Long-term stability of the French WISC-IV: Standard and CHC index scores

Stabilité à long terme des indices standard et CHC du WISC-IV

Sotta Kieng^{a,b}, Jérôme Rossier^c, Nicolas Favez^{a,b}, & Thierry Lecerf^{a,b}

^aFaculty of Psychology and Educational Sciences (FPSE), University of Geneva, 40, bd du Pont-d'Arve, 1205 Geneva, Switzerland

^bDistance Learning University, Sierre, Switzerland

^cInstitute of Psychology, University of Lausanne, Quartier UNIL-Mouline, bâtiment Geopolis 2635, 1015 Lausanne Switzerland

Author Note

This work was supported by Grant 100014_135406 awarded by the Swiss National Science Foundation (*Long-term stability of the WISC-IV: Standard and CHC composite scores*; main applicant: T. Lecerf; co-applicants: N. Favez & J. Rossier).

Correspondence concerning this article should be addressed to Sotta Kieng, FPSE – Psychology, University of Geneva, 40, Bd. du Pont-d'Arve, CH-1205 Geneva, Switzerland.

Tel: +41 22 379 92 37

Fax: +41 22 379 92 29

E-Mail: sotta.kieng@unige.ch

Abstract

Introduction. – The assumption of the stability of intelligence is the source of the predictive value of the Intelligence Quotient (e.g, Full Scale IQ). However, few studies have investigated the long-term stability of one of the most frequently used tests in the field of cognitive assessment: the Wechsler Intelligence Scale for Children – 4th edition (WISC-IV).

Objective. – For a deeper understanding and a better use of intelligence test scores, this study examined the long-term stability of the standard index scores and five CHC composite scores of the French WISC-IV.

Method. – A test-retest procedure was used, with an average retest interval of 1.77 year ($SD = 0.56$ year). This study involved 277 French-speaking Swiss children aged between 7 and 12 years. Three types of stability analysis were conducted: (a) mean-level changes, (b) rank-order consistency and change, and (c) individual-level of change.

Results. – The observed pattern of mean-level changes suggested a normative mean-level stability for the Verbal Comprehension Index (VCI), the Perceptual Reasoning Index (PRI), the General Ability Index (GAI), Comprehension-Knowledge (Gc), and Visual Processing (Gv). Regarding individual differences stability, only the FSIQ and the GAI reached a reliability of .80 required for making decisions about individuals. Using a two standard errors of measurement confidence interval ($\pm 2 SEM$), we examined individual-level stability. Results indicated that more than 70% of the children presented stable performances for the GAI, Gc, and Gv scores.

Conclusion. – Together, nomothetic and idiographic perspectives suggested that the GAI, Gc, and Gv were the most stable scores in our non-clinical sample.

Résumé

Introduction. – L’hypothèse de la stabilité de l’intelligence est à l’origine de la valeur prédictive du Quotient Intellectuel (p.ex. QI Total). Or, peu d’études ont été conduites sur la stabilité à long terme des scores de l’une des batteries les plus utilisées dans le domaine de l’évaluation cognitive : la 4^{ème} édition de l’Échelle d’Intelligence de Wechsler pour Enfants et Adolescents (WISC-IV).

Objectif. – Afin de favoriser une compréhension approfondie et une meilleure utilisation des scores des tests d’intelligence, cette étude examine la stabilité à long terme des indices standards et de cinq indices CHC, estimés à partir de l’adaptation française du WISC-IV.

Méthode. – La stabilité à long terme des différents scores a été évaluée par le biais d’une procédure test – retest avec un intervalle moyen de 1,77 an ($ET = 0,56$ an) entre les deux passations. L’échantillon comprend 277 enfants suisses francophones âgés de 7 à 12 ans. La stabilité des scores a été évaluée sous trois angles : (a) la stabilité du niveau moyen du groupe, (b) la stabilité différentielle et (c) la stabilité intra-individuelle.

Résultats. – Les comparaisons de moyennes entre les deux passations suggèrent une stabilité du niveau moyen pour l’Indice de Compréhension Verbale (ICV), l’Indice de Raisonnement Perceptif (IRP), l’Indice d’Aptitude Générale (IAG), les scores compréhension-connaissance (Gc) et traitement visuel (Gv). Concernant la stabilité différentielle, seuls le QI Total et l’IAG atteignent un seuil de fidélité de .80 recommandé pour les décisions au niveau individuel. La stabilité intra-individuelle est examinée en définissant un intervalle de confiance de 2 erreurs types de mesure ($\pm 2 ETM$). Les résultats montrent que plus de 70% des enfants présentent des performances stables pour les scores de l’IAG, de Gc et de Gv.

Conclusion. – Globalement, la perspective nomothétique et la perspective idiographique suggèrent que l’IAG, Gc et Gv sont les scores les plus stables dans notre échantillon non clinique.

Keywords: WISC-IV; long-term stability; reliability; CHC, intelligence

Mots clés : WISC-IV; stabilité à long terme; fidélité; CHC, intelligence

1. Introduction

Previous studies suggested that intelligence is a steady and enduring trait across time (e.g., Deary, Pattie, & Starr, 2013; Deary, Whalley, Lemmon, Crawford, & Starr, 2000; Hertzog & Schaie, 1986; McCall, 1977). Indeed, apart from temporary fluctuations occurring in intellectual development, the cognitive performances of individuals are assumed to be relatively stable from childhood through adulthood. The stability of individual differences in intelligence confers a predictive value to the Full Scale Intelligence Quotient (FSIQ). Hence, intelligence tests like the Wechsler Scales are commonly used for diagnostic and intervention purposes. Because high-stakes decisions (e.g., grade-skip or admission to special education programs) are frequently based on the FSIQ and the index scores, it is essential to formulate diagnostic hypotheses and interventions based on reliable and stable intelligence test scores.

The reliability – and more particularly, the internal consistency – is routinely assessed for intelligence test scores. According to the Classical test theory, reliability / precision is the foundation for the validity of test score interpretation (AERA, APA, & NCME, 2014). Typically, a test-retest procedure is used to assess the reliability / precision of intelligence test scores across time (i.e., longitudinal studies with two assessments at least). With this procedure, the same test is administered to the same individuals twice with a defined retest interval, and test-retest correlations are computed to assess the stability.

Most longitudinal studies indicated that when individuals are tested again after an interval of several days or several years (with the same measure or alternate forms), their performance at the second assessment was higher than their performance at the first assessment (e.g., Calamia, Markon, & Tranel, 2012; Hausknecht, Halpert, Di Paolo, & Moriarty Gerrard, 2007; Salthouse, Schroeder, & Ferrer, 2004). For instance, studies conducted with a retest interval of 1 year or less reported retest gains between 0.10 and 0.60

Time 1 SD^1 (see review in Benedict & Zgaljardic, 1998; and Salthouse et al., 2004).

Furthermore, these studies indicated that retest gains varied with tasks. Typically, crystallized abilities demonstrated higher stability than fluid abilities (Schwartzman, Gold, Andres, Arbuckle, & Chaikelson, 1987). These studies also demonstrated that tests with problem-solving components were subject to greater practice effects than those with fewer such demands (Calamia et al., 2012; Dikmen, Heaton, Grant, & Temkin, 1999). Similarly, with a short retest interval (from 3 to 6 months) retest gains tended to be greater for simple speed task (e.g., processing speed subtests such as Coding or Symbol Search) compared to verbal ones (e.g., verbal comprehension subtests such as Vocabulary or Information; Calamia et al., 2012; Estevis, Basso, & Combs, 2012). In the longitudinal study conducted with adults (between 18 and 58 years), and tested twice after an interval of a few days to 35 years, Salthouse and colleagues (2004) demonstrated that seven or more years were needed to remove the positive retest effects.

Two main factors may explain the longitudinal change: age and retest effects (i.e., practice effects; Ferrer, Salthouse, McArdle, Stewart, & Schwartz, 2005; Salthouse et al., 2004). While age effects refer to maturation (aging processes), “*retest effects refer to influences on the difference in performance between the first and a subsequent measurement occasion that are attributable to the previous assessment*” (Salthouse, 2009, p. 509).

According to Salthouse and colleagues (2004), four types of influences (specific and general retest factors) could contribute to the retest effects: (1) test-specific factors (e.g., remembering items or answers); (2) familiarity with the testing situation that could reduce the anxiety; (3) increase in the cognitive ability assessed by the test during the test-retest interval;

¹ To compare different measures, the practice effect is expressed in Time 1 standard deviation units (i.e., T1 SD units). The T1 SD is a non-dimensional unit of effect size and is calculated as the ratio of change from Time 1 to Time 2 (i.e., $T1\ SD\ units = \frac{T2-T1}{SD\ of\ T1}$).

and (4) changes that occur in the environment of the individual. These authors assumed that the fourth influence is more relevant for general information or vocabulary tests.

Several methods are used to estimate age and retest effects. For instance, the comparisons of performances between children tested twice and children with the same age tested once allow assessing retest effects. However, because some longitudinal studies demonstrated that retest effects could contaminate age effects, it is necessary to distinguish effects due to the age or to the retest (Ferrer et al., 2005; Salthouse et al., 2004). Indeed, with adult samples, Salthouse and colleagues (2004) suggested that positive retest effects could obscure negative age effects. One method to distinguish these effects is to vary the retest interval among participants. Thus, there will be no more perfect correlation between the increase of age and the increase of retest interval. This procedure has only rarely been used. In the present study, because the retest interval varies among children, we will be able to decompose age and retest effects.

