technical and interpretive manual*. NCS Pearson.

Weiner, I. B. (1989). On competence and ethicality in psychodiagnostic assessment. *Journal of Personality Assessment*, *53*(4), 827–831. https://doi.org/10.1207/s15327752jpa5304_18