DOI: bpa.2022.00004

Threshold values for significant changes in test-retest difference scores for the Wechsler Intelligence Scale for Children – Fourth Edition

Lina Pezzuti[1], James Dawe[1], Marco Lauriola[2]

[1] Department of Dynamic and Clinical Psychology, and Health Studies, Sapienza University of Rome, Italy
[2] Department of Social and Developmental Psychology, Sapienza University of Rome, Italy

lina.pezzuti@uniroma1.it

ABSTRACT. La presente ricerca si proponeva di studiare l'effetto della pratica a breve termine della *Wechsler Intelligence Scales for Children - Quarta Edizione (WISC-IV)* e di fornire valori soglia che consentano ai professionisti di valutare se esistono vere differenze nelle prestazioni individuali o se tali differenze siano dovute al caso. A un campione di 440 soggetti è stata somministrata la WISC-IV due volte con un intervallo medio di 30 giorni. I risultati mostrano che la pratica è più pronunciata quando si utilizzano i punteggi grezzi dei subtest rispetto a quelli ponderati. Sono stati ricavati dei valori soglia per valutare i cambiamenti significativi nei subtest e negli indici, consentendo così ai professionisti di valutare con maggiore precisione il significato clinico dei cambiamenti osservati durante una doppia somministrazione a breve termine della WISC-IV.

SUMMARY. One of the purposes of administering intelligence scales is to assess changes in cognitive functioning over time, from a few days to several years, to determine whether the examinee has progressed or regressed after treatment or other events. (e.g., an accident, a rehabilitation, etc.). The present research aimed to study the short-term practice effect of the Wechsler Intelligence Scales for Children – Fourth Edition and provide threshold values that allow practitioners to assess whether there are true differences in individual performance or whether these differences are due to chance. A sample of 440 subjects was administered the WISC-IV twice with an average interval of 30 days. The results show that practice is more pronounced when using raw subtest scores than when using weighted scores. Threshold values for assessing significant change in subtests and indices were obtained. For example, for the Full-Scale Intelligence Quotient, a difference between 6 and 27 IQ points between the first and second administration indicates a practice effect. Conversely, if the difference is equal to or less than 5 IQ points, then there was a decline, while if it is equal to or greater than 28 IQ points, there was an increase in performance not due to the practice effect. Therefore, these data should allow practitioners to more accurately assess the clinical significance of observed changes during a short-term dual administration.

Keywords: *WISC-IV, Test-retest, Individual changes*

INTRODUCTION

The first aim of administering one of the Wechsler intelligence scales (WPPSI-IV, WISC-IV, or WAIS-IV) is to assess an individual's cognitive functioning based on observed performance. A second aim is to evaluate change in cognitive functioning over time, from a few days (short term) to several years (long term), to determine whether the examinee has progressed or regressed after treatment or other events (e.g., an accident, a rehabilitation, etc.).

Intelligence is a psychological construct presumed to be relatively stable; thus, intelligence tests must produce similar scores from one time to another (Canivez & Watkins, 1998; Conley, 1984; Deary, Pattie & Starr, 2013; Heilbronner et al., 2010; Hunt, 2010; Mackintosh, 1998; Moffitt, Caspi, Hakness & Silva, 1993; Reeve & Bonaccio, 2011; Revelle, 2010; Simonton, 2011; Strauss, Sherman & Spreen, 2006; Wright, 2011). Accordingly, intelligence test scores were found to be relatively stable from childhood through adulthood (Chen & Siegler, 2000; Deary, Whalley, Lemmon, Crawford & Starr, 2000; Johnson, Gow, Corley, Starr & Deary, 2010) for both average and above-average samples (Reeve & Bonaccio, 2011; Simonton, 2011).

The information concerning the short-term test-retest stability of the *Wechsler Intelligence Scale for Children (WISC*; Wechsler, 1949), *WISC-Revised* (*WISC-R*; Wechsler, 1974), *WISC-Third Edition* (*WISC-III*; Wechsler, 1991), and *WISC-Fourth Edition* (*WISC-IV*; Wechsler, 2003a; 2003b) is available in their respective test manuals and has typically been conducted with nondisabled youths across retest intervals of fewer than 3 months. Likewise, there is extensive literature dealing with the stability of these WISC editions for a variety of test-retest intervals using healthy children, gifted children, and those with learning disabilities, attention deficit disorder, mental retardation, or other handicapping conditions (e.g., Anderson, Cronin & Kazmierski, 1989; Bauman, 1991; Canivez & Watkins, 1998, 1999, 2001; Ellzey & Karnes, 1990; Truscott, Narrett & Smith, 1994).

Most of these studies have indicated significant increases in verbal intelligence quotient (VIQ), performance intelligence quotient (PIQ), and full scale intelligence quotient (FSIQ) scores, with the largest increases in PIQ during short test-retest intervals (e.g., 30 days). The practice effects tend to disappear with longer retest intervals (e.g., 1-3 years). Canivez and Watkins (1999) concluded that the FSIQ of the WISC-III is the only score that possesses sufficient stability for interpreting individual cases.

The *Wechsler Intelligence Scale for Children – Fourth Edition* (*WISC-IV*; Wechsler 2003a; 2003b; Wechsler, 2012) is currently used in Italy for clinical practice with children and adolescents. Because approximately 60% of the items in its core subtests are new or revised (Watkins, 2010), the internal consistency and test-retest reliability of the WISC-IV cannot be assumed to be equivalent to previous editions and must be studied again.

