NeuroRehabilitation 41 (2017) 581–592
DOI:10.3233/NRE-172275
IOS Press

Methodology for the development of normative data for Spanish-speaking pediatric populations

D. Rivera[a] and J.C. Arango-Lasprilla[a,b,*]
[a]*BioCruces Health Research Institute, Cruces University Hospital, Barakaldo, Spain*
[b]*IKERBASQUE, Basque Foundation for Science, Bilbao, Spain*

Abstract.
OBJECTIVE: To describe the methodology utilized to calculate reliability and the generation of norms for 10 neuropsychological tests for children in Spanish-speaking countries.
METHOD: The study sample consisted of over 4,373 healthy children from nine countries in Latin America (Chile, Cuba, Ecuador, Guatemala, Honduras, Mexico, Paraguay, Peru, and Puerto Rico) and Spain. Inclusion criteria for all countries were to have between 6 to 17 years of age, an Intelligence Quotient of $\geq$80 on the Test of Non-Verbal Intelligence (TONI-2), and score of <19 on the Children's Depression Inventory. Participants completed 10 neuropsychological tests. Reliability and norms were calculated for all tests.
RESULTS: Test-retest analysis showed excellent or good- reliability on all tests (r's>0.55; p's<0.001) except M-WCST perseverative errors whose coefficient magnitude was fair. All scores were normed using multiple linear regressions and standard deviations of residual values. Age, age^2, sex, and mean level of parental education (MLPE) were included as predictors in the models by country. The non-significant variables ($p > 0.05$) were removed and the analysis were run again.
CONCLUSIONS: This is the largest Spanish-speaking children and adolescents normative study in the world. For the generation of normative data, the method based on linear regression models and the standard deviation of residual values was used. This method allows determination of the specific variables that predict test scores, helps identify and control for collinearity of predictive variables, and generates continuous and more reliable norms than those of traditional methods.

Keywords: Methodology, Spanish-speaking, normative data, pediatric population

1. Introduction

Neuropsychology is a discipline of psychology that studies cognitive processes and their relationship with behavior in both healthy populations and individuals with brain injuries and/or pathologies (Lezak, Howieson, & Loring, 2004). One of the most important specialties within this discipline is child neuropsychology, which is responsible for studying behavior in relation to the development of different brain structures and systems in children and adolescents (APA, Division 40, 2001).

There are currently a number of disorders that affect the development and functioning of cognitive processes in children, such as autism, attention deficit/hyperactivity disorder, specific learning disorders, motor and behavioral disorders, traumatic brain injury, cerebral palsy, and other neurodevelopmental conditions (American Psychiatric Association,

*Address for correspondence: Juan Carlos Arango Lasprilla, Ph.D., BioCruces Health Research Institute, Cruces University Hospital, IKERBASQUE, Basque Foundation for Science, Plaza de Cruces s/n. 48903, Barakaldo, Bizkaia, Spain. Tel.: +34 946006000 /Ext. 7963; E-mail: jcalasprilla@gmail.com.

2014). Due to the variety of symptoms associated with each of these disorders, clinicians perform neuropsychological evaluations to accurately diagnose them, document natural changes over time, and track improvements associated with specific interventions.

The use of neuropsychological tests during assessment is essential to obtaining an objective measure of the child's cognitive functioning in different domains, such as memory, attention, language, visuospatial skills, executive functioning, motor skills, and behavioral and emotional functioning (Baron, 2004; Lezak et al., 2004; Strauss, Sherman, & Spreen, 2006). Because test performance is known to be influenced by many factors, having normative data adjusted to the sociodemographic (e.g. age, education, sex) and cultural characteristics of individuals is of vital importance in order to adequately interpret a person's score on a particular neuropsychological test (Lezak, et al., 2004; Strauss et al., 2006; Van der Elst, Molenberghs, Van Boxtel, & Jolles, 2013). For this reason, researchers from different countries like the United States of America (Goodman, Delis, & Mattson, 2010), Korea (Kim & Na, 2008), Taiwan (Shu, Tien, Lung, & Chang, 2000), Italy (Cianchetti, Corona, Foscoliano, Conty, & Sannio-Fancello, 2007), Holland (Hiuzinga & Smidts, 2010), Australia (Davies, Field, Andersen, & Pestell, 2011), Portugal (Townes, Martins, Castro-Caldas, Rosenbaum, & Derouen, 2008), Israel (Vakil, Greenstein, & Blachstein, 2010), and Iran (Yousefi, et al., 1992), among others, have established normative data for some of the primary neuropsychological tests used with children and adolescents.

In Latin America and Spain, some of the neuropsychological tests most commonly used with children and adolescents are the Trail Making Test, the Digit and Symbols Test, the Stroop Color-Word Interference Test, the Wisconsin Card Sorting Test, the California Verbal Learning Test, the Rey Complex Figure Test, The Boston Naming Test, the Peabody Picture Vocabulary Test, and the Verbal Fluency Test (Arango-Lasprilla, Stevens, Morlett Paredes, Ardila, & Rivera, 2016; Olabarrieta-Landa et al., 2016). Unfortunately, despite their popular use, many of these tests have not been validated in research and lack country-specific normative data for the majority of these Spanish-speaking countries. Because of this, clinical neuropsychologists report using: 1) normative data from other countries, 2) personalized procedures through clinical practice, and 3) raw scores without a comparison to normative data (Arango-Lasprilla et al., 2016; Olabarrieta-Landa et al., 2016). Therefore, the objective of this study was to describe the methodological procedure used to develop normative data for 10 commonly used neuropsychological tests in a group of children and adolescents (6 to 17 years of age) from nine Latin American countries and Spain.

2. Method

2.1. Sample

In the present study consisted of neuropsychological evaluations of 6,030 clinically healthy children and adolescents from ten Latin American countries (Chile, Colombia, Cuba, Ecuador, Guatemala, Honduras, Mexico, Paraguay, Peru, and Puerto Rico) and Spain. This article will present the methodology used to generate normative data for this study. Since the normative data for Colombia ($n = 1{,}657$) are published in Arango-Lasprilla, Rivera, and Olabarrieta-Landa (2017), the sample for the present study consisted of 4,373 children and adolescents ages 6 to 17 who were evaluated in 21 centers in the 9 other Latin American countries and Spain (see Table 1). The demographic characteristics (age, sex, type of school, and parent education) by country can be found in Table 2.

All participants met the following inclusion criteria: a) being between the ages of 6 and 17 years old, b) being born and currently living in the country where the protocol was administered, c) having Spanish as primary language, d) having an Intelligence Quotient (IQ) of $\geq$80 according to the Test of Non-Verbal Intelligence (TONI-2; Brown, Sherbenou, & Johnsen, 2009), e) having a score of <19 on the Children's Depression Inventory (CDI, Kovacs, 1992), and f) being enrolled in a regular private or public school.

