RESEARCH ARTICLE

ACTA Vol. 20, No. 4, 2022, 445-470
NEUROPSYCHOLOGICA

Received: 16.06.2022
Accepted: 02.12.2022

A – Study Design
B – Data Collection
C – Statistical Analysis
D – Data Interpretation
E – Manuscript Preparation
F – Literature Search
G – Funds Collection

The Indonesian Neuropsychological Test Battery (INTB): Psychometric properties, preliminary normative scores, the underlying cognitive constructs, and the effects of age and education[1]

Shinta Estri Wahyuningrum[1,2[A,C,D,E,F,G]], Augustina Sulastri[3[B,D,E,G]], M.P.H. Hendriks[1,4[D,E,F]], Gilles van Luijtelaar[1[A,C,D,E,F]]

[1] Donders Centre for Cognition, Radboud University, Nijmegen, The Netherlands
[2] Computer Science Faculty, Soegijapranata Catholic University, Semarang, Indonesia
[3] Psychology Faculty, Soegijapranata Catholic University Semarang, Indonesia,
[4] Academic Centre for Epileptology (ACE), Kempenhaeghe, Heeze, The Netherlands

SUMMARY

Background: Indonesia lacks standardized and adapted neuropsychological tests, which hampers their use in clinical practice. Recently, an Indonesian Neuropsychological Consortium has initiated the adaptation of ten internationally commonly used tests for use in Indonesia. Here, we report the analyses of the psychometric properties, including preliminary normative data, the reliability, the underlying cognitive constructs, and the effects of age and education on these constructs as validity indicators.

Material/ Methods: Four hundred ninety healthy adults living on Java Island participated in this study. All subjects completed the Indonesian Neuropsychological Test Battery (INTB) for diagnosis of various cognitive functions. The test-retest reliability was determined in a parallel study with fifty participants.

Results: Underlying cognitive constructs were assessed with Principal Component Analysis (PCA) revealed seven constructs that accounted for 62.84% of the total variance, and the goodness of fit of the model was good. ANOVA's showed significant effects of age on six constructs (i.e., speed of visuospatial information processing, auditory short-term and working memory, speed and inhibitory control, and verbal learning ability). Age effects were not found for executive internal language. All constructs showed effects of education, except for recall and verbal learning ability.

Conclusions: Interestingly, as expected, not all constructs showed the same age-dependent decline, and if present, all seem to be unique. It is concluded that the psychometric properties of the INTB justify their usage for the Indonesian population.

Key words: Cognitive construct, Indonesian Neuropsychology Test Battery, Psychometric, age-effect, education-effect

[1] The Directorate of Higher Education General of Indonesia with number 0317/AK.04/2022.

INTRODUCTION

A neuropsychological test is a tool that assesses a person's specific cognitive abilities (Elkana et al., 2015; Kessels & Hendriks, 2021; Lezak, Howieson, Bigler, &Tranel, 2012). With rapid modern technologies in brain imaging, neuropsychological tests (NPTs) are imperative to establish disturbances in any specific cognitive function domains. NPTs are ideally suited to establish or complement cognitive diagnostics (Zillmer, Spiers & Culbertson, 2008). Assessment using NPTs also can be used to illustrate the strength and weaknesses of an individual's cognitive patterns by comparing his/her score with scores from healthy subjects with similar demographic characteristics, called normative scores (Zucchella et al., 2018).

Representative normative scores must be recently collected, preferably locally, and based on a sufficient number of healthy subjects (Bridges & Holler, 2007). Locally also implies representing a cross-section of the society regarding demographic characteristics. This is important since age and education factors are two robust characteristics that may affect a client's cognitive functioning (Elkana et al., 2015; Lövdén, Fratiglioni, Glymour, Lindenberger, & Tucker-Drob, 2020; Ktaiche, Fares & Abou-abbas, 2021).Consequently, normative scores for most cognitive tests are corrected for age and level of education, while the need to correct sex effects is less universal and test-specific (Kern et al., 2008). These corrected scores make sure that a client score can be compared with those of a reference group, mimicking the client 's demographics as closely as possible. Each population may have its distinctive cognitive patterns.