The WISC-IV includes 10 core subtests (Bock design, Similarities, Digit span, Picture concepts, Coding, Vocabulary, Letter-number sequencing, Matrix reasoning, Comprehension and symbol search) and 5 supplemental subtests (Picture completion, Cancellation, Information, Arithmetic, and Word reasoning). The interpretation of the WISC-IV is mainly based on the full-scale intelligence quotient (FSIQ) and four index scores: the verbal comprehension index (VCI); the perceptual reasoning index (PRI); the working memory index (WMI); and the processing speed index (PSI). Two other indices, the general ability index (GAI) and the cognitive proficiency index (CPI), can be derived.

The stability of the WISC-IV scores across time has been investigated in several studies, mainly with long time intervals. The only short-term reliability study dates back to the USA standardization of the WISC-IV (Wechsler, 2003b), in which 243 children (52.3% female and 47.7% male) were tested twice, with a time interval ranging from 13 to 63 days (Mean = 32 days). Observing the results, the stability coefficients were satisfactory for the indices (from .80 to .90). Still, the short-term practice effects were observed with gains ranging from 2.1 points for the verbal comprehension index, to 7.1 points for processing speed index. On average, the increase was 5.6 points for the FSIQ.

All the other studies were carried out with long time intervals for the second administration of WISC-IV and are briefly described in Table 1. Such studies showed that: subtest long-term stability coefficients were consistently lower than the short-term stability coefficients reported for the normative samples; the long test-retest reliability coefficients for the subtests were generally lower than the scores for the four indices and the FSIQ; the FSIQ exhibited a higher long-term stability coefficient respect four WISC-IV indices. Some studies showed differences of 1-2 points in the subtest scores and up to 9-10 points in the four indices and FSIQ in a high

Table 1 – Summary of the long test-retest studies by the WISC-IV

Authors	**Mean time interval**	**Sample and age at first administering**	**Results:** *test-retest stability coefficients*	**Results:** *test-retest mean differences*
Ryan, Glass & Bartels, 2010	11 months	43 elementary and middle school children (mean age = 7.77 years, *SD* = 1.91)	Ranged from .54 for the PSI to .88 for the FSIQ.	42% of the FSIQ scores increased by 5 or more points on retest.
Lander, 2010	36 months	131 students with a learning disability	Ranged from .52 for the PSI to .65 for the FSIQ.	
Watkins & Smith, 2013	34 months	344 students aged 6,1 to 14,3 years	They were .72, .76, .66, .65, and .82 for the VCI, PRI, WMI, PSI, and FSIQ respectively.	The subtest scores did not differ by more than 1 point, and the index scores did not differ by more than 2 points. 44% of the students' VCI, PRI, WMI, and PSI scores increased by 10 or more points.
Bartoi, Issner, Hetterscheidt, January, Kuentzel & Barnett, 2015	22 months	51 children aged 8 to 16 years	Ranged from .58 for the PSI to .86 for the FSIQ.	78.4% of the children had test-retest differences up to 9 points for the FSIQ; similarly, 68.6%, 56.9%, 54.9%, and 54.9% of the children increased up to 9 points for the VCI, PRI, WMI, and PSI, respectively.
Kieng, Kieng & Geistlich, 2017	21 months	277 children aged 7 to 12 years	Ranged from .63 for the WMI to .80 for the FSIQ.	Half of the subjects shift from one intelligence classification category to the higher category.
Okada, Kawasaki, Shinomiya, Hoshino, Ino, Sakai, ... & Niwa, 2021	31 months	138 children with autism spectrum disorder (aged 5,5 to 16,8 years)	.83 for FSIQ, ranged from .62 to .79 for the four WISC-IV indices.	The mean of the FSIQ and VCI scores increased by 3.4 and 4.6 points in the second test.

percentages of the subjects tested. Some authors (Ryan, Glass & Bartels, 2010; Watkins & Smith, 2013) concluded that given this variability, it could not be assumed that the WISC-IV scores are consistent across long test-retest intervals.

In summary, most research has focused on the study of long-term stability, while there are no short-term reliability studies except for the US WISC-IV standardization data. Therefore, assessing the clinical significance of changes in retest performance must be carefully considered for short intervals. The test-retest procedure (from a few weeks to several years) can be used to address this. Indeed, numerous studies have shown that performance on a second test is superior to performance on the first test (Estevis, Basso & Combs, 2012; Salthouse, 2014; Sherman, Brooks, Iverson,

Slick & Strauss, 2011). However, to determine whether the test-retest difference score represents a significant change, threshold values must be obtained with correction for the effect of practice, measurement error, and regression to the mean, as done in a paper by Lecerf, Kieng and Geistlich (2017) on the WISC-IV.

If psychologists had instruments with perfect reliability, the performance observed on the test should be the same as that obtained on their retest. Nevertheless, no score of psychological measures has perfect reliability. Every test has a bias that makes it difficult to interpret the differences in observed scores between test and retest only in terms of cognitive functioning change. Therefore, determining if there has been a true change involves taking into account various psychometric phenomena, such as measurement error (i.e., the source of inaccuracies in test scores), and providing information about the reliability of test scores and practice effects, which reflect changes associated with repeated test administration.

It is worth noting that assessing change requires distinguishing statistical significance from the clinical relevance of a test-retest difference score (Brooks, Sherman, Iverson, Slick & Strauss, 2011; Jacobson & Truax, 1991). The 5% threshold ($p<.05$; for discussion Cohen, 1994; Reuchlin, 1992) is regularly used in psychology to determine whether a difference is statistically significant. However, statistically significant differences in intelligence tests do not mean clinically significant ones. Differences are considered of clinical interest if they are rare in the population. For some authors, this corresponds to a difference observed in less than 5% of the population (Chelune, 2003), for others in less than 10% (Kaufman & Kaufman, 2008), still others in less than 15% (Sattler, 2008). In this paper, we will use the 10% threshold.