To our knowledge, the distinction between “short-term” and “long-term” retest interval is not clearly defined in the literature. Sattler (2008) considers a period less than one year as a short time interval (< 1 year). Similarly, for Watkins and Canivez (2004), a long time interval is a retest interval of more than one year (> 1 year). Close to these definitions, we consider that a period of one year or more is a long-term interval (≥ 1 year).

To date, very few studies have investigated the long-term stability of the *Wechsler Intelligence Scale for Children—fourth edition* (WISC-IV), and as far as we know, none with the French version. More than short-term stability, long-term stability is needed to provide complete evidence to support the predictive value of high-stakes decisions based on test scores (i.e., decision consistency). The stability of the previous editions of the WISC has been explored with several test-retest intervals and with various groups of children (non-clinical or clinical groups). Most longitudinal studies conducted with the U.S. WISC / WISC-R / WISC-

III indicated that the FSIQ was fairly stable (i.e., $r > .70$) with clinical samples (e.g., Bauman, 1991; Canivez & Watkins, 1998, 2001; Oakman & Wilson, 1988; Stavrou, 1990; Truscott, Narrett, & Smith, 1994; Vance, Blixt, Ellis, & Debell, 1981). Because of many changes in the WISC-IV, the previous findings are obsolete. The clinical interpretation of this fourth edition is currently based on a *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).

Table 1 reports results for some short- and long-term longitudinal studies conducted with the WISC-IV. Stability coefficients corrected for the variability of the WISC-IV normative sample (Allen & Yen, 2001; Guilford & Fruchter, 1978; Magnusson, 1967), and stability coefficients corrected for combined or additive effect of content and time sampling error (Macmann & Barnett, 1997) were reported. The standardized mean difference (i.e., d) is the difference of the two test means divided by the pooled standard deviation (Cohen, 1977).

In the technical and interpretative manual of the French WISC-IV (Wechsler, 2005b), short-term stability of the scores was evaluated with a sample of 93 non-clinical children, who were tested twice one month apart (mean test-retest interval = 27 days). The corrected test-retest coefficients ranged from .78 (WMI) to .91 (FSIQ) for the index scores, and from .64 (Picture Concepts) to .83 (Symbol Search) for the subtest scores (see Table 1). The short-term stability of the U.S. WISC-IV scores was evaluated in a sample of 243 children with a retest interval between 13 and 63 days (mean test-retest interval = 32 days). As reported by Williams, Weiss, and Rolfhus (2003), corrected stability coefficients ranged from .86 (PSI) to .93 (VCI and FSIQ). Regarding the subtest scores, corrected stability coefficients ranged from .76 (Picture Concepts) to .92 (Vocabulary) (see Table 2 in Williams et al., 2003). These results indicated that the French WISC-IV corrected stability coefficients were slightly lower than those reported for the U.S. WISC-IV scores.

A longer retest interval (11 months) was considered by Ryan, Glass, and Bartels (2010), who investigated the stability of the U.S WISC-IV scores with a sample of 43 voluntary children from a private school (see Table 1). Except for the PRI ($r = .68$) and the PSI ($r = .54$), corrected stability coefficients of index scores were above .70. Stability coefficients of the subtest scores ranged from .26 (Picture Concepts) to .84 (Vocabulary). These results indicated that the stability coefficients of subtest scores were lower than those of the composite scores. At an individual level, Ryan and colleagues found that 42% of children changed their FSIQ by more than 5 points between both assessments.

As mentioned before, most short-term stability studies conducted on intelligence tests have revealed retest effects. For the WISC-IV, several studies have shown that the practice effects were more pronounced for the PRI and the PSI than for the WMI and the VCI (Flanagan & Kaufman, 2009; Ryan et al., 2010; Wechsler, 2005b). Flanagan and Kaufman (2009) also observed that “practice effects are largest for ages 6 to 7 and become smaller with increasing age” (p. 32). In addition, Ryan et al. (2010) found that children with higher performances benefited more from the second testing than children with lower performances. However, because this study was conducted with a small sample and with a short retest interval (< 1 year), these findings must be taken with caution and could not be extrapolated to all children.

As far as we know, only three studies have investigated the stability of the WISC-IV scores with a long test-retest interval (i.e., ≥ 1 year). First, a long-term stability study was conducted by Lander (2010). This study involved a sample of 131 children with learning disabilities. The test-retest interval was 2.89 years. Except the FSIQ, the uncorrected long-term stability coefficients were lower than .70 (see Table 1). Concerning the subtest scores, the long-term stability coefficients ranged from .28 (Symbol Search) to .62 (Block Design). In addition, Lander found a significant mean decrease from the first to the second assessment

for the PSI (-2.14 points), but the associated effect size was weak ($d = -.18$). In order to examine the intraindividual stability, Lander analyzed the change in individual scores between test and retest by computing a confidence interval based on a ± 2 Standard Error of Measurement ($\pm 2 SEM$). Lander found that 78% (VCI and FSIQ), 73% (PRI), 70% (WMI), and 73% (PSI) of the children remained stable across time with this $\pm 2 SEM$ confidence interval. Thus, Lander stated that “*there were many individuals who changed more than would be expected due to error*” (p. 75). These results might be explained by the fact that selecting a specific sample of children with learning disabilities restricted the range and hence lowered test-retest correlations.

The second long-term stability study on the WISC-IV scores was conducted by Watkins and Smith (2013). Three hundred and forty-four children evaluated for special education eligibility were tested twice with a retest interval around 3 years ($M = 2.84$ years, $SD = 0.75$ year). Except the PSI ($r = .65$), corrected stability coefficients ranged from .70 (WMI) to .84 (FSIQ; see Table 1). Again, corrected stability coefficients for the subtest scores were lower than those for the composite scores. Regarding intraindividual stability, Watkins and Smith found that 71% and 75% of the children had test-retest differences less than or equal to 9 points for the VCI and the FSIQ, respectively. These percentages were 61%, 63%, and 56% for the PRI, WMI, and PSI, respectively. Therefore, Watkins and Smith concluded, “*even the most reliable WISC-IV score, the FSIQ, may not be sufficiently stable for longitudinal individual decisions*” (p. 4).

The third long-term stability study on the WISC-IV scores was conducted by Bartoi et al. (2015). Participants in this study were 51 clinically referred children aged from 8 to 16 years. The average retest interval was 1.84 year ($SD = 0.50$ year). The uncorrected stability coefficients for index scores ranged from .58 (PSI) to .86 (FSIQ), and from .35 (Letter-Number Sequencing) to .81 (Vocabulary) for subtest scores (see Table 1). Individual

variation in scores showed that 78.4% of the children had test-retest differences less than or equal to 9 points for the FSIQ; similarly 68.6%, 56.9%, 54.9, and 54.9% of the children had test-retest differences less than or equal to 9 points for the VCI, PRI, WMI, and PSI, respectively. Overall, these results were consistent with those reported by Watkins and Smith (2013).

2. Aims of the study

Despite the increasing use of the Wechsler Intelligence Scales, there is a general lack of research investigating the long-term stability of subtest and composite scores. Thus, the specific objective of the present research was to examine the long-term stability of the French WISC-IV scores with young children. The most interesting features of the present study were the large non-clinical French-speaking Swiss sample and the consideration of both perspectives, nomothetic (group level) and idiographic (intraindividual level). Firstly, *mean-level changes* provided information on the stability of the mean of the group for each index / subtest score. The mean-level changes refer to the extent to which index / subtest scores change over time. Secondly, *rank-order consistency and change* provided information on the stability of interindividual differences. Rank-order consistency describes the degree to which the relative differences among individuals remain invariant across time. Some degree of rank-order stability is a defining feature of psychological traits. Indeed, if individual differences are not stable enough, it is considered that performance is related mostly to states rather than to traits. Thirdly, *individual-level of change* provided information on intraindividual stability / instability. This analysis of intraindividual-level of change refers to the magnitude of increase or decrease exhibited by a child on an index / subtest score. From a clinical point of view, this type of analysis is crucial because mean-level changes (i.e. group level) do not give

any information on specific individuals. Individual-level change may be masked in the mean-level analysis because equal numbers of children may increase or decrease on a score.

In addition, in the first set of supplemental analyses, we examined the extent to which retest effect in a cognitive ability was associated with retest effect on other cognitive abilities. Second and as mentioned before, a design in which there is variability among participants in the retest interval makes it possible to distinguish between “retest effects” and “age effects”. In other words, because children varied in the retest intervals, the increase in age was not perfectly correlated with the increase in the retest interval. Hence, in the second set of analyses, the difference between the second assessment and the first assessment for a given score was correlated with the duration of the retest interval. In the third set of analyses, we investigated if the difference between both assessments for a given score was correlated with age at initial testing. Finally, we examined the relation between initial level and subsequent change.

3. Method

3.1. Participants

The sample was composed of French-speaking Swiss children aged from seven to twelve years, attending school at the canton of Geneva, Switzerland. Children were tested twice with the WISC-IV during school hours. Participant selection was restricted to primary students because in the Geneva school system, students change schools during the transition from primary to secondary school. The participation in this academic clinical study², which aims to assess the long-term stability of the WISC-IV scores with non-clinical children, was voluntary and the informed parental consent was obtained for all children. The planned sample was 250 children. There are 164 public elementary schools in the canton of Geneva.