Participants were excluded according to the following criteria: a) having a history of a central nervous system disease that is associated with neuropsychological problems (e.g. epilepsy, brain injury, movement disorders, multiple sclerosis, brain tumor, stroke), b) having a history of alcohol abuse and/or consumption of psychotropic substances, c) having some type of active or uncontrolled systemic disease associated with cognitive impairment (e.g. diabetes mellitus, hypothyroidism, vitamin B12 deficiency), d) having a history of psychiatric illness (e.g. mayor depression, bipolar mood disorder, psychosis), e) having severe sensory deficits (e.g. loss of vision

Table 1
Sample distribution by center and country

	Center	Frequency	Percent
Chile	Chimbarongo/Rancagua/ San vicente	284	73.4%
	Chillan/Curico/Talca	103	26.6%
	Total Chile	387	100.0%
Cuba	Havana	381	100.0%
Ecuador	Quito	302	100.0%
Guatemala	Guatemala	203	100.0%
Honduras	Tegucigalpa	300	100.0%
Mexico	Mexico D.F.	239	25.6%
	Guadalajara	264	28.3%
	Mexicali	232	24.8%
	Monterrey	199	21.3%
	Total Mexico	934	100.0%
Paraguay	Asuncion	79	26.3%
	Central Department	221	73.7%
	Total Paraguay	300	100.0%
Peru	Arequipa	348	100.0%
Puerto Rico	Ponce	66	30.7%
	San German	123	57.2%
	San Juan	26	12.1%
	Total Puerto Rico	215	100.0%
Spain	Alicante	226	22.5%
	Almeria	188	18.7%
	Granada	211	21.0%
	Madrid	190	18.9%
	Seville	188	18.7%
	Total Spain	1003	100.0%

and/or hearing) that affect the administration of or performance on the tests, f) being on psychiatric or other medications that could alter cognitive performance, g) having intellectual or learning disability or other neurodevelopmental disorders, h) having a history of pre- peri-, and post-natal problems (e.g. hypoxia, jaundice, seizures, hydrocephalus, spine bifida, neuromuscular disorders), i) having a score of >5 on the Alcohol Use Disorders Identification Test (AUDIT-C) for participants 12 years of age and older, and j) using psychoactive substances such as heroin, barbiturates, amphetamines, methamphetamines, or cocaine in the last 6 months for participants 12 years of age and older.

2.2. Instruments

To determine if participants met the inclusion and exclusion criteria, the parents (or guardians) of each potential child or adolescent participants answered a sociodemographic and inclusion/exclusion questionnaire. All children and adolescents also completed the TONI-2 to determine their IQ and the CDI to assess depressive symptomatology. Those 12 years of age and older completed the AUDIT-C to assess alcohol consumption and answered questions regarding recent psychoactive substance use.

2.3. Clinical and demographic interview for participants

A questionnaire was created to collect information about the child or adolescent related to the health status and clinical history. It was completed by the parent or guardian. With this information, it was possible to identify participants who met the exclusion criteria proposed for the present study. In the interview, the following information was obtained: demographic data, pregnancy, childbirth, and possible complications; motor, language, visual, and auditory problems; assistance received by different professionals (e.g. neurologist, psychiatrist, medical rehabilitation professional, occupational therapist, speech therapist, psychologist), the existence of psychological disorders, and pharmacological treatment. The parents/guardians also responded to a demographic questionnaire about themselves that included

Table 2
Sample distribution by country, age, sex, type of school, and MLEP

	Sample		Age	Sex		Type of school		MLPE
	n Total	Maximum error (Accuracy level)	Mean (SD)	Girls *n* (%)	Boys *n* (%)	Public *n* (%)	Private *n* (%)	Mean (SD)
Chile	387	0.050 (95.0%)	11.5 (3.5)	194 (50.1%)	193 (49.9%)	195 (50.4%)	192 (49.6%)	12.3 (3.0)
Cuba	381	0.050 (95.0%)	11.5 (3.5)	190 (49.9%)	191 (50.1%)	381 (100%)	0 (0.0%)	16.2 (1.8)
Ecuador	302	0.056 (94.4%)	11.4 (3.5)	175 (57.9%)	127 (42.1%)	159 (52.6%)	143 (47.4%)	14.4 (3.6)
Guatemala	203	0.069 (93.1%)	10.7 (2.5)	94 (46.5%)	108 (53.5%)	112 (55.2%)	91 (44.8%)	10.5 (4.1)
Honduras	300	0.056 (94.4%)	11.2 (3.2)	161 (53.7%)	139 (46.3%)	155 (51.7%)	145 (48.3%)	12.8 (3.7)
Mexico	934	0.032 (96.8%)	11.4 (3.5)	481 (51.5%)	453 (48.5%)	574 (61.5%)	360 (38.5%)	13.1 (3.9)
Paraguay	300	0.056 (94.4%)	11.6 (3.5)	161 (53.7%)	139 (46.3%)	141 (47.0%)	159 (53.0%)	14.1 (2.9)
Peru	348	0.053 (94.7%)	12.0 (3.3)	171 (49.1%)	177 (50.9%)	187 (53.7%)	161 (46.3%)	12.5 (2.4)
Puerto Rico	215	0.067 (93.3%)	12.2 (3.6)	120 (55.8%)	95 (44.2%)	133 (61.9%)	82 (38.1%)	14.5 (2.6)
Spain	1003	0.031 (96.9%)	11.3 (3.4)	518 (51.6%)	485 (48.4%)	546 (54.4%)	457 (45.6%)*	14.0 (4.0)

Note: MLPE: Mean Level Parental Education; *Private/Concerted (private school partially publicly funded).

information on their age, education, monthly income, and occupation.

2.4. Screening tests

2.4.1. Test of Non-Verbal Intelligence (TONI-2; Brown et al., 2009)

The TONI-2 is a test that evaluates intelligence in people between the ages of 5 and 85 years. The duration of the test is between 15 and 20 minutes. The test attempts to eliminate any influence of language, motor ability, and culture by assessing the ability to solve abstract tests. The TONI-2 positively correlates with the Performance IQ of the Wechsler Intelligence Scale for Children - 3rd edition (WISC-III), which provides support for adequate concurrent validity. The TONI-2 also has demonstrated support for good construct validity established by correlations between the TONI-2 and five of six WISC-III subtests, as well as the predictive value of the spelling, mathematical problem-solving, and reasoning tasks of the Stanford Achievement Test (SAT) (Bostantjopoulou, Kiosseoglou, Katsarou, & Alevriadou, 2001; Mackinson, Leigh, Blennerhassett, & Anthony, 1997).