As stated above, intelligence is a psychological construct that is assumed to be stable; so intelligence tests should produce similar scores from one time to the other. However, test-retest stability is not only characteristic of the test but may vary depending on: the type of the sample assessed (e.g., clinical or healthy); the size of the sample, i.e., Charter (2003) recommends a minimum of 400 participants for test-retest studies and Watson (2004) suggested a sample of at least 300, and possibly 400; the time interval between test and retest (short or long interval); the statistical methods used to calculate the reliability of the test-retest (e.g., Pearson correlation or intraclass correlations) and the practical effects (e.g., Cohen's *d*, Anova, or reliable change index). Finally, the type of scores examined influences the test-retest results, e.g., in the study by Lemay and colleagues (Lemay, Bedard, Rouleau& Tremblay, 2004), the reliabilities of the raw score on the Wechsler's scales were higher than the reliabilities assessed by demographically adjusted scores (i.e., scaled or weighted scores). Although the conversion from raw score to scaled score may be helpful to compare results gathered from different age groups, it also tends to induce a greater variability in the performance, limiting its usefulness in repeated administration, leading to a slight decrease in reliability (Lemay et al. 2004).

In the present paper, we studied the practice effect of repeated administration of the WISC-IV at a short time interval. We proposed threshold values to assess whether there are true differences in individual performance or whether these differences are due to chance. These threshold values are estimated using a sample of 440 subjects taking the WISC-IV twice. The threshold values reported here consider the effects of practice and measurement error. Therefore, these data allow practitioners to assess the clinical significance of observed changes more accurately during a short-term dual administration.

METHOD

Participants

We recruited 440 children and adolescents (219 girls and 221 boys) equally distribuited into the 11 age groups from 6 to 16 years of age reported in the WISC-IV manual. For each participant, biographical data and the education level of the parents were recorded according to four categories: primary, secondary, high school and university degree. As was done for the WISC-IV Italian standardization, the national data with which to compare the data of the present research were derived from estimates taken from a representative sample of the Italian population carried out on a sample of 7977 households with a total of 19907 individuals (Census of Italy Bank, 2010). The distribution of the individuals by age, gender, and parents' education is comparable to those of the above-mentioned survey. All of the children and adolescents examined had no previous psychological diagnosis and assessment, nor were they undergoing psychological treatment of any kind.

Instrument

The *Wechsler Intelligence Scale for Children – Fourth Edition* (*WISC-IV*; Wechsler 2003a; 2003b) is one of the most frequently used tests to assess the general intellectual functioning of Italian-speaking children. The WISC-IV is an individually administered test of intelligence for children ages 6 years 0 months through 16 years 11 months. It includes 10 core subtests (Bock design, Similarities, Digit span, Picture concepts, Coding, Vocabulary, Letter-number sequencing, Matrix reasoning, Comprehension and Symbol search) and 5 supplemental subtests (Picture completion, Cancellation, Information, Arithmetic, and Word reasoning). Each subtest has a standardized mean of 10 and a standard deviation of 3. The Italian WISC-IV was standardized on a nationally representative sample (N = 2200), closely approximating the 2010 database from the Census of Italy Bank (2010) on gender, parents' socioeconomic status, and parents' professional class. Currently, interpretation of the WISC-IV is mainly based on the full-scale intelligence quotient (FSIQ) and four index scores. The verbal comprehension index (VCI) is derived from the sum of Similarities, Vocabulary, and Comprehension scores; the perceptual reasoning index (PRI) from the sum of Block design, Picture concepts, and Matrix reasoning scores; the working memory index (WMI) from the sum of Digit span, and Letter-number sequencing scores; and the processing speed index (PSI) from the sum of Coding, and Symbol search scores. Two other Indices as the general ability index (GAI: the subtests are those of the VCI and PRI indices) and the cognitive proficiency index (CPI: the subtests are those of the WMI and PSI indices), can be derived. Finally, the FSIQ is obtained by adding the ten core subtest scores.

To interpret the outcomes at the WISC-IV, the raw scores of the subtests are converted to age-weighted scores, and the sum of the age-weighted scores of the subtests belonging to the indices is converted to the standard point IQ.

The reliability study of the Italian edition of the WISC-IV was mainly conducted through the split-half method, which is helpful in studying the homogeneity of the items composing the subtest. On the contrary, the reliability coefficients for 3 of the 15 subtests (Coding, Symbol search, and Deletion) and 2 of the 7 process scores (random deletion strategy and structured deletion strategy) were calculated using test-retest method. The reliability of the composite scores was instead studied using Mosier's (1943) formula.

The Italian WISC-IV standardization manual (Wechsler, 2012) reported average reliability indices varying between .74 for Symbol search and to .90 for Vocabulary, similar to the US edition values, which ranged between .79 for Symbol search and Cancellation to .90 for Letter-number sequencing. The average reliability of the four indices varied between .84 for the processing speed index to .94 for the verbal comprehension index in the Italian standardization sample; similarly, in the US edition, it varies between .88 for the processing speed index to .94 for the verbal comprehension index. Further, the average reliability of IQ was .96 in the Italian standardization and .97 in the US edition.

Procedure

Research participants were approached via primary and secondary schools in central Italy. Informed consent was requested from both parents and also information regarding any previous diagnoses and/or psychological treatment given to their children.