² This research was funded through grant from the Swiss National Science Foundation to T. Lecerf, N. Favez, and J. Rossier (grant 100014-135406: « Long-term stability of the WISC-IV: Standard and CHC composite scores »).

The Public Instruction Department gave permission to contact school principals of 23 elementary schools in different areas. A total of 18 schools and 40 classes volunteered to participate in this research. The school principals forwarded the application to teachers. Only teachers who volunteered to participate forwarded the request, in turn, to parents. No data is available on the number of teachers or parents who have received the information for the first assessment. According to the progress of their class in the school program or number of courses that engage the whole class, teachers decided how many children in the class could participate. In the first assessment, 283 children were tested. Because of the refusal of 6 parents, 277 children were tested at the second assessment. Thus, the final sample includes 277 children (132 boys and 145 girls).

All children who completed both assessments were included in the sample. Parents who agreed to participate were asked to fill out a questionnaire about sociodemographic information. Compared to the Geneva census, our sample demonstrated an over-representation of children from higher-income households (32% vs. 19%) and from middle-income households (45% vs. 43%). Conversely, our sample demonstrated an under-representation of children from lower-income households (23% vs. 38%). Note that the supplemental Picture Completion subtest was not administered to 29 children of the sample.

3.2. Instrument and Index Scores

The WISC-IV is one of the most frequently used tests to assess the general intellectual functioning of French-speaking children. The WISC-IV is an individually administered test of intelligence for children of ages 6 years 0 months through 16 years 11 months. The WISC-IV includes 10 core subtests (Block 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 French WISC-IV was standardized on a nationally representative sample ($N = 1,103$) closely approximating the 1999 French census on gender, parents' socioeconomic status, and French geographic region.

Currently, interpretation of the WISC-IV is mainly based on the Full Scale Intelligence Quotient (FSIQ) and on 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. Finally, the FSIQ is derived from the sum of the ten core subtest scores.

Two additional standard index scores were computed: the General Ability Index and the Cognitive Proficiency Index. The General Ability Index (GAI) has been introduced as an alternative to the FSIQ with which it correlates highly (for the French sample: $r = .92$; Lecerf, Reverte, Coleaux, Favez, & Rossier, 2010; Lecerf et al., 2011). The GAI is a composite score derived from the Verbal Comprehension and the Perceptual Reasoning subtest scores (i.e., 6 subtest scores). By reducing the influence of basic cognitive processes like working memory and processing speed, the GAI should provide a better estimate of general intellectual functioning for some children (e.g., gifted children or children with ADHD; Lecerf, Bovet-Boone, Peiffer, Kieng, & Geistlich, 2016). The Cognitive Proficiency Index (CPI) is the counterpart of the GAI and is derived from the Working Memory and the Processing Speed subtest scores (i.e., 4 subtest scores). The CPI reflects the proficiency at which an individual process cognitive information. Several studies have demonstrated the clinical benefit in using the GAI and the CPI for assessing children referred for psychoeducational difficulties

(Bremner, McTaggart, Saklofske, & Janzen, 2011; Lecerf et al., 2016; Prifitera, Saklofske, & Weiss, 2008; Saklofske, Zhu, Coalson, Raiford, & Weiss, 2010). The index scores and the FSIQ retain an age-blocked population mean of 100 and standard deviation of 15.

Because several Confirmatory Factor Analyses (CFAs) demonstrated that models based on the CHC theory better fitted the U.S. and the French WISC-IV data, the long-term stability of five CHC composite scores was assessed (e.g., Chen, Keith, Chen, & Chang, 2009; Golay, Reverte, Rossier, Favez, & Lecerf, 2013; Keith, Fine, Taub, Reynolds, & Kranzler, 2006; Lecerf, Rossier, Favez, Reverte, & Coleaux, 2010; Reverte, Golay, Favez, Rossier, & Lecerf, 2014). Comprehension-Knowledge (Gc) reflects knowledge emerging from the education, experiences, and interactions with the environment (Ghisletta & Lecerf, in press). This index score is derived from the sum of Similarities and Comprehension scores. Fluid Reasoning (Gf) reflects the skills of inductive and deductive reasoning, forming concepts, and adaptability to new problems. This index score is derived from the sum of Picture Concepts and Matrix Reasoning scores. Visual Processing (Gv) reflects the ability to mentally manipulate visual stimuli, analyze, and move them. This index score is derived from the sum of Block Design and Picture Completion scores. Associated with the immediate storage, Short-Term Working Memory (Gwm) reflects the ability to maintain information and to recall it. This index score is derived from the sum of Digit Span and Letter-Number Sequence scores. Finally, Processing Speed (Gs) reflects the ability to perform quickly simple tasks, to maintain attention and concentration on visual stimuli. This index score is derived from the sum of Coding and Symbol Search scores.

Through a statistical approximation procedure, Lecerf et al. (2012) have developed French CHC tables that were used in the present study. It should be noted that Gwm tables are close to the WMI norms reported in the French technical manual of the WISC-IV, although scores are not strictly similar (± 1 point in some cases). Similarly, Gs tables are

close to the PSI norms. All the CHC composite scores (i.e., Gc, Gf, Gv, Gwm, and Gs) have a mean of 100 and a standard deviation of 15.

3.3. Procedure

According to the WISC-IV standard procedures for administration and scoring (Wechsler, 2005a), the ten core subtests of the French WISC-IV and the supplemental subtest Picture Completion were administered twice to each child. Six licensed psychologists, supervised by the same doctoral level psychologist, conducted the administrations during school hours. Stamped in advance with the approval from an ethics commission and the Public Instruction Department of Geneva, the six licensed psychologists administered the WISC-IV to the child presenting a properly executed parental permission slip. Only children who had proficiency in speaking, understanding and reading French, and who were in the school grade appropriate to their chronological age were recruited.

To minimize potential bias, the examiners were blind to previous scores. Furthermore, at least two psychologists scored the 554 WISC-IV protocols and agreement was reached for all ambiguous responses.

The length of the retest intervals was not defined in advance. Children were retested randomly, when retest interval was at least twelve months. To ensure the cooperation and to minimize teacher's constraints, the assessments were scheduled at their convenience. The retest interval ranged from 1 to 3.25 years ($M = 1.73$ year, $SD = 0.56$ year, $Mdn = 1.67$). The mean age at the first testing was 8.87 years ($SD = 0.82$ year; range = 7 years 0 month – 11 years 1 month) and the mean age at the second testing was 10.64 years ($SD = 1.11$ year; range = 8 years 1 month – 12 years 7 months).

4. Results

Descriptive statistics (Means, *SDs*, Mean score differences, and Cohen's *d*) and long-term stability coefficients of the WISC-IV index and subtest scores are reported in Table 2. First and as expected, IQ scores and subtest scores are close to the theoretical means (i.e., 100 and 10) and to the theoretical standard deviations (i.e., 15 and 3). For the first assessment, the means IQs ranged from 95.18 (WMI) to 104.90 (VCI), with standard deviations between 13.88 (PSI) and 15.23 (VCI). For the second assessment, the means IQs ranged from 97.20 (WMI) to 107.20 (PSI), with standard deviations between 13.60 (PRI) and 15.17 (VCI). Thus, the children in the current sample have an average level of functioning similar to that of the French standardization sample, and approximately the same degree of variability.

There was some variability in the retest intervals, but no controlled randomization because it depended on the teachers' schedule. We computed the correlations between age at the first assessment and the retest interval. This correlation was significant and positive ($r = .26$, $SE = .06$), indicated that the retest interval was a little bit longer for the older children. Although a non-significant correlation would be expected if retest interval were completely randomized, this correlation is weak ($r^2 = 6.76\%$).

4.1. Mean-level changes

Dependent *t*-tests were conducted to examine mean-level stability from test to retest. To ensure that the overall experimentwise error rate remained at .05, Holm-Bonferroni correction was applied (Aickin & Gensler, 1996; Gaetano, 2013; Holm, 1979). As can be seen from Table 2, there were significant mean-level increases between test and retest for the WMI (+2.02), PSI (+5.79), FSIQ (+2.53), CPI (+4.87), Gf (+2.08), Gwm (+2.01), and Gs scores (+5.51). In contrast, no significant mean-level differences between the first and the second assessment were found for the VCI, PRI, GAI, Gc, and Gv scores. When scores

increased significantly from the first to the second assessment, the magnitude of differences was small to medium, suggesting small practice effects (Cohen, 1992). This is the case for the PSI, FSIQ, CPI, and Gs scores. Regarding subtests, Coding (+0.82) and Symbol Search scores (+1.09) showed significant mean-level increases from test to retest, with a small to medium magnitude. This finding suggests small practice effects for these subtests.

4.2. Rank-order consistency

Three types of long-term stability coefficients were computed: raw test-retest stability coefficients (r_{12}), stability coefficients corrected for variability of the normative sample (r_c) (Magnusson, 1967), and stability coefficients corrected for the combined or additive effect of content and time sampling error (r_t) (Macmann & Barnett, 1997) (see Table 2). First, it should be noted that raw coefficients (r_{12}) were relatively similar to the coefficients corrected for the variability of the normative sample (r_c). As expected, the FSIQ and GAI scores were the most stable ($r = .83$). Regarding the CHC composite scores, Gc and Gv were the most stable scores ($r = .73$ and $.78$, respectively). When both internal consistency and long-term stability were taken into account (i.e., r_t), the FSIQ and GAI scores remained the most stable ($r = .76$). Consistent with previous studies, stability coefficients were lower for the subtest scores than their corresponding index score.