2.4.2. Children's Depression Inventory (CDI; Kovacs, 1992)

The CDI is a self-report test composed of 27 items that measure depressive symptoms in children and adolescents. This measure is composed of five subscales: negative humor, ineffectiveness, low self-esteem, social withdrawal, and pessimism. Each item has three possible responses (0, 1, or 2) depending on the degree of depression, with 0 representing an absence of symptomatology and 2 representing severe depressive symptomatology. The suggested cut-off is 19 points, and the maximum score is 54 points.

2.4.3. The AUDIT Alcohol Consumption Questions (AUDIT-C; Bush, Kivlahan, McDonell, Fihn, & Bradley, 1998)

The AUDIT-C is the modified version of the AUDIT, which consists of three items from the traditional 10-item version. Each item is scored on a scale of 0 to 4 points. The AUDIT-C has demonstrated support for validity as a primary care screening test for heavy drinking and/or active alcohol abuse or dependence. In this study, only those participants 12 years of age and older completed this measure and those with a score >5 were excluded from participation.

2.4.4. Checklist for use of psychoactive substances

This measure is a list containing the most commonly used psychoactive substances. Participants were excluded from the study if they reported having consumed any type of substance in the last six months, including heroin, barbiturates, amphetamines, methamphetamines, and cocaine. These questions were administered only to participants 12 years of age and older.

2.5. Neuropsychological tests

Participants who met the inclusion criteria were administered the following neuropsychological tests:

1. Rey Osterrieth Complex Figure Test (ROCF; Rey, 2009)
2. Stroop Color-Word Interference Test (Golden, 2010)
3. Modified Wisconsin Card Sorting Test (M-WCST; Schretlen, 2010)
4. Trail Making Test A-B (TMT A-B; Reitan & Wolfson, 1985)
5. Symbol Digit Modalities Test (SDMT; Smith, 2002)
6. Shortened version of Token Test (De Renzi & Faglioni, 1978)
7. Concentration Endurance Test (d2) (Brickenkamp, 2009)
8. Phonological and Semantic Verbal Fluency Tests (Benton & Hamsher, 1989)
9. Peabody Picture Vocabulary Test - PPVT-III (Dunn, Dunn, & Arribas, 2010)
10. Learning and Verbal Memory Test (TAMV-I; Rivera, Olabarrieta-Landa, & Arango-Lasprilla, 2017)

2.6. Procedure

The present study began by recruiting the local researchers who would carry out the study in their individual countries. Universities, institutions, and research centers from Latin America (Chile, Cuba, Ecuador, Guatemala, Honduras, Mexico, Paraguay, Peru, and Puerto Rico) and Spain were contacted and invited to participate in the multi-center study. Next, the centers that agreed to participate in the study requested approval from their institution/center's ethics committee. After approval, any copyrighted test materials (manuals, booklets, and stimulus cards)

were purchased. In each center, a neuropsychology researcher was appointed to coordinate data collection for the study. Randomized lists were used to determine the order of test administration for each participant, aiming to avoid order bias and cognitive conditioning. For the creation of the list, the function $f_x = \text{RANDOM}()$ in Microsoft Excel© was used, and this configuration considered the interaction of verbal fluency tests with the TAMV-I verbal memory test. To enter the data, a template was designed in Microsoft Excel©, which was designed using the following configuration options to reduce data entry bias: *Data validation = custom* (numeric variables), *drop-down lists* (categorical variables), and *configuration formats*. All the neuropsychological tests were administered according to the specific manual guidelines of each test.

A pilot test was performed with 20 participants to ensure proper functioning and comprehension of Spanish-language test instruction. Data from the pilot test were eliminated from data analysis. Data collection began in January 2016 and finished in May 2017. The neuropsychological battery was administered individually in a single day to children and adolescents at schools and/or universities. Administration lasted approximately 120 minutes per participant. Prior to initiating the screening tests and battery of neuropsychological tests, parent questionnaires were received and reviewed, all parents/guardians and children 12 years of age and older signed the informed consent, and children under 12 years of age signed the assent. Informed consent included information related to the aim of the study, rights of participants, duration/place of the assessment, contact information for the local researcher responsible for the study, and the possibility to be re-evaluated in a set period of time (60–120 days).

In each country, a randomly selected group of children and adolescents with a size of 5–10% of the sample collected were selected for re-evaluation 60 to 120 days after the initial evaluation. Randomization was done using *random sample cases* in SPSS. The same procedures, including random test order generation, were used for the re-evaluation session. Unfortunately, 5 of the countries (Guatemala, Honduras, Paraguay, Peru, and Puerto Rico) began data collection late and were unable to re-evaluate any of the participants within the original study time frame. Thus, test-retest data for this sample was available from a sub-sample of 233 participants from Chile, Cuba, Ecuador, Mexico, and Spain.

2.7. Statistical analysis

2.7.1. Accuracy of the final sample

The accuracy of the total sample size by country was established using classical estimation assuming infinite (very large) population sizes (Arrufat, Guàrdia-Olmos, & Blanxart, 1999), where the case of maximum uncertainty was assumed ($\pi = 1 - \pi = 0.5$) and a confidence interval of 95%.

2.7.2. Reliability

The reliability of each neuropsychological test's scores for the entire sub-sample was calculated through the test-retest method, which evaluates the temporal stability of the test scores. This type of reliability assumes that the scores of examinees will not have unexpected or fluctuating changes over time (Abad, Díez, Gil, & García, 2011). The reliability was calculated with intraclass correlation coefficients ($ICC_{xtest-retest}$). The intraclass correlation coefficient is frequently used to report reliability, especially with variables that share metric and variance (McGraw & Wong, 1996; Weir, 2005).

2.7.3. Normative data

In order to determine if there were significant differences between countries on the performance of each of the 10 neuropsychological tests, a multivariate analysis of variance (MANOVA) was performed to examine the effect of *country* on the test scores. Country was dummy coded and used as a fixed factor, and each of test scores served as dependent variables. Bonferroni adjustment alpha level of 0.005 to 10 pairwise comparisons was use (0.05/10).

Linear regression models and standard deviations of the residual values of the models (Van Breukelen & Vlaeyen, 2005) were used to generate normative data for each country. A multivariate regression model was fitted to the data. The multivariate regression model assumes that $Y_i = X_i\beta + \varepsilon_i$, where Y_i is the vector of the measurements for children, X_i the design matrix for the fixed effects, $\boldsymbol{\beta}$ the vector of the regression coefficients (fixed effects), and $\boldsymbol{\varepsilon_i}$ the vector of the residual components. The main model included age, age^2, sex of the child, and the mean level parental education (MLPE) as predictor variables. Age was centered (= Age – average age of the sample in each country), and then age^2 was calculated from the centered age to avoid multicollinearity (Aiken & West, 1991). Sex was coded as boys = 1

and girls = 0. The MLPE variable was coded as 1 if the participant's parent(s) had >12 mean years of education or 0 if participant's parent(s) had ≤12 mean years of education. The cut-off of 12 years of education was selected because it provided a relatively standard reference point and was defined as a crucial cut-off point for higher education in other studies with Spanish-Speaking countries in Latin America (Guàrdia-Olmos, Peró-Cebollero, Rivera, & Arango-Lasprilla, 2015) and Spain (Peña-Casanova et al., 2009).