Indonesia is an archipelago with a unique culture and a large diverse population (Ananta, Arifin, Sairi, Handayani, & Pramono, 2015), and adapted tests and normative scores for most of the cognitive tests are lacking. This is also the case for the Wechsler Adult Intelligence Scale-IV, that has been adapted in 2014 (Suwartono, Halim, Hidajat, Hendriks, & Kessels, 2014). An Indonesian Neuropsychological Consortium started by developing an Indonesian Neuropsychological Tests Battery (INTB) including collecting normative data on Java Island. The availability of recently collected data in Indonesia is relevant for both experimental and clinical settings. The choice of the ten tests in the INTB is based on their clinical utility and the coverage of relevant cognitive domains, such as various executive functions, reception and production of language, various types of learning and memory, both verbal and visuospatial, and attention and concentration.

Occasionally, subtests from one NPT can reflect more than one cognitive domain (Bialystok, Craik, Binns, Ossher, & Freedman, 2014; Mengual-Macenlle, Marcos, Golpe, & Gonzales-Rivas, 2015; Santos et al., 2015; Tucha,Aschenbrenner, Koerts, & Lange 2012; Nielsen et al., 2018). Previous studies grouped NPTs variables to reveal a specific domain (Chapman et al., 2011; Siedlecki, Honig, & Stern, 2008). For illustration, the construct validity of the Montreal Cognitive Assessment battery, some variables of different tests measure the same cognitive construct (Vogel, Banks, Cummings, & Miller, 2015). Many methods can be used to explore the underlying cognitive constructs of a series of tests. One of them is

Principal Component Analysis (PCA). PCA is data-driven and based solely on analysing the data itself, without having to make any assumptions about the underlying mechanisms. It may reduce high-dimensional data sets to a small number of modes or constructs. The constructs depend on the dataset (Santos et al., 2015).Therefore, they will be specifically meant for interpretation and mainly restricted to the population being studied (Chapmanet al., 2011a; Fong, van Patten, & Fucetola, 2019).

Previous studies employing PCA have shown that clinically significant cognitive constructs are related to particular aspects of cognitive functioning in patients. For example, patients who are diagnosed with various cognitive deficits such as language impairment (Fong et al., 2019), in neurological diseases like Alzheimer's (Chapman, et al., 2011b), in patients with traumatic brain injury (Ravdin & Katzen, 2013) and in a relatively large sample of an outpatient neurology clinic specializing in neurodegenerative diseases (Vogel et al., 2015). These constructs may vary across the life span since normal aging contributes to cognitive decline. However, the consistency of the declining pattern of cognitive abilities has not yet been confirmed in the Indonesian context. The decline is relevant since the trend of aging has increased significantly globally. Indonesia is predicted to have more than 70 million inhabitants aged 60 or older and over 10 million aged 80 or older in 2050 (Adioetomo & Mujahid, 2014). Earlier studies outside Indonesia showed that some domains related to aging, such as processing speed, attention, memory, and executive functioning might decrease with age (Friedman, Miyake, Young, DeFries, Corley, & Hewitt, 2008). At the same time, language functions and visuospatial abilities/construction remain relatively stable (Cohen, Marsiske, & Smith, 2019; Glisky, 2007; Peña-Casanova et al., 2009). Establishing age-related changes also assists in diagnosing various types of pathological declines. Next, the level of education is a significant determinant of cognitive performance, and its effects on the constructs will be explored as well.

This study has three aims: first, to present preliminary normative data on the tests of the INTB and investigate some psychometric properties of the tests in the battery, including the test-retest reliabilities. Secondly, to investigate the underlying cognitive factor structure of the INTB for a healthy population. The third aim is to investigate the effects of age and education on the underlying cognitive constructs, as measured with the INTB. Next is to contribute to the validity of the battery and investigate whether the aging process influences the cognitive constructs.

METHOD

Participants

Four hundred and ninety participants were recruited from three different cities in Java (West Indonesia) to collect normative data and investigate the underlying cognitive factors in the data set. The data were collected in collaboration with two

Table 1. Demographic data of 490 participants, the number of female/male and mean and SD of education level for six age categories

	Age group (year old)					
	16 – 19 (n = 63)	20 – 29 (n = 203)	30 – 39 (n = 73)	40 – 49 (n = 55)	50 – 59 (n = 66)	60 – 69 (n = 30)
Gender (F/M)	36/27	123/80	43/30	41/14	33/33	18/12
Education (M/SD)	13.30 (2.06)	14.96 (1.97)	13.96 (2.85)	13.89 (2.82)	13.36 (3.61)	10.70 (3.50)

universities as consortium members (Wahyuningrum, van Luijtelaar, & Sulastri, 2021). Exclusion criteria were age less than 16 years and self-reported history of any developmental or acquired brain injury or brain-related diseases. All participants had been informed of their privacy rights and knew that their data would be used for scientific purposes. The ethical issue compliance was in line with the WMA Declaration of Helsinki and approved by the local ethics committee of Soegijapranata Catholic University. The participants received seventy-five thousand rupiahs (equal to five US dollars) after finishing series of tests.