Participants were individually administered all 15 subtests of the Italian WISC-IV version twice, with test-retest intervals ranging from 17 to 38 days (M = 30; SD = 2.8 days).

Data analysis

- *Practice effect of repeated administration with the WISC-IV.* Cohen's d is used to estimate the effect size of practice effects due to repeated administrations for raw and weighted scores, for subtests and total scores, and for indices and FSIQ. An effect size ≥.80 is considered as a large practice effect; .50-.79, medium; .20-.49, small; and <.20, trivial (Cohen, 1988).
- *Reliable change index and threshold values for detecting decline or progression at the second evaluation*. Several statistical procedures exist to assess the significance of changes in test-retest performance and to account for bias and error (Basso, Carona, Lowery & Axelrod, 2002; Brooks, Strauss, Sherman, Iverson & Slick, 2009; Estevis et al., 2012). In this paper, we use the method initially proposed by Jacobson and colleagues (Jacobson, Follette & Revenstrof, 1984; Jacobson & Truax, 1991) and revised by Chelune and colleagues (Chelune, Naugle, Lüders, Sedlak & Awad, 1993), which results in the calculation

of a reliable change index (RCI). This method provides threshold values determining the magnitude of test-retest changes required for significant differences (≤10%). The RCI is calculated using simple descriptive statistics: mean and standard deviation values of the test and retest scores and test-retest correlations (i.e., a reliability coefficient). For each individual, the test-retest difference score is calculated. If the test-retest difference is greater than the RCI, it is considered a significant difference, rarely observed in the population.

In calculating the RCI, we use the standard error of a difference (SEM_{diff}) formula. According to Chelune et al. (1993), SEM_{diff} considers that measurement errors are unlikely to be the same at test and retest administrations (and therefore, reliability could not be the same) and that there might be an effect of practice. Thus, modifying the original formula of Jacobson and Truax (1991), who assumed that the measurement errors were the same at test and retest and that there was no effect of practice, the formula proposed by Chelune et al. (1993) is: SEM_{diff} = Squared root $[(SEM_{test})^2 + (SEM_{retest})^2]$.

In this formula, the SEM_{test} is the standard error of measurement at the first evaluation (test), and the SEM_{retest} is the standard error of measurement at the second evaluation (retest). The SEM_{diff} value is multiplied by 1.645 and 1.96 to obtain the 90% and 95% confidence intervals respectively. The SEM_{diff} value is 0 for a test with perfect reliability and stability; when reliability decreases, SEM_{diff} increases.

In a later step, Chelune et al. (1993) incorporated practice effects (provided by the mean difference between test and retest) to identify threshold values for significant decline or progression at 90% respectively through the formulae: (SEM_{diff}(90%) - Practice effect) and (SEM_{diff}(90%) + Practice effect).

In summary, this SEM_{diff} procedure corrects the distribution of observed change scores by firstly taking into account measurement error and secondly taking into account the effects of the practice.

RESULTS

Examining Table 2 about the subtests, reveals a more pronounced practice effect for the raw subtest scores (*Mean d* = –.61) than for the weighted scores M*ean d* = –.50). More specifically, a small practice effect emerged for Letter-number sequencing and Arithmetic subtests, and a large practice effect for Coding and Picture completion subtests. For all other subtests, the effect was medium.

Regarding the WISC-IV indices (see Table 3), the same effect was found for both the sum of weighted and standardized points (i.e., IQs), with the impact of practice ranging from .55 (medium) for the working memory index, to 1.39 (very large) for full-scale intelligence quotient.

Tables 4 and 5 show the SEM_{diff} values and thresholds at 90% and 95% confidence intervals for deciding whether a significant change in the direction of decline or progression has occurred at a second short-term administration.

DISCUSSION AND CONCLUSION

Few studies have investigated the short-term stability of one of the most frequently used tests in the field of cognitive administration: the WISC-IV.

Given the distortions in psychological tests, particularly in intelligence scales, the mere difference between the scores observed on the examination and the retest cannot provide information on a possible change in cognitive functioning. The present study shows that correcting the test-retest difference scores for the effects of practice and measurement error present in the two administrations is crucial.

Comment for the first results of the study on the effect of practice in a short-time test-retest WISC-IV with 440 children is that this effect is stronger when using the raw scores of the subtests, which corresponding to the sums of corrected items, with respect to weighted scores that are age-corrected scaled scores transformed using WISC-IV Italian norms; a result that confirms the research of Lemay et al. (2004). The explanation for this result lies in the continuity of the raw scores and the age-weighted scores classification: i.e., if a subject obtains a raw score of 8 on the first administration and a score of 10 on the second administration, the change is evident, but if the raw scores of 8 and 10 correspond to the same age-weighted score, no difference between test and retest will emerge. So, it may be more beneficial for the practitioners to assess changes in subtest by using the first type of scores (raw) than the second (weighted).