A final regression model was conducted $\hat{y}_i = B_0 + B_1 \cdot (Age - \bar{X}_{Age\ by\ country})_i + B_2 \cdot (Age - \bar{X}_{Age\ by\ country})_i^2 + B_3 \cdot Sex_i + B_4 \cdot MLPE_i$. If predicted variables were not statistically significant in the multivariate model with an alpha of 0.05, the non-significant variables were removed, and the model was run again. The established regression model and the standard deviation of the residual values (SD_e) provided by the regression model were subsequently used to norm the score (Van Der Elst, Van Boxtel, Van Breukelen, & Jolles, 2006a; 2006b; Van der Elst, Hurks, Wassenberg, Meijs, & Jolles, 2011; Van der Elst, Dekker, Hurks, & Jolles, 2012). Standard deviation of the residual (SD_e) value is calculated as follows: $\left(SD_e = \sqrt{MSE}\right)$, where Mean Squared Error (MSE) is $\sum (y_i - \hat{y}_i)^2 / n - k$. Using each country's dataset, these models were applied to each neuropsychological test's scores separately.

For all multiple linear regression models, the following assumptions were evaluated: a) collinearity by a Variance Inflation Factor (VIF) not greater than 10 and a collinearity tolerance values not greater than 1 (Kutner, Nachtsheim, Neter, & Li, 2005), b) normality using Q-Q plots and histograms of residual values, and c) the existence of influential values by calculating the Cook's distance. The maximum Cook's distance value was related to a $F(p, n - p)$ distribution, where p is the number of regression parameter, including constant, and n is the sample size. Influential values are considered when percentile value is equal or higher than 50 (Cook, 1977; Kutner et al., 2005). The scores should be transform to other metrics (e.g. root square) if severe violation of normality assumptions happens. In case of multicollinearity, the main model should be change, and in case of influential values, outliers' cases should be excluded for the analysis. All analyzes were performed using SPSS version 23 (IBM Corp., Armonk, NY).

3. Results

3.1. Accuracy of the final sample

The maximum error (e) of sample sizes range from 0.069 (Guatemala) to 0.031 (Spain). All country sample size maximum error (e) and accuracy levels can be found in Table 2. Regarding the sample, non-proportionate quota sampling was used, looking for a symmetrical distribution for strata of age, sex, and type of school for each country, except for Cuba, where the education is entirely public (Gasperini, 2000).

3.2. Reliability

The subsample used to calculate the test-retest reliability consisted of 232 participants from Chile ($n = 29$; 12.4%), Cuba ($n = 38$; 16.3%), Ecuador ($n = 16$; 6.9%), Mexico ($n = 74$; 31.8%), and Spain ($n = 76$, 32.6%). Sixty percent were boys and the average age was 11.5 (SD = 3.6) years. The average time elapsed between the two tests administrations was 78.2 (SD = 11.7; range = 60–115) days. In Table 3 shows the interclass coefficients for each neuropsychological test score for the entire sub-sample.

The scores demonstrated temporal stability, with interclass coefficients greater than 0.54 ($r = 0.54$ to $r = 0.95$) that were statistically significant (p's<0.001). In general, good and high magnitudes of temporal stability were observed for the scores of all tests examined, indicating a consistency of the measures across time periods, which fulfilled the assumption that participant scores would not demonstrate unexpected or fluctuating changes (Abad et al., 2011; Cicchetti, 1994; Strauss et al., 2006).

3.3. Normative data

MANOVAs using country as a fixed factor and all neuropsychological test scores as dependent variables showed a significant difference across the ten countries (Wilks' Lambda = 0.349, $F_{(252,31359)} = 17.210$, $p < 0.001$, partial $\eta^2 = 0.116$). Subsequent analyses of the group differences within each of these dependent variables were examined in tests of Between-Subjects Effects (see Table 4).

3.3.1. Verification of the assumptions in the regression models for the scores

The summary of verification of assumptions for the final models is shown in Table 5. The visual

Table 3
Intraclass Correlation coefficients to test – retest

Test scores	$ICC_{Xtest\ Xre-test}$		
	Coefficient	95% CI	
FOCR Copy	0.89**	0.86	0.92
FOCR Immediate Recall	0.88**	0.85	0.91
Stroop Word	0.90**	0.87	0.93
Stroop Color	0.91**	0.88	0.93
Stroop Word-Color	0.87**	0.83	0.90
M-WCST Correct categories	0.80**	0.74	0.85
M- WCST Perseveration errors	0.54**	0.41	0.65
M-WCST Total errors	0.79**	0.73	0.84
TMT – A	0.68**	0.58	0.75
TMT – B	0.76**	0.69	0.82
SDMT	0.88**	0.84	0.90
Token Test	0.75**	0.68	0.81
Letter F	0.87**	0.82	0.90
Letter A	0.80**	0.74	0.84
Letter S	0.79**	0.73	0.84
Category Animals	0.78**	0.72	0.83
Category Fruits	0.81**	0.75	0.85
Peabody Test	0.95**	0.93	0.96
TAMVI Total Recall	0.70**	0.61	0.77
TAMVI Delayed Recall	0.68**	0.59	0.76
TAMVI Recognition	0.70**	0.61	0.77
d2 – TN	0.77**	0.70	0.82
d2 – CR	0.89**	0.85	0.91
d2 – TP	0.79**	0.72	0.84
d2 – CP	0.81**	0.75	0.85

Note: ** $p < 0.001$; ICC: Intraclass correlation coefficient; TN: Total number of items processed, CR: Total number of correct responses, TP: Total performance, CP: Concentration performance.

analysis of the Q-Q normality diagrams and histograms of the residual values for the final models fulfilled the assumption of normality and indicated a non-problematic distribution in each of the models studied. No evidence of multicollinearity was found since all VIF values were lower than 1.311 and tolerance values did not exceed 1. No influential cases were present since the Cook distance values were less than 0.251. The maximum Cook's distance value was 0.251 in a $F_{(3,212)}$ distribution which correspond to percentile 14. Because the percentile value did not exceed 50 the influence of outliers was not deemed sufficient to require remedial measures.