4.3. Individual-level stability

Individual variations in scores across the retest interval indicated that the difference between test and retest scores varied from -25 to +28 points for the FSIQ (see Table 3). Only 18 children earned identical FSIQs at both assessments.

According to the procedure used in previous studies, a two Standard Error of Measurement confidence interval ($\pm 2 SEM$) was determined to assess individual-level

stability / instability (e.g., Canivez & Watkins, 2001; Lander, 2010). Theoretically, less than 5% of children will present a variation higher than two *SEM* (i.e., 95% confidence interval). If the score at the second assessment falls within the ± 2 *SEM* interval determined on the basis of the score of the first assessment, a child's performance will be considered stable. In contrast, if the score at the second assessment does not fall within the ± 2 *SEM* interval determined based on the score of the first assessment, a child's performance will be considered unstable.

The *SEM* for the standard index and the subtest scores were provided in the French technical manual of the WISC-IV (Wechsler, 2005b, p. 32, Table 4.2), and was derived from the internal consistency reliability coefficients (Wechsler, 2005b, p. 32, Table 4.1). Lecerf, Reverte, and colleagues (2010) and Lecerf and colleagues (2012) computed the *SEM* for the GAI, the CPI, and the CHC composite scores. Applied to the WISC-IV, a ± 2 *SEM* confidence interval suggests that performance at the second assessment should fall within ± 9.96 , ± 10.48 , ± 10.34 , ± 12.02 , ± 7.26 , ± 8.28 , and ± 9.96 points for the VCI, PRI, WMI, PSI, FSIQ, GAI, and CPI, respectively. For the CHC composite scores, and with this 95% confidence interval (i.e., ± 2 *SEM*), the score at the second assessment should fall within the interval defined by ± 11.85 , ± 11.62 , ± 11.42 , ± 9.90 , and ± 11.70 points for Gc, Gf, Gv, Gwm, and Gs, respectively. To our knowledge, no criterion has been proposed in the literature to determine whether an individual's stability is satisfactory or not.

Results indicated that the GAI was the most stable index score at an individual level. More precisely, no change has occurred across time for 73.6% of children, because the child's GAI remained within a ± 2 *SEM* (± 8.28 points). Instead of the 95% theoretically expected, 69% (VCI), 69.7% (PRI), 63.9% (WMI), 59.6% (PSI), 58.8% (FSIQ), and 57% (CPI) of children remained stable across time. For the CHC composite scores, Gc and Gv were the most stable scores at an individual-level; precisely, 72.6% and 74.2% of children

remained stable across assessments. For Gf, Gwm, and Gv scores, 64.6%, 63.5%, and 57.8% of the children remained stable, respectively. Regarding the subtest scores, results indicated that more than 80% of the children remained stable across time for Picture Concepts, Coding, and Comprehension scores.

We can also mention that for the FSIQ, 32.49% of the children remained stable with a ± 1 SEM confidence interval (± 3.63 points), instead of the 68% theoretically expected. If we consider a ± 3 SEM confidence interval (± 10.89 points), our data showed that 77.26% of the children remained stable for the FSIQ during the retest interval, instead of the 99% theoretically expected.

4.4. Correlates of changes among cognitive abilities and contribution of age and retest

The first set of analyses aimed to investigate the extent to which retest effect in a given cognitive ability was associated with retest effect on another cognitive ability. In other words, we examined whether changes in cognitive abilities covary or whether they change independently. One way to investigate the correlations between changes among cognitive abilities (i.e., scores) is to compare the difference between the second and the first assessment for a given score ($\Delta_{\text{index A}} = I_{A2} - I_{A1}$) with the difference between the second and the first assessment for a second score ($\Delta_{\text{index B}} = I_{B2} - I_{B1}$). Results indicated that only changes in VCI (i.e., Δ_{VCI}) were significantly and positively correlated with changes in the three other cognitive abilities; however, these correlations were weak ($\Delta_{\text{VCI}} \times \Delta_{\text{PRI}} = .14$, $SE = .06$; $\Delta_{\text{VCI}} \times \Delta_{\text{WMI}} = .16$, $SE = .06$; $\Delta_{\text{VCI}} \times \Delta_{\text{PSI}} = .15$, $SE = .06$). Thus, the rates of change of the various cognitive functions did not substantially correlate.

To distinguish between age and retest effects, the difference between the second and the first assessment for a given score ($\Delta_{\text{index A}} = I_{A2} - I_{A1}$) was correlated with the duration of the retest interval ($\Delta_T = T_2 - T_1$). Correlations were significant and negative for the FSIQ and

the GAI ($\Delta_{FSIQ} \times \Delta_T = -.16$, $SE = .06$; $\Delta_{GAI} \times \Delta_T = -.15$, $SE = .06$). This finding suggested that children who were assessed twice with a shorter retest interval benefited slightly more from prior administration of the WISC-IV than did children who were assessed twice with a larger retest interval. However, these correlations are too small to have any clinical consequence.

The third set of analyses examined the age effect. To deal with this question, the difference between the second and the first assessment for a given score ($\Delta_{index\ A} = I_{A2} - I_{A1}$) was correlated with age at initial testing (A_1). We found that for the PRI, PSI, CPI, and FSIQ, the gain from test to retest was significantly and positively associated with age at initial testing ($\Delta_{PRI} \times A_1 = .13$, $SE = .06$; $\Delta_{PSI} \times A_1 = .17$, $SE = .06$; $\Delta_{CPI} \times A_1 = .20$, $SE = .06$; $\Delta_{FSIQ} \times A_1 = .19$, $SE = .06$). Regarding the FSIQ, this finding suggested that older children at the first assessment tended to benefit a bit more from prior administrations of the WISC-IV than did younger children.

In the fourth set of analyses, the relation between initial level and subsequent change was tested by correlating the difference between the second and the first assessment for a given score with the score at the initial testing (e.g., $\Delta_{FSIQ} \times FSIQ_1$; $\Delta_{VCI} \times VCI_1$; etc.). Correlations were significant and negative for all scores (ranging from $-.32$ for $\Delta_{VCI} \times VCI_1$ to $-.45$ for $\Delta_{PRI} \times PRI_1$). These findings suggested that children with a higher level at the first assessment tended to present a smaller difference between scores, while children with a lower level at the first assessment tended to present a larger difference between the first and the second score. In contrast to Ryan's and colleagues (2010), we found a significant negative correlation between the FSIQ at initial testing and the amount of gain on the retest ($r = -.44$; $SE = .05$). In other words, children with lower FSIQ at the first assessment tended to benefit more from prior administration of the WISC-IV than did children with higher ability at the first assessment. This finding could be due to the regression to the mean effect (Kieng et al. 2015).

5. Discussion

The present study investigated the long-term stability of the French WISC-IV scores with non-clinical children tested twice. To our knowledge, only three studies have examined the long-term stability of the WISC-IV scores: Lander (2010), Watkins and Smith (2013), and Bartoi et al. (2015). These studies were conducted with U.S. clinical samples. In contrast, our sample involves non-clinical French-speaking Swiss children. Thus, the present study provides useful information about the generalization of the long-term stability of the WISC-IV scores. Furthermore, we also reported the long-term stability coefficients for five CHC composite scores.

First, dependent samples *t*-tests indicated significant increases between test and retest scores for the WMI, PSI, FSIQ, CPI, Gf, Gwm, and Gs. However, the effect sizes were negligible or small to medium (i.e., $d < 0.50$). According to Salthouse and colleagues (2004), it can be suggested that the increase in the PSI is due to an increase in the relevant cognitive ability (i.e., processing speed). Indeed, results indicated a significant increase in both Coding and Symbol Search scores. According to research in field dependence-independence (FDI), there is a relation between FDI and performances with the WISC, and more particularly Block Design and Picture Completion (Huteau, 1985). Field independence individuals (FI) tended to make greater use of spatial strategy, while field dependence individuals (FD) “*made more use of the strategies induced by the items*” (Rémy and Gilles, 2014, p. 81). The practice effect could also be explained because FI individuals may recall better strategies. Most importantly, the increases were smaller than those reported in the French technical manual with a short-term test-retest interval, and more particularly for the PSI and the FSIQ (see Table 1: $d = 0.81$ and 0.59 , respectively). We assumed that the smaller practice effects are due to the longer retest interval. In addition, no mean-level change was found in the present

study for the VCI, PRI, GAI, Gc, and Gv. These findings are relatively consistent with those reported in previous studies, and hence, we agree with Canivez and Watkins (2001): although practice effects occur in the long-term retest study, the effect sizes tend to be too small to have any real practical consequences.