The demographic information of the participants is presented in Table 1. Participants were categorized into six age categories by decades, except for the first and last category. The level of education consisted of four categories, according to the Indonesian educational system: zero to nine years (n = 49), ten to twelve years (n = 153), thirteen to fifteen years (n = 267), and over seventeen years (n = 21). The number of participants from the three cities was: Jakarta (n = 192), Semarang (n = 197), and Surabaya (n = 101). Fifty additional participants, 26 females and 24 males, were recruited to investigate the test-retest reliability. The mean age of these latter participants was 37.46 years (SD = 11.93, min = 21, max = 64), and the mean score of years of education was 16.70 years (SD = 2.70, min = 9, max = 22).

Measurement and Assessment

The INTB consists often paper and pencil-based tests conducted using the official Indonesian language, “Bahasa Indonesia”. All participants can speak and read Bahasa Indonesia at least at a primary school level. Participants were allotted approximately two hours to complete the series of the ten tests. The tests are presented in the same order and highly standardized for all participants. Trained research assistants were students in the final year of their psychology study, called the tester of the INTB. All ten tests were administered and scored according to standard instructions.These tests were presented in the following order:

Digit Span Test (DS). The Lezak et al. (2004) version was used. The participant was asked to repeat increasing spans of digits in the right order as instructed by the tester (Lamar, Swenson, Penney, Hospital, & Libon, 2018). The dependent variables are the correct number of correct recall of the digits in the forward, backward, and sequence conditions. The DS was used to measure working memory for auditory stimuli.

Rey Auditory-Verbal Learning Test was adapted from RAVLT. The test measures verbal episodic memory, learning over trials, and short- and long-term recall (after 20 minutes delay) (Strauss & Spreen, 1991). There were seven trials in two parts: part one was the learning and memorizing condition of 15 words nouns (list A) presented five times (A1-A5), followed by a distraction in the form of the presentation of 15 different words (list B). In part two, subjects were asked to recall the 15 words from list A (A6), and after a 20-minute delay, they were asked once more to recall list A (A7). Four variables were used: first, the mean score of trial 1 to trial 5. Second, a learning score over the five trials (LOT). This was calculated by dividing the difference in the scores between trial five and trial one by five. The third variable (A6) is short-term retention (STPR), expressed as the percentage of recalled items of A5. The last variable is long-term retention (LTPR); this delayed recall score was expressed as a percentage of A5 as well. The test was adapted for Indonesia by the replacement of the majority of the words with more familiar ones to Indonesian people. We translated and adapted the words of the RAVLT created by Geffen et al. (1994) using direct translation from English, and some words were chosen by their closest meaning rather than the number of syllables/pronunciations/ phonemes of the words and also by the familiarity of the terms for Indonesians (Utami et al., 2022).

The long version of the *Boston Naming Test (BNT)* originated from Kaplan et al. 2001 and was adapted for Indonesia by Sulastri et al. (2019). This test, measuring verbal naming ability, consists of 60 black and white drawings of objects. Participants are asked to name the objects within 20 seconds. If they fail, a phonemic and semantic cue is given, respectively. Both the total number of correct items and the total time to complete the test are the most commonly used dependent variables.

Ruff's Five Point Test (FP) is an executive function test. A participant is instructed to connect two or more dots with straight lines to create a unique design. The time is limited to 3 minutes. The goal score is the total number of unique designs created by the participant, as well as the number of perseveration errors (the number of repeated designs).

The Trail Making Test (TMT) (Reitan & Wolfson, 1955) is used to measure executive functioning and divided attention. TMT is divided into two parts, part A and part B, and both measure the time spent completing the test. In part A, the participant draws a line connecting circles within numbers 1-25. Furthermore, in part B, the participant must connect two sets of stimuli (number and letter) in an alternating sequence. The dependent variables were both time consumption of part A and part B.