The methodology to generate normative data based on linear regression provides normative data calculations using a four-step procedure described by Van Breukelen and Vlaeyen (2005) and Van der Elst et al. (2006a; 2006b; 2011; 2012):

1. The expected test score of tested children (i) is computed based on the fixed effect parameter estimates of the established regression model: $\hat{Y}_{\mathrm{i}} = B_0 + B_1 X_1 + B_2 X_2 \ldots B_K X_K$.
2. To obtain the residual value ($\hat{e}_i$), a subtraction between the raw score of the neuropsychological test (Y_{i}) and the predicted value previously calculated was performed $\left(\hat{Y}_{\mathrm{i}}\right)$, as shown in the following formula: $\hat{e}_i = Y_{\mathrm{i}} - \hat{Y}_i$.
3. Using the residual standard deviation (SD_e) value provided by the regression model, residuals were standardized: $\hat{z}_{\mathrm{i}} = \hat{e}_i / SD_e$.
4. Finally, the exact percentile corresponding to the z-score previously calculated using the cumulative probability function of the standard normal distribution must be found. This can be done using conversion tables (e.g. Strauss et al., 2006) or using online calculators. In the online calculator, the z-score is entered, and the percentile of the unilateral test is then chosen. For example, the corresponding percentile of $\hat{z}_{\mathrm{i}} = -0.296$ is 38.

3.4. User-friendly normative data

The four-step normative procedure explained above offers clinicians the ability to determine the exact percentile of a child's test score. However, this method can be prone to human error due to the number of hand computations required. To enhance user-friendliness, the authors have computed these steps for a range of raw test scores based on age, sex, and MLPE as necessary, and created tables for clinicians to more easily obtain a percentile range/estimate associated with a given raw score on a specific test. The percentile obtained using these user-friendly tables could be slightly different from the hand-calculation (which although prone to human error, is a more precise method) because the user-friendly table is based on a limited number of percentile values.

4. Discussion

The purpose of this paper was to describe the methodology and procedures involved in generating normative data for 10 neuropsychological tests for children and adolescents from nine countries in Latin America (Chile, Cuba, Ecuador, Guatemala, Honduras, Mexico, Paraguay, Peru, and Puerto Rico) and Spain.

Test-retest reliability was also examined. The interclass coefficients obtained in this study have good (0.60–0.74) or excellent (0.75–1.00) magnitudes of reliability according to Cicchetti (1994), except

Table 4
Differences between countries by raw scores

Dependent Variable	*F*	Sig.	η^2 (90% CI)
FOCR Copy	47.896	<0.001	0.108 (0.09; 0.12)
FOCR Immediate Recall	28.595	<0.001	0.067 (0.05; 0.08)
Stroop Word	10.239	<0.001	0.025 (0.01; 0.03)
Stroop Color	7.160	<0.001	0.018 (0.01; 0.02)
Stroop Word-Color	10.001	<0.001	0.025 (0.01; 0.03)
M-WCST Correct categories	32.255	<0.001	0.075 (0.06; 0.09)
M-WCST Perseveration errors	13.063	<0.001	0.032 (0.02; 0.04)
M-WCST Total errors	42.470	<0.001	0.097 (0.08; 0.11)
TMT – A	16.386	<0.001	0.040 (0.03; 0.05)
TMT – B	15.622	<0.001	0.038 (0.03; 0.05)
SDMT	9.672	<0.001	0.024 (0.01; 0.03)
Token	37.775	<0.001	0.087 (0.07; 0.10)
Letter F	9.973	<0.001	0.024 (0.01; 0.03)
Letter A	2.464	0.012	0.006 (0.00; 0.01)
Letter S	4.058	<0.001	0.010 (0.00; 0.01)
Category Animals	19.984	<0.001	0.048 (0.03; 0.06)
Category Fruits	10.752	<0.001	0.026 (0.02; 0.03)
Peabody	26.638	<0.001	0.063 (0.05; 0.07)
TAMVI Total Recall	25.869	<0.001	0.061 (0.05; 0.07)
TAMVI Delayed Recall	14.073	<0.001	0.034 (0.02; 0.04)
TAMVI Recognition	15.024	<0.001	0.036 (0.02; 0.05)
d2 – TN	11.238	<0.001	0.027 (0.02; 0.04)
d2 – CR	6.443	<0.001	0.016 (0.01; 0.02)
d2 – TP	7.272	<0.001	0.018 (0.01; 0.02)
d2 – CP	7.425	<0.001	0.018 (0.01; 0.02)

Note: TN: Total number of items processed, CR: Total number of correct responses, TP: Total performance, CP: Concentration performance.

M-WCST perseverative errors, whose coefficient magnitude is fair ($r = 0.54$; $p < 0.001$). Thus, almost all of the neuropsychological test scores showed temporal test stability. In addition, the stability coefficients obtained in this study can provide valuable information about the replicability of the test results (Strauss et al., 2006). Similarly, other studies looking at the same neuropsychological tests have shown similar test-retest reliability (e.g. Hurks, 2012; Reese & Read, 2000; Tombaugh, Kozak, & Rees, 1999).

Regarding norms, researchers have pointed out the need of normative data adjusted to the demographic characteristics of a specific population (Strauss et al., 2006; Van der Elst et al., 2013) because the raw scores on neuropsychological tests can be impacted by these types of variables (e.g. age, education, sex, race, ethnicity). Despite the wide use of neuropsychological tests for the evaluation and diagnosis of cognitive problems in Latin America and Spain, to date there are few studies with normative data for neuropsychological tests for Spanish-Speaking children and adolescents (Arango-Lasprilla et al., 2016; Olabarrieta-Landa et al., 2016).

Existing normative data studies have several limitations, such as 1) the use of mean and standard deviation within each subgroup (e.g. Malloy-Diniz et al., 2007; Oliveira, Mograbi, Gabrig, & Charchat-Fichman, 2016), 2) conversion of raw scores to metrics such as Z or T values (e.g. Golden, 2010; Rey, 2009; Schretlen, 2010), and 3) having no representative samples in their studies (e.g. Malloy-Diniz et al., 2007; Matute, Rosselli, Ardila, & Morales, 2004). Each of these limitations presents significant problems, which will be explained in detail in the next paragraphs.