Secondly, from a nomothetic perspective, corrected stability coefficients ranged from .60 (Gf) to .83 (FSIQ, GAI). At this step, it is important to remember that there are no clear rules to define a satisfactory level of reliability. It depends first on how a score is being used, and whether level of reliability will describe an individual or a group (Abell, Springer, & Kamata, 2009; Nunnally & Bernstein, 2010; Salvia, Ysseldyke & Bolt, 2012; Thorndike & Thorndike-Christ, 2010). At a group level, a reliability of .70 could be sufficient in the first stage of test development, while a reliability of .80 could be required in basic research (Abell, Springer, & Kamata, 2009; Nunnally & Bernstein, 2010; Thorndike & Thorndike-Christ, 2010). At an individual level, while some authors suggested that a reliability of .80 is the minimum for relying individual high-stakes decisions (e.g., diagnosis, intervention, placement, or selection), Nunnally and Bernstein (2010) indicated that a minimum reliability of .90 is required, or even a minimum of .95 (see also Wasserman & Bracken, 2013). Charter and Feldt (2001) have examined the relation between reliability cutoff scores and correct or incorrect clinical decision. They found that *“to be confident that one will correctly classify at least 90% of the examinees truly in need of treatment one would need an r of about .98 or better”* (p. 533).

Following these different cutoff values, the current data suggested that corrected stability coefficients for the VCI, PRI, FSIQ, GAI, CPI, Gc, and Gv were sufficient at a group level. According to Thorndike and Thorndike-Christ (2010), if we considered that a reliability of .80 is the minimum level of reliability for making decisions about individuals, only the FSIQ and the GAI could be used. However, even the FSIQ did not reach a reliability of .90

recommended by Nunnally and Bernstein. To be more concrete, a reliability of .80 for the FSIQ indicated that if a first individual falls at the 75th percentile of the group, and if a second individual falls at the 50th percentile at the first assessment, we have 4 chances in 5 that the difference remains in the same direction at the second assessment. According to Charter and Feldt (2001), with stability coefficient around .80 for the FSIQ and the GAI, and with a cutoff value of about 0.5 *SD* (i.e., ± 8 points), there are about 80% of true cases testing positive. Anyway, the present findings are relatively consistent with those reported by Bartoi and colleagues (2015), and by Watkins and Smith (2013): the FSIQ is the most stable score (with the GAI).

We must also emphasize that the U.S. and the French long-term stability coefficients were smaller than the U.S. and the French short-term stability coefficients. For instance, the stability coefficient was .91 for the French FSIQ (see Table 1). Most importantly, the assumption that intelligence is a steady and enduring trait across time is mainly based on short-term studies. However, the long-term data are not really consistent with this assumption.

Regarding the subtest scores, most stability coefficients indicated low test-retest stability, except those for Block Design, Vocabulary, and Picture Completion scores (see Table 2). Overall, the long-term stability coefficients of subtest scores were lower than those for the index scores. These findings extend the conclusions of previous studies (Bartoi et al., 2015; Lander, 2010; Watkins & Smith, 2013). It should be noted that the long-term stability of Block Design and Vocabulary scores were also higher than .70 in Bartoi's and colleagues and Watkins' and Smith studies. The long-term stability of the subtest scores in the present research pointed out that the subtest scores are unreliable from a nomothetic perspective. This finding is in line with recommendations made before: we should not use subtest profiles for diagnosis because of instability.

Thirdly, because of its clinical relevance, we analyzed individual stability / instability in FSIQ, index scores, and subtest scores across the retest interval. Indeed, it is well known that interindividual stability doesn't give any information about intraindividual stability (see for instance Voelkle, Brose, Schmiedek, & Lindenberger, 2014). A stable index score at an interindividual level (nomothetic perspective) may be less stable at an individual-level (idiographic perspective), and vice versa. Take for example the FSIQ: according to Thorndike and Thorndike-Christ, a sufficient long-term stability was found ($r = .83$), but this was not really the case at the individual-level. Indeed, 41.2% of children showed changes more than $\pm 2 SEM$ on retest. In contrast, Picture Concepts score had a weak long-term stability ($r = .51$) but a better individual-level stability, because 80.1% of children remained stable across time. To perform the idiographic analysis, individual stability / instability in FSIQ, index scores, and subtest scores, a confidence interval was determined with a ± 2 standard error of measurement ($\pm 2 SEM$). Changes in the scores were considered stable across the retest interval if performance remained within $\pm 2 SEM$, that is within an interval of ± 7.26 points for the FSIQ. By using this confidence interval, the FSIQ was stable only for 58.8% of children. Taken together, our findings suggested – in line with previous studies – that index scores, and particularly the FSIQ, were less stable at the individual-level than one might infer from stability coefficients (i.e., interindividual analyses), and from the negligible practice effects. For the FSIQ, we also found that 22.74% of children demonstrated changes more than $\pm 3 SEM$ (± 10.89 points) on retest. This result is consistent with Watkins' and Smith (2013) findings, which indicated that about 25% of children showed changes in FSIQ of 10 points or more. In sum, from an idiographic perspective, we found that the GAI, Gc, and Gv were the most stable at an individual level (i.e., almost 80% of children present stable performances between both assessments).

A series of analyses were conducted to determine whether changes in cognitive abilities covary or whether they change independently, and to examine age effects and retest effects. These analyses indicated that the rates of change of the various cognitive scores did not correlate, except for the VCI. Nevertheless, the correlations were too low to have any clinical consequence. This finding suggested that longitudinal changes for index scores were not related to the development of the general factor (*g* factor). Otherwise, the rates of change of the various scores would correlate. Furthermore, additional analyses indicated, as expected for the FSIQ (and the GAI), that children who were assessed twice with a shorter retest interval benefited slightly more from prior administrations than did children who were assessed twice with a larger retest interval. This finding is consistent with the data obtained with short-term retest studies and with long-term retest studies. We also found that for the PRI, PSI, VCI, and FSIQ, the gain from the first to the second assessment was significantly and positively associated with age at initial testing. Older children at the first assessment tended to benefit slightly more from prior administrations of the WISC-IV than did younger children. This finding is not consistent with the data reported by Flanagan and Kaufman (2009) because they found that practice effects were largest for ages 6 to 7 and become smaller with increasing age. Finally, we found that children with higher initial level tended to present a smaller increase at the second assessment, while children with lower initial level tended to present a higher increase at the second assessment. This finding was not consistent with Ryan's and colleagues (2010), who found a positive correlation between retest gain and initial level. We assumed that this finding reflects the statistical phenomenon of the "regression to the mean" (Barnett, van der Pols, & Dobson, 2005). The scores at the second assessment are less extreme than those at the first assessment. Altogether, the present findings demonstrated that the magnitude of the retest effects depends on the initial age, the initial level of performance and the retest interval. However, these findings should be

considered only as tendencies because there was a significant correlation between age at the first testing and the length of the retest interval ($r = .26$).

The present study provides useful information regarding the long-term stability of the WISC-IV index scores. However, the current research has some limitations. Indeed, a generalization of the data might be limited because our sample came from a specific Swiss state (i.e., Geneva), and thus is not representative of the French-speaking population. Moreover, our results involve children restricted to the age group between seven to twelve years. Finally, the present study was not designed to investigate the presence and extent of examiner bias variance. However, Mcdermott, Rhoad, and Watkins (2014) found that all WISC-IV scores – and especially the FSIQ and the VCI – conveyed substantial amounts of variation that were not due to examinees' individual differences but to examiners. To prevent potential bias in assessment, the six psychologists were trained to administer the subtests in the same way. Careful consideration was also given to a consensus in the scoring procedure. It should also be noted that our study does not completely allow us to distinguish “retest effect” and “age effect” because of the significant correlation between age at the first assessment and retest interval. However, the correlation was relatively low ($r = .26$); thus, we assumed that it does not completely invalid the results. Furthermore, partial correlations were conducted. When the influence of age was controlled, correlations between the gain / loss and the duration of the retest interval were significant and negative for all index scores (range from $-.22$ for $\Delta_{FSIQ} \times \Delta_T$ to $-.13$ for $\Delta_{VCI} \times \Delta_T$), except the WMI. When the influence of the retest interval was controlled, correlations between the gain / loss and age at initial testing were significant and positive for all index scores (range from $.16$ for $\Delta_{PRI} \times A1$ to $.25$ for $\Delta_{FSIQ} \times A1$), except for the VCI.

Regardless of these limits, similar findings were found with clinical and non-clinical samples, and with U.S. and non-U.S. children. Given the consistency across studies, the

present results supplement and extend the previous conclusions of Ryan and colleagues (2010), Lander (2010), Watkins and Smith (2013), and Bartoi and colleagues (2015): the longer retest intervals, the lower stability coefficients will be, and more particularly for the stability at an individual level.

6. Conclusion

The current study has important implications for psychological practice. Because the GAI is relatively stable at an interindividual level and at an individual level, this index score might be the most useful for predictions. In contrast, our results showed that the FSIQ was less stable at the individual level (idiographic perspective). However, more studies are required to demonstrate the prediction validity of the GAI. Based on an idiographic perspective, it could be argued that two CHC-factors – Gc and Gv – could also be used for predictions. However, from a nomothetic perspective, the stability of Gc and Gv was not sufficient for important individual decisions according to Nunnally and Bernstein, or to Charter and Feldt ($r = .73$ and $r = .78$, respectively). Thus, the present study indicated that caution must be exercised when interpreting the FSIQ, index scores, or subtest scores for diagnostic decisions. When Deary et al. (2000) stated, “*in the absence of disease processes, we might expect broad stability of individual differences in mental abilities across the human lifespan*” (p. 54), they considered only interindividual differences, not intraindividual stability. It is not adequate to assume that the WISC-IV scores will remain stable across time for *all* children, and that intelligence is such a steady and enduring trait across time.