A second reflection is on the threshold values found in the present study, which are higher than those found in the research by Lecerf et al. (2017) and Kieng et al. (Kieng, Rossier, Favez, Geistlich & Lecerf, 2015) resulting from the

Table 2 – Mean and SD for the early and late WISC-IV administration, and tests of mean differences, using raw and standard scores for the subtest (n = 440)

Test score	Early administration		Late administration		Test of mean difference		
	M	*SD*	*M*	*SD*	*t*	Diff. (or practice effect)	Cohen's *d*
Subtest				*Raw scores*			
Block design	34.66	13.66	39.10	14.23	−14.19	−4.45	−.68
Similarities	20.60	7.60	22.35	7.73	−12.75	−1.76	−.61
Digit span	17.18	4.07	18.25	4.51	−10.89	−1.07	−.52
Picture concepts	16.07	4.04	17.58	4.03	−12.89	−1.51	−.61
Coding	52.57	14.85	58.56	15.86	−17.12	−5.99	−.82
Vocabulary	36.82	9.44	38.72	9.45	−11.51	−1.91	−.55
Letter–number sequencing	17.75	4.31	18.62	3.94	−8.92	−.88	−.43
Matrix reasoning	19.30	6.09	21.03	6.16	−13.07	−1.76	−.62
Comprehension	19.25	6.07	20.70	6.04	−12.05	−1.46	−.58
Symbol search	28.43	8.13	31.09	8.19	−11.34	−2.67	−.54
Picture completion	21.60	6.38	24.13	6.63	−19.10	−2.53	−.91
Cancellation	83.79	23.34	91.84	24.16	−14.44	−8.06	−.69
Information	18.27	4.83	19.22	5.11	−11.90	−.95	−.57
Arithmetic	22.84	5.57	23.64	5.34	−8.97	−.80	−.43
Word reasoning	13.63	3.46	14.67	3.60	−12.89	−1.04	−.61
Subtest				*Weigthed scores*			
Block design	10.71	3.01	11.88	2.84	−11.16	−1.17	−.53
Similarities	10.49	2.43	11.44	2.52	−11.38	−.95	−.54
Digit span	11.19	2.54	12.03	2.63	−9.99	−.84	−.48
Picture concepts	10.24	2.85	11.69	2.81	−12.29	−1.45	−.59
Coding	10.31	2.60	11.93	2.80	−15.55	−1.62	−.74
Vocabulary	10.53	2.31	11.18	2.37	−8.02	−.65	−.38
Letter–number sequencing	10.97	2.91	11.69	2.79	−6.94	−.73	−.33
Matrix reasoning	10.62	2.69	11.84	2.82	−11.99	−1.22	−.57
Comprehension	9.89	2.72	10.80	2.61	−10.06	−.91	−.48
Symbol search	11.18	3.26	12.54	3.25	−9.37	−1.36	−.45
Picture completion	10.10	2.95	11.64	2.93	−16.68	−1.54	−.79
Cancellation	10.89	2.75	12.50	4.35	−7.98	−1.61	−.38
Information	10.14	2.39	10.88	2.40	−8.84	−.73	−.42
Arithmetic	11.19	2.77	11.77	2.68	−6.65	−.58	−.32
Word reasoning	10.35	2.57	11.35	2.41	−11.10	−1.01	−.53

Note. All comparisons are significant at $p<.00001$.

Table 3 – Mean and SD for the early and late WISC-IV administration, and tests of mean differences, using the sum of weighted scores and standard scores (IQ) for indices (n = 440)

Test score	Early administration		Late administration		Test of mean difference		
	M	*SD*	*M*	*SD*	*t*	Diff. (or practice effect)	Cohen's *d*
Indices			*Sum of weighted scores*				
Verbal comprehension index	30.91	5.89	33.43	6.01	−14.86	−2.52	−.71
Perceptual reasoning index	31.57	6.14	35.40	6.31	−19.52	−3.83	−.93
Working memory index	22.15	4.72	23.72	4.69	−11.52	−1.57	−.55
Processing speed index	21.48	5.05	24.47	5.11	−15.52	−2.98	−.74
Full−scale IQ	106.11	14.56	117.41	15.18	−29.18	−11.30	−1.39
General ability index	62.42	9.83	68.83	10.25	−24.11	−6.41	−1.15
Cognitive proficiency index	43.58	7.35	48.12	7.53	−20.16	−4.54	−.96
Indices			*Standard scores (IQ)*				
Verbal comprehension index	101.81	11.80	106.85	12.02	−14.86	−5.04	−.71
Perceptual reasoning index	103.18	13.34	111.53	13.73	−19.54	−8.35	−.93
Working memory index	106.46	14.15	111.17	14.07	−11.52	−4.71	−.55
Processing speed index	104.36	14.94	113.13	14.97	−15.48	−8.77	−.74
Full−scale IQ	104.84	11.85	114.08	12.34	−29.36	−11.30	−1.40
General ability index	102.72	11.54	110.24	12.04	−24.11	−7.53	−1.15
Cognitive Proficiency Index	106.64	13.72	115.11	14.06	−20.16	−8.48	−.96

Note. All comparisons are significant at $p<.00001$.

Table 4 – Practice effects, SEM_{diff}, and RCI threshold values for significant test-retest differences for the raw and weighted WISC-IV subtest scores