The use of mean and standard deviations within subgroups can have two major drawbacks (Van Breukelen & Vlaeyen, 2005). First, normative data tables are generated, assuming that predictive variables of the test scores are known, as not all variables assumed are relevant. For example, normative data generated by sex may not be important for a test if this variable does not influence performance. Second, taking into account traditional demographic variables (e.g. sex and age) and dividing the sample into subgroups implies a considerable loss of information. For example, when dividing the sample by sex (males vs. females), the sample size is reduced by approximately 50%. In addition, if the sample is divided into five age groups, it reduces the size of the sample per subgroup to 10% of the total sample. As a result, the distribution of test scores, including its mean and standard

Table 5
Summary of Collinearity Statistics and Cook's Distance values for end models in all scores of different countries values

Test	Maximum VIF	Maximum Cook's Distance
FOCR Copy	1.097	0.131
FOCR Immediate Recall	1.060	0.080
Stroop Word	1.116	0.151
Stroop Color	1.068	0.123
Stroop Word-Color	1.072	0.154
M-WCST Correct categories	1.015	0.094
M-WCST Perseveration errors	1.051	0.158
M-WCST Total errors	1.048	0.155
TMT – A	1.008	0.242
TMT – B	1.012	0.115
SDMT	1.054	0.222
Token	1.311	0.216
Letter F	1.025	0.082
Letter A	1.011	0.117
Letter S	1.018	0.091
Category Animals	1.033	0.087
Category Fruits	1.050	0.130
Peabody	1.131	0.184
TAMVI Total Recall	1.063	0.238
TAMVI Delayed Recall	1.066	0.190
TAMVI Recognition	1.066	0.251
d2 – TN	1.199	0.159
d2 – CR	1.010	0.081
d2 – TP	1.201	0.094
d2 – CP	1.022	0.079

Note: TN: Total number of items processed, CR: Total number of correct responses, TP: Total performance, CP: Concentration performance.

deviation, are not reliable since it can lead to random trends in norm tables, where the average can vary and make large jumps across the age groups. Conversion of raw scores to other metrics (e.g. standard or z-scores) are common practice; but, this simple conversion of raw scores has no effect on the shape of the distribution. If raw scores are normally distributed, the resulting z-scores will be as well. However, if the raw scores have an asymmetric distribution, then the z-scores will be similarly skewed (Crawford, 2004).

Finally, non-representative sample sizes can lead to problems when generating normative data to be generalized to the general population. The sample should reflect its composition—that is, be representative—to ensure the external validity of the research and thus the replicability of the results.

In order to address these limitations, the present study used a method based on multiple regression models and standard deviation of residual values (Van Breukelen & Vlaeyen, 2005). This method provides information on which variables predict test scores and which variables are relevant for the development of valid normative data. In this study, age, age^2, sex, and MLPE were used as predictors as part of a main model for each of the test scores. The final models were constructed from the hierarchical elimination of non-significant predictor variables (Van der Elst, Molenberghs, van Tetering, & Jolles, 2017), or variables with a *p value* greater than 0.05 ($p > 0.05$). However, most predictors in the final models had a p value less than 0.001 ($p < 0.001$), controlling for possible type I errors (Van der Elst et al., 2006a).

Another advantage for the regression approach to developing normative data is that norms are continuous and more reliable than those obtained by tabulating the mean and standard deviation of the score scales for different age groups, level of education, and sex. In addition, normative data based on regression allows for exploration of more than one type of function (linear vs. quadratic) for age. For example, a curvilinear effect of age on cognitive processes has been shown in past research. This is of special relevance in studies with children, given that a single year can have a tremendous difference in terms of rapid cognitive development compared to adults. In addition, cognitive development is not always present in linear functions depending on the cognitive process under study (Grady, 2012). Another important variable that should be considered in developmental studies is parental education. Parental education has been found to predict cognitive development (Meador et al., 2011; Schady, 2011), educational level (Dubow, Boxer, & Huesmann, 2009; Ermisch, & Pronzato, 2010) and occupation (Dubow et al., 2009) of children in the future. That is why, unlike other normative data studies, parents' educational level was introduced in the models as a predictive variable to generate normative data. In a study by Van der Elst et al. (2011), parental level of education was used as a predictive variable in their normative study, and results indicated that parents' education influences verbal fluency performance among children.

Another point in favor of this method is the identification and control of collinearity in the predictive variables through VIF. In the present study, VIF values were less than 1.311 and tolerance values did not exceed 1, well below threshold values. Finally, this method allows for the standardization of raw scores. Transforming raw scores into percentiles allows z-scores to be standardized by using a table of the areas under the normal curve (Crawford, 2004). This method has been used recently in different studies to generate normative data for several neuropsychological tests (Guàrdia-Olmos et al., 2015; Van der Elst et al., 2011, 2012).

Finally, the present study used a total sample of 4,373 participants from nine Latin American countries and Spain, making this the first and largest normative data study done in pediatric and adolescent populations in the world. Moreover, the subsamples of each country have precision levels that oscillate between 93.1% (Guatemala) and 96.9% (Spain), which approaches the best possible representation for each population.

4.1. Limitations

This study has several limitations. First, although the method used in this study allows knowing which variables predict test performance in children (even after having included MLPE as a predictor), there may be other variables that affect test performance that were not presently examined. Therefore, future studies should take into account other potentially important variables (e.g. bilingualism etc.).

Second, to avoid possible human error when calculating test percentiles, this study provides tables with approximate percentiles. It is hoped that such tables will facilitate use of appropriate norms by professionals and allow more accurate interpretation of results. However, these tables are based on a limited number of percentile values and for that reason may be slightly different from hand-calculations.

Although having an overall large number of participants, there were countries with a relatively small sample (e.g. Guatemala and Puerto Rico). This, however, did not prevent the generation of valid normative data for these countries. Future studies should increase the sample size in these countries in order to reduce potential bias due to sampling error. Moreover, excluding samples from Chile, Mexico, Paraguay, Puerto Rico, and Spain, in the remaining countries the samples were collected from only one geographical area, potentially affecting generalization of norms to the entire country. Finally, all data was collected from urban areas. Future studies should expand data collection to other geographical areas of these countries and assess children from rural schools to improve representativeness and generalizability.

5. Conclusions

Despite these limitations, this is the largest normative data study in the world with a total sample of 4,373 Spanish-speaking children and adolescents in 9 Latin American countries and Spain. The selection of neuropsychological tests in this study was according to the frequency of use by neuropsychologists in Latin America and Spain (Arango-Lasprilla et al., 2016; Olabarrieta-Landa et al., 2016). For the generation of normative data, the method based on linear regression models and the standard deviation of residual values was used. This method allows determination of the variables that predict test scores, helps identify and control for collinearity of predictive variables, and generates continuous and more reliable norms than those of traditional methods (e.g. obtaining just the mean and standard deviations). In addition, to generate normative data, the MLPE was included as a predictive variable, which is especially relevant in child development studies. Finally, this study describes the methodology used to generate normative data for 10 neuropsychological tests for children and adolescents in nine countries from Latin America and Spain. Futures studies can utilize the presented procedures to create normative data for these tests in other Spanish-speaking countries or to develop new norms for other neuropsychological tests with the goal of continuing the development and improvement of clinical practice in these countries.

Conflict of interest

None to report.

References

Abad, F. J., Díez, J. O., Gil, V. P., & García, C. G. (2011). *Medición en ciencias sociales y de la salud*. Madrid: Editorial Síntesis.

Aiken, L. S., & West, S. G. (1991). *Multiple regression: Testing and interpreting interactions*. Sage Publications.