Conflict of interest: none.

References

- Abell, N., Springer, D. W., & Kamata, A. (2009). *Developing and validating rapid assessment instruments*. New York, NY: Oxford University Press.
- Aickin, M., & Gensler, H. (1996). Adjusting for multiple testing when reporting research results: The Bonferroni vs Holm methods. *American Journal of Public Health*, 86, 726–728. <http://doi.org/10.2105/AJPH.86.5.726>
- American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2014). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association
- Barnett, A. G., van der Pols, J. C., & Dobson, A. J. (2005). Regression to the mean: What it is and how to deal with it. *International Journal of Epidemiology*, 34, 215–220. <http://doi.org/10.1093/ije/dyh299>
- Bartoi, M., Issner, J. B., Hettterscheidt, 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://doi.org/10.1080/21622965.2013.811075>
- Bauman, E. E. (1991). Stability of WISC-R scores in children with learning difficulties. *Psychology in the Schools*, 28, 95–100. [http://doi.org/10.1002/1520-6807\(199104\)28:2<95::AID-PITS2310280203>3.0.CO;2-9](http://doi.org/10.1002/1520-6807(199104)28:2<95::AID-PITS2310280203>3.0.CO;2-9)
- Benedict, R. H. B., & Zgaljardic, D. J. (1998). Practice effects during repeated administrations of memory tests with and without alternate forms. *Journal of Clinical and Experimental Neuropsychology*, 20, 339–352. <http://doi.org/10.1076/jcen.20.3.339.822>

- Bremner, D., McTaggart, B., Saklofske, D., & Janzen, T. (2011). WISC-IV GAI and CPI in psychoeducational assessment. *Canadian Journal of School Psychology, 26*, 209–219. <http://doi.org/10.1177/0829573511419090>
- Calamia, M., Markon, K., & Tranel, D. (2012). Scoring higher the second time around: Meta-analyses of practice effects in neuropsychological assessment. *The Clinical Neuropsychologist, 26*, 543–570. <http://doi.org/10.1080/13854046.2012.680913>
- Canivez, G. L., & Watkins, M. W. (1998). Long-term stability of the Wechsler Intelligence Scale for Children–Third Edition. *Psychological Assessment, 10*, 285–291. <http://doi.org/10.1037/1040-3590.10.3.285>
- 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., & Feldt, L. S. (2001). Meaning of reliability in terms of correct and incorrect clinical decisions: The art of decision making is still alive. *Journal of Clinical and Experimental Neuropsychology, 23*, 530–537.
- Chen, H.-Y., Keith, T. Z., Chen, Y.-H., & Chang, B.-S. (2009). What does the WISC-IV measure? Validation of the scoring and CHC-based interpretative approaches. *Journal of Research in Education Sciences, 54*, 85–108.
- Cohen, J. (1977). *Statistical power analysis for the behavioral sciences* (rev.). Mahwah, NJ: Lawrence Erlbaum Associates, Inc.
- Cohen, J. (1992). A power primer. *Psychological Bulletin, 112*, 155–159. <http://doi.org/10.1037/0033-2909.112.1.155>

- 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://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://doi.org/10.1016/S0160-2896\(99\)00031-8](http://doi.org/10.1016/S0160-2896(99)00031-8)
- Dikmen, S. S., Heaton, R. K., Grant, I., & Temkin, N. R. (1999). Test–retest reliability and practice effects of Expanded Halstead–Reitan Neuropsychological Test Battery. *Journal of the International Neuropsychological Society*, 5, 346–356.
<http://doi.org/10.1017/S1355617799544056>
- 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://doi.org/10.1080/13854046.2012.659219>
- Ferrer, E., Salthouse, T. A., McArdle, J. J., Stewart, W. F., & Schwartz, B. S. (2005). Multivariate modeling of age and retest in longitudinal studies of cognitive abilities. *Psychology and Aging*, 20, 412–422. <http://doi.org/10.1037/0882-7974.20.3.412>
- Flanagan, D. P., & Kaufman, A. S. (2009). *Essentials of WISC®-IV assessment* (2nd ed.). Hoboken, NJ: John Wiley & Sons, Inc.
- Gaetano, J. (2013). *Holm-Bonferroni sequential correction: An EXCEL calculator-ver. 1.2*.
- Ghisletta, P., & Lecerf, T. (n.d.). Intelligence, crystallized. In S. K. Whitbourne & D. Simonton (Eds.), *The encyclopedia of adulthood and aging*. Hoboken, NJ: Wiley-Blackwell.

- Golay, P., Reverte, I., Rossier, J., Favez, N., & Lecerf, T. (2013). Further insights on the French WISC–IV factor structure through Bayesian structural equation modeling. *Psychological Assessment*, 25, 496–508. <http://doi.org/10.1037/a0030676>
- Guilford, J. P., & Fruchter, B. (1978). *Fundamental statistics in psychology and education* (6th ed.). New York, NY: McGraw Hill Higher Education.
- Hausknecht, J. P., Halpert, J. A., Di Paolo, N. T., & Moriarty Gerrard, M. O. (2007). Retesting in selection: A meta-analysis of coaching and practice effects for tests of cognitive ability. *Journal of Applied Psychology*, 92, 373–385. <http://doi.org/10.1037/0021-9010.92.2.373>
- Hertzog, C., & Schaie, K. W. (1986). Stability and change in adult intelligence: I. Analysis of longitudinal covariance structures. *Psychology and Aging*, 1, 159–171. <http://doi.org/10.1037/0882-7974.1.2.159>
- Holm, S. (1979). A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, 6, 65–70. <http://doi.org/10.2307/4615733>
- Huteau, M. (1985). *Les conceptions cognitives de la personnalité*. Paris: Presses Universitaires de Paris.
- Keith, T. Z., Fine, J. G., Taub, G. E., Reynolds, M. R., & Kranzler, J. H. (2006). Higher order, multisample, confirmatory factor analysis of the Wechsler Intelligence Scale for Children–Fourth Edition: What does it measure? *School Psychology Review*, 35, 108–127.
- Kieng, S., Rossier, J., Favez, N., Geistlich, S., & Lecerf, T. (2015). Stabilité à long terme des scores du WISC-IV : forces et faiblesses personnelles. *Pratiques Psychologiques*, 21, 137–154. <http://doi.org/doi: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. *Dissertation Abstracts International: Section A. Humanities and Social Sciences*.
- Lecerf, T., Bovet-Boone, F., Peiffer, E., Kieng, S., & Geistlich, S. (2016). WISC-IV GAI and CPI profiles in healthy children and children with learning disabilities. *Revue Européenne de Psychologie Appliquée/European Review of Applied Psychology*, 66, 101–107. <http://doi.org/10.1016/j.erap.2016.04.001>
- Lecerf, T., Golay, P., Reverte, I., Senn, D., Favez, N., & Rossier, J. (2012). Scores composites CHC pour le WISC-IV : Normes francophones. *Pratiques Psychologiques*, 18, 37–50. <http://doi.org/10.1016/j.prps.2011.04.001>
- Lecerf, T., Kieng, S., & Geistlich, S. (2015). Cohésion–non-cohésion des scores composites : Valeurs seuils et interprétabilité. L'exemple du WISC-IV. *Pratiques Psychologiques*, 21, 155–171. <http://doi.org/10.1016/j.prps.2015.02.001>
- Lecerf, T., Reverte, I., Coleaux, L., Maillard, F., Favez, N., & Rossier, J. (2011). Indice d'aptitude général et indice de compétence cognitive pour le WISC-IV : Normes empiriques versus normes statistiques. *European Review of Applied Psychology/Revue Européenne de Psychologie Appliquée*, 61, 115–122. <http://doi.org/10.1016/j.erap.2011.01.001>
- Lecerf, T., Rossier, J., Favez, N., Reverte, I., & Coleaux, L. (2010). The four- vs. alternative six-factor structure of the French WISC-IV: Comparison using confirmatory factor analyses. *Swiss Journal of Psychology/Schweizerische Zeitschrift Für Psychologie/Revue Suisse de Psychologie*, 69, 221–232. <http://doi.org/10.1024/1421-0185/a000026>

- 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. <http://doi.org/10.1037/h0088959>
- Magnusson, D. (1967). *Test theory*. Reading, MA: Addison-Wesley .
- McCall, R. B. (1977). Childhood IQ's as predictors of adult educational and occupational status. *Science*, 197, 482–483. <http://doi.org/10.1126/science.197.4302.482>
- 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, 207–214. <http://doi.org/10.1037/a0034832>
- Nunnally, J. C., & Bernstein, I. H. (2010). *Psychometric theory* (3rd ed.). New York, NY: Tata 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. [http://doi.org/10.1002/1520-6807\(198804\)25:2<118::AID-PITS2310250204>3.0.CO;2-T](http://doi.org/10.1002/1520-6807(198804)25:2<118::AID-PITS2310250204>3.0.CO;2-T)
- Prifitera, A., Saklofske, D. H., & Weiss, L. G. (2008). *WISC-IV clinical assessment and intervention*. San Diego, CA: Elsevier Academic Press.
- Rémy, L., & Gilles, P.-Y. (2014). Relationship between field dependence-independence and the g factor: What can problem-solving strategies tell us? *Revue Européenne de Psychologie Appliquée/European Review of Applied Psychology*, 64, 77–82. <http://doi.org/10.1016/j.erap.2014.02.001>
- Reverte, I., Golay, P., Favez, N., Rossier, J., & Lecerf, T. (2014). Structural validity of the

Wechsler Intelligence Scale for Children (WISC-IV) in a French-speaking Swiss sample. *Learning and Individual Differences*, 29, 114–119.