The *Verbal Fluency test (VF)* evaluates the spontaneous production of words under restricted search conditions (Strauss & Spreen, 1991), and the phonemic version that has been used in this study is also considered as an executive function test. In this test, the participants have to produce words of three phonemic categories. For each category, the subject is given 60 seconds. Participants are

asked to produce as many words that begin with the letter K, second with the letter S, and third with the letter T. The constraints are not to mention a name of a person or place. The number of the correct words for each phonemic category was used as the dependent variable, as well as the sum of the three categories. The test was adapted for Indonesia, and the choice of the three letters was based on Hendrawan and Hatta (2010).

The *Stroop Colour Word Test (S)*, based on Stroop (1935), is the coloured-word interference version and measures the inhibition of an overlearned response, the sensitivity to interference, and mental flexibility. It is considered to be an executive function test. The procedure consists of three conditions. In the first condition, when presented with the first sheet or card, the participant is instructed to name colour patches. When the second card is presented, the participant is asked to read words that denote colours printed in black ink. Card 3 shows colour names printed in a different colour, and the participants are asked to name the ink colour in which colour names are printed (card 3) (Strauss & Spreen, 1991).

The *Figural Reproduction (FR)* test was based on and adapted from the Visual Reproduction test (de Brito-Marques, Cabral-Filho, & Miranda, 2012). The test consists of three geometrical pictures. This test measures the skills of short-term visual-spatial memory and reproduction of visual stimuli. Total score (correct reproduced items) and total time to finish the test were used as dependent variables.

The *Bourdon Wiersma test (BW)* measures concentration and sustained attention. The test consists of 50 lines for each with 25 groups of dots with a varying (three to five) number. Participants were instructed to mark the group with four dots. The dependent variables were the (mean) time to finish the rows and the number of errors (both misses and false positives).

The *Token Test (T)* originates from De Renzi and Vignolo (1962). A language comprehension test assesses comprehension of verbal commands of increasing complexity (Strauss & Spreen, 1991). The participant needs to follow the instruction given by the tester correctly. The Token test's material consists of 20 circles and squares with four different colours. As a dependent variable, we used the number of error items.

Statistical Analysis

The reliability of the tests was determined in three different ways. First, by using *a test-retest method* (n = 50), with an interval of seven to fourteen days. The first and second test administration were correlated (*Pearson's correlation coefficient*) and compared using Student's *t-test* for dependent groups. Next, the *Intraclass Correlation Coefficient* (ICC) and *Standard Error of Measurement* (SEM) were used as reliability coefficients. ICC was used to measure the internal consistency reliability or discrepancy between subjects, and SEM was defined as the determination of the amount of variation or spread in the measurement errors for a test and is subsequently an indicator of the reliability of a test (Geerinck, Alekn, Beaudart, Bautmans, Cooper, De Souza Orlandi, 2019). Low levels

of SEM (close to 0) indicate high levels of score accuracy, and high level of SEM (close to the SD of the observed score) indicates low levels of score accuracy. SEM for test and retest was calculated by the difference of standard deviation between the scores of the two assessments divided by the square root of 2 (Geerinck et al., 2019; Palta et al., 2011). ICC values were interpreted as 'excellent' when > 0.90, as 'good' between 0.75-0.90, as "moderate" between 0.50 and 0.75 and 'poor' when < 0.50 (Koo & Li, 2016).

Both data sets (n = 490 and n = 50) were checked upon impossible scores and coding errors. Both errors were traced by scatter plots of each variable and by checking the minimum and maximum scores. Outliers regarding poor performance were identified with the 3-times *standard deviation* (SD) rule (Hermens et al., 2013). Among 490 participants we identified 0.2% (64 cells) with data more than 3-times the SD and replaced it with a mean score corresponding to scores of those with a similar age and education level.

Preliminary normative scores are presented using *mean*, *standard deviation*, *median*, and *percentile scores* of twenty-four variables of the ten NPTs. All INTB variables were standardized to *z-scores* (with a mean of zero and an SD of 1) as our dataset for factor analysis. Furthermore, eleven variables representing error scores or time to complete the test were reversed by multiplying the z-score value by minus one.