American Psychiatric Association. (2014). *Guía de consulta de los criterios diagnósticos del DSM-5®: Spanish Edition of the Desk Reference to the Diagnostic Criteria From DSM-5®*. American Psychiatric Pub.

APA Division 40 (2001). An educational pamphlet brought to you by the Public Interest Advisory Committee, Division 40 (Clinical Neuropsychology) American Psychological Association. Recuperated from www.div40.org/pdf/pedneuropscyhBroch3.pdf

Arango-Lasprilla, J. C., Rivera, D., & Olabarrieta-Landa, L. (2017). *Neuropsicología infantil*. Bogotá, Colombia: Editorial Manual Moderno.

Arango-Lasprilla, J. C., Stevens, L., Morlett Paredes, A., Ardila, A., & Rivera, D. (2016). Profession of neuropsychology in Latin America. *Applied Neuropsychology: Adult*, *24*(4), 318-330.

Arrufat, A. S., Guàdia-Olmos, J. G., & Blanxart, M. F. (1999). *Introducción a la estadística en Psicología* (Vol. 27). Edicions Universitat Barcelona.

Baron, I. S. (2004). *Neuropsychological evaluation of the child*. New York, New York, Oxford University Press.

Benton, A. L., & Hamsher, K. D. (1989). *Multilingual Aphasia Examination*. Iowa City, IA: AJA Associates.

Bostantjopoulou, S., Kiosseoglou, G., Katsarou, Z., & Alevriadou, A. (2001). Concurrent validity of the Test of Nonverbal Intelligence in Parkinson's disease patients. *The Journal of Psychology*, *135*(2), 205-212.

Brickenkamp, R. (2009). Manual del test de atención d2. Madrid: Tea Ediciones.

Brown, L., Sherbenou, R. J., & Johnsen, S. K. (2009). *Test of Nonverbal Itelligence*. Austin, TX: Pro-Ed.

Bush, K., Kivlahan, D. R., McDonell, M. B., Fihn, S. D., & Bradley, K. A. (1998). The AUDIT alcohol consumption questions (AUDIT-C): An effective brief screening test for problem drinking. *Archives of Internal Medicine*, *158*(16), 1789-1795.

Cianchetti, C., Corona, S., Foscoliano, M., Conty, D. & Sannio-Fancello, G. (2007). Modified Wisconsin Card Sorting Test (MCST, MWCST): Normative Data in Children 4-13 years old, according to Classical and New Types of Scoring. *The Clinical Psychologist*, *21*, 456-478.

Cicchetti, D. V. (1994). Guidelines, criteria, and rules of thumb for evaluating normed and standardized assessment instruments in psychology. *Psychological Assessment*, *6*(4), 284.

Cook, R. D. (1977). Detection of influential observation in linear regression. *Technometrics*, *19*(1), 15-18. doi: 10.2307/1268249

Crawford, J. R. (2004). Psychometric foundations of neuropsychological assessment. In L. H. Goldstein & J. McNeil (Eds.), *Clinical Neuropsychology: A Practical Guide to Assessment and Management for Clinicians*. Chichester: Wiley

Davies, S., Field, A., Andersen, T., & Pestell, C. (2011). The Ecological validity of The Rey-Osterrieth Complex Figure: Predicting everyday problems in children with neuropsychological disorder. *Journal of Clinical and Experimental Neuropsychology*, *33*(7), 820-831.

De Renzi, E., & Faglioni, P. (1978). Development of a shortened version of the Token Test. *Cortex*, *14*, 41-49.

Dubow, E. F., Boxer, P., & Huesmann, L. R. (2009). Long-term effects of parents' education on children's educational and occupational success: Mediation by family interactions, child aggression, and teenage aspirations. *Merrill-Palmer quarterly (Wayne State University. Press)*, *55*(3), 224-249.

Dunn, L. M., Dunn, L. M., & Arribas, D. (2010). *PPVT-III Peabody. Test de vocabulario en imágenes* (2da ed.). Madrid: TEA.

Ermisch, J., & Pronzato, C. (2010). *Causal effects of parents' education on children's education* (No. 2010-16). ISER Working Paper Series.

Golden, C. J. (2010). *Manual de test de colores y palabras. Publicaciones de psicología aplicada*. Madrid: TEA Ediciones.

Goodman, A., Delis, D., & Mattson, S. (2010). Normative Data for 4 year old children on the California Verbal Learning Test-Children's Version. *The Clinical Psychologist*, *13*(3), 274-282.

Grady, C. (2012). The cognitive neuroscience of ageing. *Nature Reviews Neuroscience*, *13*(7), 491-505.

Guàrdia-Olmos, G., Pero-Cebollero, M., Rivera, D., & Arango-Lasprilla, J. C. (2015). Methodology for the development of normative data for ten Spanish-language neuropsychological tests in eleven Latin American countries. *NeuroRehabilitation*, *37*(4), 493-499.

Harrison, J. E., Buxton, P., Husain, M., & Wise, R. (2000). Short test of semantic and phonological fluency: Normal performance, validity and test-retest reliability. *British Journal of Clinical Psychology*, *39*(2), 181-191.

Hiuzinga, M., & Smidts, D. (2010) Age-related changes in executive Function: A normative study with the Dutch Version of the Behavior Rating Inventory of Executive Function (BRIEF). *Child Neuropsychology: A journal on Normal and Abnormal Development in Childhood and Adolescence*, *17*(1), 51-66.

Hurks, P. P. (2012). Does instruction in semantic clustering and switching enhance verbal fluency in children? *The Clinical Neuropsychologist*, *26*(6), 1019-1037.

Kim, H., & Na, L. D. (2008). A normative study of the Boston Naming Test in 3-to-14 year-old Korean Children. *Psychology Press*, *22*, 84-97.

Kovacs, M. (1992). *The Children's Depression Inventory manual.* Toronto, Ontario, Canada: Multi-Health Systems.

Kutner, M. H., Nachtsheim, C. J., Neter, J., & Li, W. (2005). *Applied linear statistical models (5th ed.)*. New York: McGraw Hill.

Lezak, M. D, Howieson, D. B, & Loring, D. W, (2004). *Neuropsychological assessment* (4th ed.). New York: Oxford University Press New

Mackinson, J. A., Leigh, I. W., Blennerhassett, L., & Anthony, S. (1997). Validity of the TONI-2 with deaf and hard of hearing children. *American Annals of the Deaf*, *142*(4), 294-299.

Malloy-Diniz, L. F., Bentes, R. C., Figueiredo, P. M., Brandão-Bretas, D., da Costa-Abrantes, S., Parizzi, A.M., . . . & Salgado, J. V. (2007). Normalización de una batería de tests para evaluar las habilidades de comprensión del lenguaje, fluidez verbal y denominación en niños brasileños de 7 a 10 años: Resultados preliminares. *Revista de Neurología, 44*(5), 275-280.