WISC-IV subtest	Practice effect	SEM_{diff}	SEM_{diff} (90%)	RCI SEM_{diff} (90% decline)	RCI SEM_{diff} (90% progression)	RCI threshold (90% decline)	RCI threshold (90% progression)	SEM_{diff} (95%)	RCI SEM_{diff} (95% decline)	RCI SEM_{diff} (95% progression)	RCI threshold (95% decline)	RCI threshold (95% progression)
Subtest raw scores												
Block design	4.45	7.74	12.69	8.24	17.14	8	17	15.16	10.72	19.61	11	20
Similarities	1.76	3.34	5.47	3.72	7.23	4	7	6.54	4.78	8.30	5	8
Digit span	1.07	2.28	3.74	2.67	4.80	3	5	4.47	3.40	5.53	3	6
Picture concepts	1.51	2.79	4.58	3.07	6.09	3	6	5.47	3.96	6.99	4	7
Coding	5.99	9.12	14.95	8.97	20.94	9	21	17.87	11.88	23.86	12	24
Vocabulary	1.91	3.92	6.43	4.52	8.34	5	8	7.68	5.78	9.59	6	10
Letter-number sequencing	.88	2.21	3.62	2.74	4.49	3	4	4.33	3.45	5.20	3	5
Matrix reasoning	1.76	3.26	5.34	3.58	7.10	4	7	6.39	4.63	8.14	5	8
Comprehension	1.46	2.89	4.74	3.28	6.21	3	6	5.67	4.21	7.14	4	7
Symbol search	2.67	5.46	8.95	6.28	11.61	6	12	10.70	8.03	13.36	8	13
Picture completion	2.53	3.62	5.94	3.41	8.47	3	8	7.10	4.57	9.63	5	10
Cancellation	8.06	13.80	22.64	14.58	30.70	15	31	27.06	19.00	35.11	19	35
Information	.95	1.90	3.12	2.17	4.07	2	4	3.73	2.78	4.67	3	5
Arithmetic	.80	2.03	3.33	2.53	4.14	3	4	3.98	3.18	4.79	3	5
Word reasoning	1.04	1.94	3.18	2.14	4.22	2	4	3.80	2.76	4.84	3	5

continued on next page

continued

WISC-IV subtest	Practice effect	SEM_{diff}	SEM_{diff} (90%)	RCI SEM_{diff} (90% decline)	RCI SEM_{diff} (90% progression)	RCI threshold (90% decline)	RCI threshold (90% progression)	SEM_{diff} (95%)	RCI SEM_{diff} (95% decline)	RCI SEM_{diff} (95% progression)	RCI threshold (95% decline)	RCI threshold (95% progression)
Subtest weighted scores												
Block design	1.17	2.39	3.92	2.75	5.09	3	5	4.69	3.52	5.85	4	6
Similarities	.95	1.93	3.16	2.21	4.12	2	4	3.78	2.83	4.74	3	5
Digit span	.84	1.91	3.13	2.29	3.97	2	4	3.74	2.90	4.59	3	5
Picture concepts	1.45	2.69	4.41	2.97	5.86	3	6	5.28	3.83	6.72	4	7
Coding	1.62	2.50	4.10	2.48	5.72	2	6	4.90	3.28	6.52	3	7
Vocabulary	.65	1.79	2.94	2.28	3.59	2	4	3.51	2.86	4.16	3	4
Letter-number sequencing	.73	2.28	3.74	3.01	4.46	3	4	4.46	3.74	5.19	4	5
Matrix reasoning	1.22	2.34	3.84	2.62	5.06	3	5	4.59	3.37	5.81	3	6
Comprehension	.91	2.05	3.36	2.45	4.27	2	4	4.01	3.10	4.92	3	5
Symbol search	1.36	3.20	5.25	3.88	6.61	4	7	6.27	4.91	7.63	5	8
Picture completion	1.54	2.32	3.81	2.27	5.35	2	5	4.55	3.01	6.09	3	6
Cancellation	1.61	4.22	6.92	5.31	8.53	5	9	8.27	6.66	9.88	7	10
Information	.73	1.85	3.03	2.29	3.76	2	4	3.62	2.88	4.35	3	4
Arithmetic	.58	1.89	3.11	2.53	3.68	3	4	3.71	3.13	4.29	3	4
Word reasoning	1.01	2.06	3.38	2.38	4.39	2	4	4.04	3.04	5.05	3	5

Table 5 – Practice effects, SEM_{diff}, and RCI threshold values for significant test-retest differences for raw (sum of weighted scores) and standard (IQs) WISC-IV indices

WISC-IV indices	Practice effect	SEM_{diff}	SEM_{diff} (90%)	RCI SEM_{diff} (90% decline)	RCI SEM_{diff} (90% progression)	RCI threshold (90% decline)	RCI threshold (90% progression)	SEM_{diff} (95%)	RCI SEM_{diff} (95% decline)	RCI SEM_{diff} (95% progression)	RCI threshold (95% decline)	RCI threshold (95% progression)
Sum of weighted scores												
Verbal Comprehension index	2.52	4.17	6.84	4.32	9.36	4	9	8.18	5.66	10.69	6	11
Perceptual reasoning index	3.83	5.16	8.46	4.62	12.29	5	12	10.11	6.28	13.94	6	14
Working memory index	1.57	3.17	5.20	3.63	6.77	4	7	6.22	4.65	7.79	5	8
Processing speed index	2.98	4.63	7.59	4.61	10.57	5	11	9.07	6.09	12.05	6	12
Full-scale IQ	11.30	12.25	20.09	8.79	31.39	9	31	24.01	12.71	35.31	13	35
General ability index	6.41	7.74	12.69	6.28	19.09	6	19	15.16	8.76	21.57	9	22
Cognitive Competency index	4.54	6.01	9.86	5.32	14.40	5	14	11.78	7.24	16.32	7	16
Standard scores (IQ)												
Verbal comprehension index	5.04	8.35	13.69	8.65	18.74	9	19	16.28	11.24	21.32	11	21
Perceptual reasoning index	8.35	11.23	18.41	10.06	26.76	10	27	21.89	13.54	30.24	14	30
Working memory index	4.71	9.52	15.61	10.90	20.32	11	20	18.56	13.85	23.27	14	23
Processing speed index	8.77	13.63	22.36	13.59	31.13	14	31	26.59	17.82	35.35	18	35
Full-scale IQ	11.30	9.98	16.37	5.08	27.67	5	28	19.47	8.17	30.76	8	31
General ability index	7.53	9.09	14.90	7.38	22.43	7	22	17.72	10.19	25.24	10	25
Cognitive competency index	8.48	11.22	18.41	9.93	26.88	10	27	21.89	13.41	30.36	13	30

long-term (more than 1 year) WISC-IV test-retest studies, where the practice effects are lower.