All dependent test variables were chosen considering the often studied and used in clinical practice, cognitive domains of executive function, learning and memory (both verbal and visuospatial), language (both production and comprehension), and various attention tasks measuring both speed and accuracy. The underlying cognitive constructs of the INTB were examined using *a PCA with orthogonal rotation* (*varimax*). Before conducting the PCA, three assumptions were tested: *the Kaiser-Meyer-Olkin* (KMO) test as a measure of sampling adequacy, *Bartlett's test of sphericity* to demonstrate the viability of the analysis, and *the communality* reflecting the shared variance among the included variables (it should be higher than 0.30). *Bartlett's method* to reveal unbiased scores was chosen to calculate factor scores, next the rotation using the *Orthogonal method*. Decision regarding the number of factors included in the model were Eigenvalues higher than one, the amount of variance accounted for, as well as the interpretability of the constructs. JASP (an open-source program for statistical analysis supported by the University of Amsterdam, https://jasp-stats.org/) was used to conduct confirmatory factor analysis on the chosen model obtained with PCA.

Finally, *a two-factor Multivariate Analysis of Variance* (MANOVA) followed one factor ANOVA's and *Bonferroni post-hoc tests*were employed to investigate the effect of age and education as between-subject factors on the underlying cognitive constructs. Considering the age groups' education differences as shown in Table 1 and the age differences between the education groups (data not presented) we used ANCOVA's to control for these factors in the *post-hoc*one factor ANOVA'*s. The post-hoc tests* were aimed to establish both global differences between age groups (global decline) as well as significant declines between two

neighbouring age categories, namely a fast decline. The latter is indicative of rather large declines between adjacent age groups. The *mean* and the *standard error* of each age and education category for each cognitive construct or factor are presented in Figure 1. In addition, the *orthogonal polynomial contrasts* were used to describe the age-dependent changes for each of the cognitive constructs, more precisely, whether a significant amount of variance could be explained by linear, quadratic, or cubic trends. As a rule of thumb, to determine the effect size, we used *eta squared* (η^2), a value of < 0.06 is a small effect, between 0.06 and 0.14 is a medium-size effect, and > 0.14 is a large effect (Cohen, 1988).

RESULTS

Preliminary Normative Score

The preliminary normative scores for Indonesian peopleliving in urban parts of Java for INTB tests are shown in Table 2. The table gives the mean, standard deviation, median, minimum, maximum, and 5 and 95 percentile scores.

Reliability measures

Performance at the individual level of the ten tests in the two trials and reliability coefficients are presented using ICC, SEM, and Pearson correlation 'r' in Table 3.

Nine tests showed moderate to excellent reliability on the ICC index (0.60-0.91). However, it was poor for three of the four RAVLT variables learning over trials, short-term percent retention, and long-term percent retention (under 0.25). Only the mean performance A1-A5 (0.87) showed excellent reliability, which confirms the outcomes of a previous study (de Sousa Magalhaes, Malloy-Diniz, & Hamdan, 2012). We used the same version of the RAVLT to determine the test-retest reliability, others used a parallel version. Therefore, learning over trials and both recall scores were affected by the performance one or two weeks earlier. Table 3 also shows the SEM indices for all variables. All the SEM values indicate lower values than the standard deviation of the experimental test. The variables that measured "time" seemed slightly higher, but all of them were below their *standard deviation*.

Pearson correlation, conducted to explore the correlation between scores in assessments one and two, presented in Table 3, shows moderate to strong correlations for all 22 variables used. Only the scores of three variables on the RAVLT test have no high correlation between the two tests.

The differences showed a significant improvement for all variables of the TMT and BNT. While other significantly different between test and retest emerged in variables RAVLT (*Mean* A1-A5 ($t(49) = 11.46$), STPR ($t(49) = 2.45$), LTPR ($t(49) = 2.36$), FR score ($t(49) = 2.72$), FPT Unique number ($t(49) = 5.05$), the total errors of Bourdon-Wiersma ($t(49) = 3.89$), Stroop time card 3 ($t(49) = 3.65$), VFT K ($t(49) = 2.26$) and S ($t(49) = 3.12$).

Table 2. Psychometric of ten tests (N=490)

