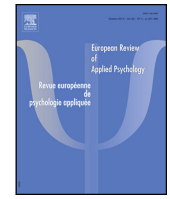
G Model
ERAP-401; No. of Pages 10

ARTICLE IN PRESS

Revue européenne de psychologie appliquée xxx (2016) xxx–xxx

Disponible en ligne sur
ScienceDirect
www.sciencedirect.com

Elsevier Masson France
EM|consulte
www.em-consulte.com/en



Original article

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

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

S. Kieng[a,b,*], J. Rossier[c], N. Favez[a,b], T. Lecerf[a,b]

[a] *Faculty of Psychology and Educational Sciences (FPSE), University of Geneva, 40, boulevard du Pont-d'Arve, 1205 Geneva, Switzerland*
[b] *Distance Learning University, Sierre, Switzerland*
[c] *Institute of Psychology, University of Lausanne, Quartier UNIL-Mouline, bâtiment Geopolis 2635, 1015 Lausanne, Switzerland*

ARTICLE INFO

Article history:
Received 15 December 2015
Received in revised form 30 August 2016
Accepted 4 October 2016

Keywords:
WISC-IV
Long-term stability
Reliability
CHC
Intelligence

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 (± 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.

Mots clés :
WISC-IV
Stabilité à long terme
Fidélité
CHC
Intelligence

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 4e 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.

* Corresponding author at: FPSE, Psychology, University of Geneva, 40, boulevard du Pont-d'Arve, 1205 Geneva, Switzerland.
E-mail address: sotta.kieng@unige.ch (S. Kieng).

http://dx.doi.org/10.1016/j.erap.2016.10.001
1162-9088/

Please cite this article in press as: Kieng, S., et al. Long-term stability of the French WISC-IV: Standard and CHC index scores. *Rev. Eur. Psychol. Appl.* (2016), http://dx.doi.org/10.1016/j.erap.2016.10.001

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 (± 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.

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; 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; 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*

[1] To compare different measures, the practice effect is expressed in Time 1 standard deviation units (i.e., T_1 *SD* units). The T_1 *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., $T_1 = SD$ units $= \frac{T_2 - T_1}{SD \text{ of } T_1}$).