In particular, Tables 4 and 5 show the SEM_{diff} values and thresholds for deciding whether a significant change in the direction of decline or progression has occurred at a second short-term administration. For example, if we administer the Block design subtest twice, a raw score difference between 9-16 points (for 90% confidence interval) indicates that there was only a practice effect between the first and second administrations; conversely, a difference less than or equal to 8 points indicates a decline in performance, whereas a difference greater than or equal to 17 indicates a progression or increase in the ability measured by the subtest. Similarly, thresholds for indices can be used. In particular, for the FSIQ (see the standard score in Table 5), a difference between 6 and 27 IQ points between the first and second administration indicates a practice effect. Conversely, if the difference is equal to or less than 5 IQ points, then there was a decline, while if it is equal to or greater than 28 IQ points, there was an increase in performance not due to the practice effect.

This study is not without limitations: the sample consists only of children with typical development. Test-retest studies should therefore be conducted on other cultural and clinical samples to assess the generalizability of the threshold values proposed here. Another limitation is that these methods assume practice effects are the same for all children, but studies have shown this is not the case. For example, children with better intellectual abilities tend to have more significant practice effects on the retest. Despite these limitations, the threshold values proposed here should help practitioners identify whether the changes observed in a short period are significant or not, i.e., present in less than 10% of the population with typical development.

References

ANDERSON, P.L., CRONIN, M.E. & KAZMIERSKI, S. (1989). WISC-R stability and re-evaluation of learning-disabled students. *Journal of Clinical Psychology*, *45*, 941-944. doi:10.1002/1097-4679(198911)45:6_941: AID-JCLP2270450619_3.0.CO;2-P

BARTOI, M., 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*, 133-140. http://dx.doi.org/10.1080/21622965.2013.811075

BASSO, M.R., CARONA, F.D., LOWERY, N. & AXELROD, B.N. (2002). Practice effects on the WAIS-III across 3- and 6-month intervals. *The Clinical Neuropsychologist*, *16*, 57-63. http://dx.doi.org/10.1076/clin.16.1.57.8329

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

BROOKS, B.L., SHERMAN, E.M.S., IVERSON, G.L., SLICK, D.J. & STRAUSS, E. (2011). Psychometric foundations for the interpretation of neuropsychological test results. In M.R. Schoenberg & J.G. Scott (Eds.), *The little black book of neuropsychology: A syndrome-based approach*. Springer Science + Business Media.

BROOKS, B.L., STRAUSS, E., SHERMAN, E.M.S., IVERSON, G.L. & SLICK, D.J. (2009). Developments in neuropsychological assessment: Refining psychometric and clinical interpretive methods. *Canadian Psychology*, *50*, 196-209. http://dx.doi.org/10.1037/a0016066

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. (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. (2003). Study samples are too small to produce sufficiently precise reliability coefficients. *The Journal of General Psychology*, *130* (2), 117-129.

CHELUNE, G.J. (2003). Assessing reliable neuropsychological change. In R.D. Franklin (Ed.), *Prediction in forensic and neuropsychology: Sound statistical practices*. Erlbaum.

CHELUNE, G.J., NAUGLE, R.I., LÜDERS, H., SEDLAK, J. & AWAD, I.A. (1993). Individual change after epilepsy surgery: Practice effects and base-rate information. *Neuropsychology*, *7*, 41-52. https://doi.org/10.1037/0894-4105.7.1.41

CHEN, Z. & SIEGLER, R.S. (2000). Intellectual development in childhood. In R.J. Sternberg (Ed.), *Handbook of intelligence*. Cambridge University Press. doi:10.1017/CBO9780511807947.006

COHEN, J. (1994). The earth is round (p<.05). *American Psychologist*, *4* (12), 997-1003.

COHEN, J. (1988). *Statistical power analysis for the behavioral sciences*, *2nd ed*. Lawrence Erlbaum Associates.

CONLEY, J.J. (1984). The hierarchy of consistency: A review and model of longitudinal findings on adult individual differences in intelligence, personality and self-opinion. *Personality and Individual Differences*, *5*, 11-25.

DEARY, I.J., PATTIE, A. & STARR, J.M. (2013). The stability of intelligence from age 11 to age 90 years: The Lothian birth cohort of 1921. *Psychological Science*, *24*, 2361-2368. http://dx.doi.org/10.1177/0956797613486487

DEARY, I.J., WHALLEY, L.J., LEMMON, H., CRAWFORD, J.R. & STARR, J.M. (2000). The stability of individual differences in mental ability from childhood to old age: Follow-up of the 1932 Scottish Mental Survey. *Intelligence*, *28*, 49-55. http://dx.doi.org/10.1016/S0160-2896(99)00031-8

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

ESTEVIS, E., BASSO, M.R. & COMBS, D. (2012). Effects of practice on the Wechsler Adult Intelligence Scale – IV across 3- and 6-month intervals. *The Clinical Neuropsychologist*, *26*, 239-254. http://dx.doi.org/10.1080/13854046.2012.659219

HEILBRONNER, R.L., SWEET, J.J., ATTIX, D.K., KRULL, K.R., HENRY, G.K. & HART, R.P. (2010). Official position of the American Academy of Clinical Neuropsychology on serial neuropsychological assessments: The utility and challenges of repeat test administrations in clinical and forensic contexts. *The Clinical Neuropsychologist*, *24*, 1267-1278. doi:10.1080/13854046.2010.526785

HUNT, E. (2010). *Human intelligence*. Cambridge University Press.