Assessment	Variable	Mean	SD	Min; max	Median	Percentile (5 and 95)
Digit Span (DS)	Forward Backward Sequence	7.48 6.28 7.74	2.24 2.40 2.89	1; 14 0; 15 0; 16	7.00 6.00 8.00	(4.0; 11.0) (3.0; 11.0) (3.5; 14.0)
RAVLT	Learning over trials Mean A1-A5 Short term recall Long term recall	16.87 10.05 90.01 88.61	7.80 2.01 16.68 17.17	-5; 46 4; 14.6 37.50; 15 36.36; 137.5	17.0 10.3 91.7 90.9	(3.0; 29.0) (6.4; 13.0) (60.0; 116.7) (57.14; 115.4)
Boston Naming Test (BNT)	Total correct Total time	51.37 309	4.95 161	34; 60 60; 848	52.0 260	(42.0; 58.0) (125;643)
Ruff's Five Point Test (FP)	Unique number Perseverance errors	25.97 3.35	9.52 6.19	3; 58 0; 51	25.5 1.0	(10.0; 41.45) (0.0; 13.45)
Trail Making Test (TMT)	Time A Time B Time B - time A Errors A Errors B	44.99 87.52 42.53 0.50 1.48	18.79 50.67 41.75 1.82 3.64	9; 134 17; 426 -51; 340 0; 21 0; 24	41.0 75.0 32.0 0 0	(23.0; 81.0) (41.0;191.25) (6.0; 114.45) (0.; 2.) (0; 10.0)
Verbal Fluency Test (VF)	Letter K Letter S Letter T Total score	14.44 13.72 12.43 40.59	5.02 5.41 4.74 13.36	3; 41 1; 44 1; 33 6; 90	14.5 14.0 12.0 41.0	(7.0; 23.0) (6.0; 23.0) (5.0; 20.0) (6.0; 90.0)
Stroop Coulor Word Test (S)	Card 1 time Card 2 time Card 3 time Time card 3 – card 2 Card 3 correct Total Errors	49.73 58.55 86.93 28.38 97.76 3.29	17.62 17.792.28 19.57 3.01 3.56	10; 240 31; 150 18; 255 -40; 149 81; 100 0; 22	45.5 54.0 80.0 26.0 99.0 2.0	(35.0; 75.0) (41.0; 95.5) (55.6;143.4) (3.44; 62.0) (92.0; 100.0) (0.0; 11.0)
Figural Reproduction (FR)	Total correct Total time	11.87 61.21	2.54 28.54	4; 15 17; 277.9	12.0 54.55	(7.0; 15.0) (29.0; 115.9)
Bourdon Wiersma Test (B)	Mean time Number of errors SD time	11.1 10.1 2.1	2.7 9.6 1.5	6.07; 25.8 0; 58 0.78; 20.6	10.58 7.00 1.75	(7.9; 15.9) (1.0; 30.4) (0.99; 3.95)
Token Test (T)	Total correct Number of errors	146.86 14.13	22.85 17.77	9;163 0; 84	156 7.00	(98.6; 163.0) (0.0; 55.0)

Factor Loadings and Their Interpretations

Bartlett's test showed a significant result ($\chi^2 = 4228.20$; $p < .001$), and *the KMO test* was meritorious (*KMO* = .83) (Beavers et al., 2013), indicating the sampling is adequate, and factors were reliable to use. The communalities for PCA retained were above .30. PCA using the 24 variables indicated that a seven-factor solution accounted for 62.83% of the total variance. The factor loadings are presented in Table 4. Only variables with factor loadings above .30 were considered and included in the interpretation of a construct, as is commonly done.

Confirmatory factor analysis was used to establish the goodness of fit of our seven factors model with parameters of χ^2 (129, N = 490) = 225.637, p < .001; *RMSEA* = 0.040; *TLI* = 0.947 indicating that the seven-factor model fitted the data rather well.

Cognitive factor one comprises five variables: two variables from the TMT test (both times to complete the TMT A and TMT B) and three from two other tests (Five Point and Bourdon Wiersma). All variables in which time to complete the tests was the dependent variable. The last variable that loaded high on this factor was the number of unique designed figures from the Ruff's Five Point (FP) test.

Table 3. Mean and standard deviation of first and second trial, three reliability coefficients of ten tests, and t statistics regards the difference between the first and second trial (N = 50)