Matute, E., Rosselli, M., Ardila, A., & Morales, G. (2004). Verbal and nonverbal fluency in Spanish-speaking children. *Developmental Neuropsychology*, *26*(2), 647-660.

McGraw, K. O., & Wong, S. P. (1996). Forming inferences about some intraclass correlation coefficients. *Psychological Methods*, *1*(1), 30.

Meador, K. J., Baker, G. A., Browning, N., Clayton-Smith, J., Cohen, M. J., Kalayjian, L. A.,... & Loring, D. W. (2011). Relationship of child IQ to parental IQ and education in children with fetal antiepileptic drug exposure. *Epilepsy & Behavior*, *21*(2), 147-152.

Olabarrieta-Landa, L., Caracuel, A., Pérez-García, M., Panyavin, I., Morlett-Paredes, A., & Arango-Lasprilla, J. C. (2016). The profession of neuropsychology in Spain: Results of a national survey. *The Clinical Neuropsychologist*, *30*(8), 1335-1355.

Oliveira, R. M., Mograbi, D. C., Gabrig, I. A., & Charchat-Fichman, H. (2016). Normative data and evidence of validity for the Rey Auditory Verbal Learning Test, Verbal Fluency Test, and Stroop Test with Brazilian children. *Psychology & Neuroscience*, *9*(1), 54.

Peña-Casanova, J., Blesa, R., Aguilar, M., Gramunt-Fombuena, N., Gómez-Ansón, B., Oliva, R.,... & Martínez-Parra, C. (2009). Spanish multicenter normative studies (NEURONORMA project): Methods and sample characteristics. *Archives of Clinical Neuropsychology*, *24*(4), 307-319.

Reese, E., & Read, S. (2000). Predictive validity of the New Zealand MacArthur communicative development inventory: Words and sentences. *Journal of Child Language*, *27*(2), 255-266.

Reitan, R. M., & Wolfson, D. (1985). *The Halstead-Reitan neuropsychological test battery: Theory and clinical interpretation*. Tucson, AZ: Neuropsychology Press.

Rey, A. (2009). *REY: Test de copia y de reproducción de memoria de figuras geométricas complejas*. Madrid: TEA ediciones.

Rivera, D., Olabarrieta-Landa, L., & Arango-Lasprilla, J. C. (2017). Diseño y creación del Test de Aprendizaje y Memoria Verbal Infantil (TAMV-I) en población hispano hablante de 6 a 17 años de edad. En J.C. Arango-Lasprilla, Rivera, D. & Olabarrieta-Landa, L. (Eds). Neuropsicología infantil (pp. 316-338). Manual Moderno: Bogotá.

Schady, N. (2011). Parents' education, mothers' vocabulary, and cognitive development in early childhood: Longitudinal evidence from Ecuador. *American Journal of Public Health*, *101*(12), 2299-2307.

Schretlen, D. J. (2010). *Modified Wisconsin Card Sorting Test: M-WCST; Professional Manual*. Odessa, FL: Psychological Assessment Resources.

Shu, B. C., Tien, A. Y., Lung, F. W., & Chang, Y. Y. (2000) Norms for the Wisconsin Card Sorting Test in 6-to-11-year-old Children in Taiwan. *The Clinical Neuropsychologist*, *14*(3), 275-286.

Smith, A. (2002). *Manual de test de símbolos y dígitos SDMT. Publicaciones de Psicología aplicada.* Madrid: TEA ediciones.

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

Tombaugh, T. H., Kozak, J. & Rees, L. (1999). Normative data stratified by age and education for two measures of verbal fluency: FAS and animal naming. *Archives of Clinical Neuropsychology*, *14*(2), 167-177.

Townes, B. D., Martins, I. P., Castro-Caldas, A., Rosenbaum, G., & Derouen, T. (2008). Repeat test scores on neurobehavioral measures over an eight-year period in a sample of Portuguese children. *International Journal of Neuroscience*, *118*(1), 79-93.

Vakil, E., Greenstein, Y., & Blachstein, H. (2010). Normative data for composite scores for children and adults derived from the Rey Auditory Verbal Learning Test. *The Clinical Neuropsychology*, *24*, 662-677.

Van Breukelen, G. J., & Vlaeyen, J. W. (2005). Norming clinical questionnaires with multiple regression: The Pain Cognition List. *Psychological Assessment*, *17*(3), 336.

Van der Elst, W., Hurks, P., Wassenberg, R., Meijs, C., & Jolles, J. (2011). Animal verbal fluency and design fluency in school-aged children: Effects of age, sex, and mean level of parental education, and regression-based normative data. *Journal of Clinical and Experimental Neuropsychology*, *33*(9), 1005-1015.

Van der Elst, W., Molenberghs, G., Van Boxtel, M. P., & Jolles, J. (2013). Establishing normative data for repeated cognitive assessment: A comparison of different statistical methods. *Behavior Research Methods*, *45*(4), 1073-1086.

Van der Elst, W., Molenberghs, G., Van Tetering, M., & Jolles, J. (2017). Establishing normative data for multi-trial memory tests: The multivariate regression-based approach. *The Clinical Neuropsychologist*, 1-15.

Van der Elst, W., Ouwehand, C., Van der Werf, G., Kuyper, H., Lee, N., & Jolles, J. (2012). The Amsterdam Executive Function Inventory (AEFI): Psychometric properties and demographically corrected normative data for adolescents aged between 15 and 18 years. *Journal of Clinical and Experimental Neuropsychology*, *34*(2), 160-171.

Van Der Elst, W., Van Boxtel, M. P., Van Breukelen, G. J., & Jolles, J. (2006a). Normative data for the Animal, Profession and Letter M Naming verbal fluency tests for Dutch speaking participants and the effects of age, education, and sex. *Journal of the International Neuropsychological Society*, *12*(01), 80-89.

Van der Elst, W., Van Boxtel, M. P., Van Breukelen, G. J., & Jolles, J. (2006b). The Letter Digit Substitution Test: Normative data for 1,858 healthy participants aged 24–81 from the Maastricht Aging Study (MAAS): Influence of age, education, and sex. *Journal of Clinical and Experimental Neuropsychology*, *28*(6), 998-1009.

Weir, J. P. (2005). Quantifying test-retest reliability using the intr-aclass correlation coefficient and the SEM. *Journal of Strength and Conditioning Research*, *19*(1), 231.

Yousefi, F., Shahim, S., Razavieh, A., Mehryar, A. H., Hosseini, A.A., & Alborzi, S. (1992). Some normative data on the Bender Gestalt test performance of Iranian children. *British Journal of Educational Psychology*, *62*(3), 410-416.