<http://doi.org/10.1016/j.lindif.2013.10.013>

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://doi.org/10.1080/09084280903297933>

Saklofske, D. H., Zhu, J., Coalson, D. L., Raiford, S. E., & Weiss, L. G. (2010). Cognitive Proficiency Index for the Canadian Edition of the Wechsler Intelligence Scale for Children-Fourth Edition. *Canadian Journal of School Psychology*, 25, 277–286.

<http://doi.org/10.1177/0829573510380539>

Salthouse, T. A. (2009). When does age-related cognitive decline begin? *Neurobiology of Aging*, 30, 507–514. <http://doi.org/10.1016/j.neurobiolaging.2008.09.023>

Salthouse, T. A., Schroeder, D. H., & Ferrer, E. (2004). Estimating retest effects in longitudinal assessments of cognitive functioning in adults between 18 and 60 years of age. *Developmental Psychology*, 40, 813–822. <http://doi.org/10.1037/0012-1649.40.5.813>

Salvia, J., Ysseldyke, J. E., & Bolt, S. (2012). *Assessment in special and inclusive education* (12th ed.). Belmont, CA: Wadsworth Publishing.

Schwartzman, A. E., Gold, D., Andres, D., Arbuckle, T. Y., & Chaikelson, J. (1987).

Stability of intelligence: A 40-year follow-up. *Canadian Journal of Psychology/Revue Canadienne de Psychologie*, 41, 244–256. <http://doi.org/10.1037/h0084155>

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.

[http://doi.org/10.1002/1520-6807\(199004\)27:2<101](http://doi.org/10.1002/1520-6807(199004)27:2<101)

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

Truscott, S. D., Narrett, C. M., & Smith, S. E. (1994). WISC—R subtest reliability over time: Implications for practice and research. *Psychological Reports*, 74, 147–156.

<http://doi.org/10.2466/pr0.1994.74.1.147>

Vance, H. B., Blixt, S., Ellis, R., & Debell, S. (1981). Stability of the WISC-R for a sample of exceptional children. *Journal of Clinical Psychology*, 37, 397–399.

[http://doi.org/10.1002/1097-4679\(198104\)37:2<397](http://doi.org/10.1002/1097-4679(198104)37:2<397)

Voelkle, M. C., Brose, A., Schmiedek, F., & Lindenberger, U. (2014). Toward a unified framework for the study of between-person and within-Person structures: Building a bridge between two research paradigms. *Multivariate Behavioral Research*, 49, 193–213. <http://doi.org/10.1080/00273171.2014.889593>

Wasserman, J. D., & Bracken, B. A. (2013). Fundamental psychometric considerations in assessment. In J. R. Graham & J. A. Naglieri (Eds.), *Handbook of Psychology* (pp. 50–81). Hoboken, NJ, USA: John Wiley & Sons, Inc.

Watkins, M. W., & Canivez, G. L. (2004). Temporal stability of WISC-III subtest composite: strengths and weaknesses. *Psychological Assessment*, 16, 133–138.

<http://doi.org/10.1037/1040-3590.16.2.133>

Watkins, M. W., & Smith, L. G. (2013). Long-term stability of the Wechsler Intelligence Scale for Children—Fourth Edition. *Psychological Assessment*, 25, 477–483.

<http://doi.org/10.1037/a0031653>

Wechsler, D. (2005a). *WISC-IV – Manuel d'administration et de cotation*. Paris: Editions du Centre de Psychologie Appliquée.

Wechsler, D. (2005b). *WISC-IV – Manuel d'interprétation*. Paris: Editions du Centre de Psychologie Appliquée.

Williams, P. E., Weiss, L. G., & Rolfhus, E. L. (2003). *WISC-IV technical report# 2: Psychometric properties. The Psychological Corporation's WISC-IV Technical Manual* (Vol. 2). Retrieved from <http://www.pearsonassess.ca/content/dam/ani/clinicalassessments/ca/programs/pdfs/wisc-iv-technical-report-number-2-psychometric-properties.pdf>

Table 1

Test-Retest Coefficients, Mean Difference Scores (Retest – Test), and Effect Sizes for Five Stability Studies (French and American samples)

Scale	Wechsler (2005)				Ryan, Glass, and Bartels (2010)				Bartoi et al. (2015)		Lander (2010)			Watkins and Smith (2013)				
	(27 days; $N = 93$)				(10.88 months; $N = 43$)				(1.84 year; $N = 51$)		(2.89 years; $N = 131$)			(2.84 years; $N = 344$)				
	r_{12} [95% CI]	r_c	ΔM	d	r_{12} [95% CI]	r_c	ΔM	d	r_{12} [95% CI]	ΔM	r_{12} [95% CI]	ΔM	d	r_{12} [95% CI]	r_c	r_t	ΔM	d
Block Design	.72 [.61, .81]	.81	1.6	0.61	.67 [.46, .81]	.72	0.56	0.21	.78 [.64, .87]]	-0.14	.62 [.50, .72]	-0.40	-0.16	.70 [.64, .75]	.72	.59	-0.50	-0.18
Similarities	.77 [.67, .84]	.77	0.7	0.21	.63 [.41, .78]	.66	0.51	0.17	.80 [.67, .88]	-0.29	.48 [.34, .60]	0.57	0.26	.58 [.51, .65]	.63	.53	0.40	0.15
Digit Span	.59 [.44, .71]	.71	0.8	0.27	.74 [.57, .85]	.76	0.30	0.11	.63 [.43, .77]	0.10	.46 [.31, .59]	-0.14	-0.05	.60 [.53, .66]	.65	.55	-0.17	-0.07
Picture Concepts	.57 [.41, .69]	.64	0.9	0.34	.22 [.00, .49]	.26	0.17	0.07	.54 [.31, .71]	0.04	.44 [.29, .57]	0.38	0.13	.46 [.37, .54]	.43	.42	0.58	0.19.
Coding	.75 [.64, .83]	.72	2.2	0.72	.42 [.14, .64]	.40	0.24	0.08	.52 [.29, .70]	-0.18	.46 [.31, .59]	-0.90	-0.40	.52 [.44, .59]	.50	.48	-0.90	-0.30
Vocabulary	.80 [.71, .86]	.82	0.4	0.17	.81 [.67, .89]	.84	-0.02	-0.01	.81 [.69, .89]	-0.16	.56 [.43, .67]	-0.49	-0.22	.69 [.63, .74]	.73	.63	-0.15	-0.06
Letter-Num. Seq.	.66 [.53, .76]	.71	0.4	0.14	.49 [.22, .69]	.52	0.30	0.11	.35 [.08, .57]	-0.14	.31 [.15, .46]	0.08	0.01	.48 [.39, .56]	.51	.46	0.18	0.06
Matrix Reasoning	.68 [.55, .78]	.67	1.0	0.33	.40 [.11, .63]	.47	-0.53	-0.21	.74 [.58, .84]	-0.45	.48 [.34, .60]	-0.25	-0.10	.63 [.56, .69]	.64	.59	0.12	0.04
Comprehension	.80 [.71, .86]	.79	0	-0.01	.49 [.22, .69]	.47	-0.23	-0.08	.49 [.25, .67]	0.50	.55 [.42, .66]	0.19	0.23	.48 [.39, .56]	.53	.43	0.09	0.03
Symbol Search	.70 [.58, .79]	.83	1.8	0.63	.41 [.13, .63]	.50	0.31	0.31	.62 [.42, .76]	0.04	.28 [.11, .43]	0.05	0.02	.54 [.46, .61]	.51	.47	0.25	0.08
Verbal Comprehension	.88 [.82, .92]	.88	2.1	0.14	.75 [.58, .86]	.76	0.25	0.02	.81 [.66, .90]	0.02	.65 [.54, .74]	0.64	0.06	.72 [.66, .77]	.78	.69	0.55	0.04
Perceptual Reasoning	.80 [.71, .86]	.83	8.2	0.59	.58 [.34, .75]	.68	0.42	0.04	.79 [.62, .89]	-1.16	.62 [.50, .72]	-0.30	-0.02	.76 [.71, .80]	.76	.71	0.37	0.03
Working Memory	.73 [.62, .81]	.78	3.6	0.24	.73 [.55, .85]	.75	2.05	0.14	.60 [.34, .78]	-0.06	.54 [.41, .65]	-0.45	-0.04	.66 [.60, .72]	.70	.62	-0.17	-0.01
Processing Speed	.79 [.70, .86]	.83	12.0	0.81	.49 [.22, .69]	.54	2.32	0.20	.58 [.31, .76]	-0.79	.52 [.38, .63]	-2.14	-0.18	.65 [.58, .71]	.65	.60	-1.90	-0.13
Full Scale IQ	.89 [.84, .93]	.91	8.3	0.59	.80 [.66, .89]	.88	1.63	0.15	.86 [.74, .93]	-0.60	.70 [.60, .78]	-0.55	-0.06	.82 [.78, .85]	.84	.73	-0.12	-0.01

Note. WISC-IV = Wechsler Intelligence Scale for Children – 4th edition; r_{12} = uncorrected test-retest correlation; r_c = correlation corrected for the

variability of the standardization sample; r_t = correlation corrected for combined or additive effect of content and time sampling error; d = standardized mean difference.