Variables	First Test	Second test	ICC	SEM	r	t(49)
	Mean (SD)	Mean (SD)				
DS forward	7.92 (2.08)	8.12 (2.05)	0.87**	0.02	0.78**	1.02
DS backward	6.34 (2.18)	6.76 (2.45)	0.86**	0.19	0.76**	1.84
DS Sequence	8.76 (2.59)	9.46 (3.20)	0.73**	0.43	0.60**	1.87
RAVLT LOT	16.86 (6.42)	9.20 (7.17)	-0.36	0.54	-0.15	-5.24**
RAVLT Mean (A1-A5)	10.27 (1.73)	12.32 (1.98)	0.87**	0.18	0.78**	11.46**
RAVLT STPR	90.12 (15.94)	96.12 (9.22)	0.20	4.75	0.19	2.45*
RAVLT LTPR	89.81 (15.70)	96.96 (16.64)	0.21	0.66	0.10	2.36*
BNT score	56.64 (3.29)	58.54 (2.22)	0.80**	0.76	0.88**	7.82**
BNT time	421.5 (212.9)	233.0 (157.24)	0.70**	39.37	0.84**	-11.44**
FP unique	27.78 (8.23)	32.46 (7.83)	0.73**	0.28	0.67**	5.05**
FP perseverance	5.24 (8.13)	6.46 (10.27)	0.60*	1.52	0.44*	0.87
TMT time A	44.9 (16.1)	38.2 (11.3)	0.72**	3.37	0.66**	3.89**
TMT time B	87.0 (43.7)	73.1 (29.0)	0.82**	10.36	0.81**	3.72*
VF letter K	15.28 (4.60)	16.52 (4.93)	0.79**	0.24	0.67**	2.26*
VF letter S	14.22 (5.30)	16.38 (5.56)	0.71**	0.18	0.59**	3.12*
VF letter T	12.98 (4.67)	14.74 (5.58)	0.76**	0.64	0.65**	2.87*
S card 3 time	86.1 (18.7)	80.8 (19.7)	0.91**	0.73	0.86**	3.65*
S card 3-2 time	25.5 (11.4)	23.8 (12.9)	0.80**	1.05	0.68**	1.26
S card 3 score	99.00 (1.62)	99.24 (1.15)	0.65**	0.33	0.52**	1.19
S total error	1.7 (2.3)	1.2 (1.7)	0.62**	0.48	0.49**	1.76
FR score	11.92 (2.69)	12.64 (2.22)	0.82**	0.33	0.73**	2.72*
FR time	58.8 (35.1)	56.4 (26.2)	0.66**	6.31	0.51**	0.54
B meantime	10.6 (2.5)	10.4 (2.3)	0.92**	0.10	0.85**	1.04
B error	10.3 (11.1)	7.6 (9.0)	0.90**	1.45	0.86**	3.89*
T error	10.1 (18.2)	8.2 (17.1)	0.91**	0.81	0.84**	1.34

The TMT and FP are executive function tests, while the Bourdon Wiersma measures sustained attention. In all three tests, visual-spatial information is presented, and speed is relevant (also, in the FP test, the number of correct items within three minutes is crucial). Therefore, cognitive factor one is thought torepresent **the speed of visuospatial information processing and planning.**

All variables from the DS test loaded exclusively on a single factor with factor loadings ranging between 0.65 to 0.82, also TMT time B (.333), errors of Token test (.395), and RAVLT mean A1-A5 (.482) loaded on this construct. The score of DS forward is generally interpreted as a short-term auditory memory buffer or the phonological loop as part of Baddeley's working memory model and relies

Table 4. Component loadings of the seven factors extracted from PCA

Variables	PC 1	PC 2	PC 3	PC 4	PC 5	PC 6	PC 7
TMT-time A	.772						
FR-time	.709						
B-Mean time	.667						
TMT-time B	.582	.333		.402			
FP-unique number	.528						
DS-backward		.820					
DS-sequence		.739					
DS-forward		.650	.344				
VF-letter K			.836				
VF-letter T			.806				
VF-letter S			.790				
BNT-time				.814			
BNT-score				.775			
FR-score				.615			
T-error		.395		.416			
S-time diff (card 3-card2)					.865		
S-time card 3					.822		
S-score card 3					.506		
B-error					.420		.350
RAVLT-LTPR						.868	
RAVLT-STPR						.839	
RAVLT-LOT							.734
RAVLT-mean (A1:A5)		.482					.511
FP-perseveration							.347
Variance (%) **EigenValue**	**26.35** **6.32**	**7.94** **1.91**	**6.98** **1.67**	**6.19** **1.59**	**5.94** **1.49**	**4.98** **1.43**	**4.45** **1.07**

on attention. The backward and sequence scores are considered and interpreted as the result of the phonological and the central executive parts of Baddeley's working memory model to manipulate the digits from back to forth. The errors of the Token test, TMT time B, and RAVLT mean A1-A5 might also contain attention and working memory component. Cognitive factor two revealed the cognitive ability of **attention, auditory short-term, and working memory**.