JACOBSON, N.S., FOLLETTE, W.C. & REVENSTROF, D. (1984). Psychotherapy outcome research: Methods for reporting variability and evaluating clinical significance. *Behavior Therapy*, *15*, 336-352.

JACOBSON, N.S. & TRUAX, P. (1991). Clinical significance: A statistical approach to defining meaningful change in psychotherapy research. *Journal of Consulting and Clinical Psychology*, *59*, 12-19.

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

KAUFMAN, A.S. & KAUFMAN, N.L. (2008). *KABC-II batterie pour l'examen psychologique de l'enfant* (2 éd.). ECPA.

KIENG, S., ROSSIER, J., FAVEZ, N., GEISTLICH, S. & LECERF, T. (2015). Stabilité à long terme des scores du WISC-IV: Forceset faiblesses personnelles. *Pratiques psychologiques*, *21*, 137-154. http://dx.doi.org/10.1016/j.prps.2015.03.002

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

LECERF, T., KIENG, S. & GEISTLICH, S. (2017). WISC-IV: Valeurs seuils pour des changements significatifs des scores de différence test-retest [WISC-IV: Cutoff values for meaningful test-retest difference scores]. *Pratiques Psychologiques, 23* (4), 345-358. https://doi.org/10.1016/j.prps.2016.07.003

LEMAY, S., BEDARD, M.A., ROULEAU, I. & TREMBLAY, P.L. (2004). Practice effect and test-retest reliability of attentional and executive tests in middle-aged to elderly subjects. *The Clinical Neuropsychologist*, *18*, 284-302. doi:10.1080/13854040490501718

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

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? *Child Psychology & Psychiatry & Allied Disciplines*, *34* (4), 455-506. https://doi.org/10.1111/j.1469-7610.1993.tb01031.x

MOSIER, C.I. (1943). On the reliability of a weighted composite.

Psychometrika, *8*, 161-168. https://doi.org/10.1007/BF02288700

OKADA, S., KAWASAKI, Y., SHINOMIYA, M., HOSHINO, H., INO, T., SAKAI, K., ... & NIWA, S.I. (2021). Long-term stability of the WISC-IV in children with autism spectrum disorder. *International Journal of School & Educational Psychology*, 1-12. https://doi.org/10.1080/21683603.2021.1930307

ORSINI, A., PEZZUTI, L. & PICONE, L. (2012). *WISC-IV. Contributo alla taratura italiana*. Firenze: Giunti O.S. Organizzazioni Speciali.

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

REUCHLIN, M. (1992). *Introduction à la recherche en psychologie*. Nathan.

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

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. http://dx.doi.org/10.1080/09084280903297933Saklofske

SALTHOUSE, T. (2014). Frequent assessments may obscure cognitive decline. *Psychological Assessment*, *26*, 1063-1069. http://dx.doi.org/10.1037/pas0000007

SATTLER, J.M. (2008). Assessment of children. *Cognitive foundations (5th ed.)*. Jerome M. Sattler, Publisher, Inc.

SHERMAN, E.M.S., BROOKS, B.L., IVERSON, G.L., SLICK, D.J. & STRAUSS, E. (2011). Reliability and validity in neuropsychology. In M.R. Schoenberg & J.G. Scott (Eds.), *The little black book of neuropsychology*. Springer.

SHROUT, P.E. & FLEISS, J.L. (1979). Intraclass correlations: Uses in assessing rater reliability. *Psychological Bulletin*, *86* (2), 420-428.

SIMONTON, D.K. (2011). Exceptional talent and genius. In T. Chamorro-Premuzic, S. von Stumm & A. Furnham (Eds.), *Wiley-Blackwell handbook of individual differences.* Blackwell.

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

TRUSCOTT, S.D., NARRETT, C.M. & SMITH, S.E. (1994). WISC-R subtest reliability overtime: Implications for practice and research. *Psychological Reports, 74*, 147-156. http://dx.doi.org/10.2466/pr0.1994.74.1.147

WATKINS, M.W. & CANIVEZ, G.L. (2004). Temporal stability of WISC-III subtest composite: Strengths and weaknesses. *Psychological Assessment*, *16*, 133-138.

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

WATSON, D. (2004). Stability versus change, dependability versus error: Issues in the assessment of personality over time. *Journal of Research in Personality*, *38*, 319-350. doi:10.1016/j.jrp.2004.03.001

WECHSLER, D. (1949). *Manual for the Wechsler Intelligence Scale for Children*. Psychological Corporation.

WECHSLER, D. (1974). *Manual for the Wechsler Intelligence Scale for Children – Revised*. Psychological Corporation.

WECHSLER, D. (1991). *WISC-III manual.* Psychological Corporation.

WECHSLER, D. (2003a). *WISC-IV administration and scoring manual.* Psychological Corporation.

WECHSLER, D. (2003b). *WISC-IV technical and interpretive manual*. Psychological Corporation.

WECHSLER, D. (2012). *WISC-IV. Manuale di somministrazione e scoring*. It. ad. A. Orsini & L. Pezzuti (Eds.). Firenze: Giunti O.S. Organizzazioni Speciali.

WRIGHT, A. J. (2011). *Conducting psychological assessment: A guide for practitioners*.

Declaration of conflicting interests. One of the authors (L. Pezzuti) receives royalties from sales of the WISC-IV (Wechsler, 2012, It ad. Orsini & Pezzuti).