Table 2

Descriptive Statistics, Stability Coefficients, and Effect Sizes

Scale	Test		Retest		ΔM	r_{12} [95% CI]	SE	r_c [95% CI]	r_t	d
	M	SD	M	SD						
<i>Ages 7-12 (N = 277)</i>										
Block Design	10.97	3.02	10.79	3.07	-0.18	.74 [.68, .79]	0.04	.73 [.67, .78]	.62	-0.08
Similarities	11.13	3.35	11.25	2.92	0.12	.67 [.60, .73]	0.04	.59 [.51, .66]	.54	0.05
Digit Span	9.37	2.72	9.68	2.75	0.32	.58 [.50, .65]	0.05	.65 [.58, .71]	.54	0.13
Picture Concepts	9.38	2.70	9.70	2.37	0.31	.40 [.30, .49]	0.06	.51 [.42, .59]	.32	0.11
Coding	9.77	2.92	10.58	2.56	0.82*	.56 [.47, .64]	0.05	.59 [.51, .66]	.46	0.32
Vocabulary	10.57	2.83	10.61	2.84	0.04	.71 [.65, .76]	0.04	.74 [.68, .79]	.59	0.02
Letter-Number Sequencing	9.03	3.03	9.38	2.83	0.35	.42 [.32, .51]	0.05	.40 [.30, .49]	.37	0.11
Matrix Reasoning	9.29	2.76	9.78	2.54	0.48*	.56 [.47; .64]	0.05	.63 [.55, .70]	.50	0.19
Comprehension	10.80	2.70	10.47	2.94	-0.33	.59 [.51, .66]	0.05	.67 [.60, .73]	.48	-0.13
Symbol Search	10.71	2.74	11.80	2.90	1.09*	.47 [.37, .56]	0.05	.55 [.46, .63]	.45	0.38
Picture Completion ^a	9.75	2.49	9.74	2.67	-0.01	.57 [.48, .65]	.05	.70 [.63, .76]	.51	0.01
Average subtest	10.07	2.84	10.34	2.76	0.27	.57		.63	.50	
Verbal Comprehension	104.90	15.23	104.70	15.17	-0.23	.80 [.75, .84]	.04	.79 [.74, .83]	.72	-0.02
Perceptual Reasoning	99.10	14.54	100.35	13.60	1.26	.73 [.67, .78]	.04	.74 [.68, .79]	.65	0.12

Scale	Test		Retest		ΔM	r_{12} [95% CI]	SE	r_c [95% CI]	r_t	d
	M	SD	M	SD						
Working Memory	95.18	14.10	97.20	14.33	2.02*	.61 [.53, .68]	.05	.66 [.59, .72]	.58	0.16
Processing Speed	101.38	13.88	107.20	14.03	5.79*	.62 [.54, .69]	.05	.67 [.60, .73]	.57	0.48
Average standard index	100.14	14.45	102.36	14.28	2.21	.68		.71	.62	
Full Scale IQ	100.80	13.93	103.30	12.78	2.53*	.80 [.75, .84]	.04	.83 [.79, .86]	.76	0.30
General Ability	102.60	14.62	103.10	13.43	0.51	.82 [.78, .86]	.03	.83 [.79, .86]	.76	0.06
Cognitive Proficiency	97.87	13.76	102.70	13.63	4.87*	.66 [.59, .72]	.05	.71 [.65, .76]	.61	0.43
Crystallized Intelligence (Gc)	105.49	14.98	104.92	14.83	-0.57	.73 [.67, .78]	.04	.73 [.67, .78]	.63	-0.05
Fluid Intelligence (Gf)	95.96	13.79	98.04	13.47	2.08*	.53 [.44, .61]	.05	.60 [.52, .67]	.45	0.16
Visual Processing (Gv) ^a	101.83	14.10	101.32	14.60	-0.51	.75 [.69, .80]	.04	.78 [.73, .82]	.67	-0.04
Short-Term Working Memory (Gwm)	95.17	14.12	97.18	14.38	2.01*	.61 [.53, .68]	.05	.66 [.59, .72]	.58	0.16
Processing Speed (Gs)	101.40	13.61	106.90	13.39	5.51*	.62 [.54, .69]	.05	.69 [.62, .75]	.57	0.47
Average CHC index	99.97	14.12	101.67	14.13	1.70	.62		.72	.58	

Note. WISC-IV = Wechsler Intelligence Scale for Children – Fourth Edition; Average index = mean for scores and median for coefficients; ΔM = mean score difference (Retest – Test); r_{12} = uncorrected test-retest correlation; SE = standard error of the correlation; r_c = correlation corrected for the variability of the standardization sample; r_t = correlation corrected for combined or additive effect of content and time sampling error; 95% CI = 95% confidence interval; d = standardized mean difference.

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

^a $N = 248$.

Table 3

Cumulative Frequency Distributions (in Percentages) of WISC-IV Index Scores

<i>T2-T1</i>	VCI	PRI	WMI	PSI	FSIQ	GAI	CPI
-30			0.4				
-29		0.4					
-28	0.4						
-27			1.1	0.4			0.4
-25		0.7			0.4	0.4	0.7
-24		1.1	2.2	1.1	0.7		1.1
-23			4.0			0.7	
-22	0.7	1.4		1.8		1.4	
-21	1.4	1.8				1.8	
-20	1.8	2.9					1.4
-19		3.2		2.9			
-18	3.2	3.6	6.9	3.6		2.9	1.8
-17	4.0	4.3					2.2
-16	5.1			4.0	1.1	3.2	2.5
-15	5.4	6.5	10.1	5.8	1.8	4.7	4.0
-14	7.6	7.6			2.9	5.1	4.3
-13	9.7	7.9		8.3	3.2	5.8	5.4
-12	12.6	9.7	13.0	9.4	4.0	7.2	6.5
-11	14.4	11.9			5.8	8.3	7.9
-10	16.6	15.9	13.4	11.6	6.9	9.7	9.7
-9	19.1	17.3	22.0	14.1	9.0	10.8	11.2
-8	22.4	18.4		14.8	12.3	14.4	13.0
-7	26.4	20.2		16.6	14.4	16.6	14.4
-6	28.9	22.0	29.2	20.6	17.7	22.4	20.2
-5	30.0	27.1		20.9	20.9	26.0	22.4
-4	37.2	30.3	29.6	22.4	22.7	28.9	24.2

-3	39.4	34.7	40.8	26.4	27.1	34.7	25.6
-2	49.8	45.1	41.2	27.1	30.7	40.1	30.7
-1	50.2				35.4	47.7	31.4
0	57.0	53.4	51.6	37.2	41.9	54.9	35.0
1	58.1				48.0	58.5	36.8
2	64.6	57.8		37.5	51.6	61.7	43.0
3	65.0	61.0	58.8	42.6	55.2	65.7	44.8
4	69.0	65.0		44.0	58.8	70.4	49.1
5	71.5	67.1		44.4	63.5	73.3	52.0
6	75.5	71.8	66.4	49.1	68.2	77.3	55.2
7	79.1	73.3		52.0	71.1	81.2	59.6
8	83.8	74.7		52.7	76.5	84.5	64.3
9	85.6	78.7	76.9	57.8	79.8	85.9	66.8
10	87.4	81.6		64.3	83.0	87.7	69.3
11	88.4	81.9			86.3	89.5	73.3
12	91.0	84.8	81.6	67.9	88.1	91.3	75.8
13	92.4	86.6		71.1	89.5	93.1	78.7
14	94.2	89.5	81.9	71.8	93.5	94.6	80.5
15	96.0	91.3	88.8	79.1	95.3	95.3	82.3
16	97.5	91.7		85.2	96.0	96.8	84.8
17		93.9		86.3		97.5	86.6
18		94.9	93.9	87.4		98.6	88.4
19	98.2	96.0		89.2	96.8	98.9	91.3
20		96.4		89.5	97.8	99.3	93.5
21		96.8	94.6	89.9			94.6
22				92.1	98.2		94.9
23		97.5		92.8	98.6		96.0
24	98.6	97.8	96.8	93.5	99.3		
25		98.2		96.4	100		96.4
26	98.9					99.6	

27		99.3	97.8	96.8		97.5
28	99.3	99.6		97.5		
29				98.2		
30			98.9			97.8
31		100		98.9		98.2
32	99.6					98.9
33			99.3	99.3		
34	100			99.6	100	99.3
35						99.6
36				100		100
48			100			

Note. Column entries represent cumulative percentages of children's change in performance

between both assessments. WISC-IV = Wechsler Intelligence Scale for Children – Fourth Edition; Δ = mean score difference (Retest – Test); VCI = Verbal Comprehension Index; PRI = Perceptual Reasoning Index; WMI = Working Memory Index; PSI = Processing Speed Index; FSIQ = Full Scale IQ; GAI = General Ability Index; and ICC = Cognitive Proficiency Index.