The third factor comprised all three variables from the phonemic VF. VF variables loaded in one group is not surprising considering the high correlation of the scores on its three subscales, as established earlier in a smaller sample of Indonesian-speaking subjects (Pesau & van Luijtelaar, 2021). All three-factor loadings were higher than 0.790. Verbal fluency is considered an **executive internal language** function, requiring self-monitoring, inhibition, access to one's lexicon, word fluency. and working memory (Lezak et al., 2012). In that sense, it is quite possible that another executive function task in which auditory information needs to be manipulated, DS forward, also loaded on this construct.

Table 5. The outcome of ANCOVA for age and education and effect sizes for seven PCs as well as an outcome on Bonferroni post-hoc test and presence of orthogonal trends

	Factor	F	p	Eta squared	Post hoc test and presence of orthogonal trends
Age F(5,483)	PC 1	11.01	<.001	.10	16 – 59 > 60+; Linear, Quadratic
	PC 2	10.65	<.001	.10	16 – 29 > 30+ 20 – 29 > 40+ Linear
	PC 3	0.53	.754	.01	n.s.
	PC 4	3.43	<.01	.03	30 – 49 > 60+ Quadratic
	PC 5	3.63	<.01	.04	16 – 29 > 60+ Linear
	PC 6	5.38	<.001	.05	16 – 19 > 60+ 20 – 29 > 30 – 39; 50+
	PC 7	3.01	<.05	.03	20 – 39 > 60+ Quadratic
Education F(4,485)	PC 1	11.27	<.001	.065	0 – 9 < 10+
	PC 2	6.16	<.001	.037	0 – 9 < 10+
	PC 3	19.19	<.001	.106	0 – 12 < 13+
	PC 4	26.40	<.001	.140	0 – 9 < 10+ 10 – 12 < 13 - 16
	PC 5	4.66	<.01	.028	10 – 12 < 13 - 16
	PC 6	1.80	.146	.011	n.s.
	PC 7	0.63	.597	.004	n.s.

Factor four comprised two variables of the BNT, the number of errors in the Token test and the number of correct items in the Figural Reproduction test. All these dependent variables are based on visual stimuli requiring a semantic process. In the BNT, with the high coefficient loading of .814 and .775 for BNT time and the number of correct items, respectively, the production of words is based on visual perception of drawings of objects. In the FR, which loaded .615, subject had to remember a geometrical figure. In the Token test, the number of errors loaded .416 on this factor, the stimuli were rows and columns of geometrical figures with different shapes and colours. All three tests also rely on a semantic process. The fourth construct is, therefore, thought to represent a **visual cued semantic process.**

Three variables of the Stroop test were the difference in time card 3 – time card 2 (.865), time card 3 (.822), the correct number of card 3 (.506). One variable, the number of errors, in the Bourdon-Wiersma, loaded (.420) together into one factor. The high loadings on two speed-related variables, including the clinically most often used difference score between card 3 minus card 2 and the time to complete card 3, and working precise and not being distracted,is helpful in performing the Bourdon-Wiersma test. The fifth factor was thought to represent **speed and inhibitory control**.

Factor six and seven emphasized that two different aspects were measured with the RAVLT test. Two recall variables from RAVLT, LTPR, and STPR, loaded high on factor six (.868 and .839), respectively, and therefore factor six represents **recall ability.** While factor seven got high factor loadings on learning over trials (.734) and the mean scores in the five learning trials (.511), another variable loaded on factor seven was the number of errors from Bourdon test (.350). This seventh factor represents mainly **learning ability.**

MANOVA

A MANOVA was conducted to assess whether the factors (age and education) and their interaction affected the seven PCs. Bivariate scatterplots were checked for multivariate normality first. Subsequent ANOVA's and post-tests indicated more precise age and education effects.

Age Effects

Results from *the MANOVA* indicated significant effects across the PCs for the factors age (*Wilks'* λ = .76, F (35, 1945.89) = 3.71, p < .001, η^2 = .05), education (*Wilks'* λ = .69, F (21, 1327.17) = 8.70, p< .001, η^2 = .12), and the first-order interaction (*Wilks'* λ = .75, F (91, 2888.46) = 1.50, p < .01, η^2 = .04). Subsequent *two-factor ANOVA* and *one-factor ANCOVA* were conducted to explore the effects of demographic factors for each of the PCs. Details regarding which construct was sensitive for age and education, including the effect sizes, are presented in Table 5, as well as outcomes of the *Bonferroni post-hoc tests* and whether there were significant linear and quadratic trends, as established with orthogonal contrasts.

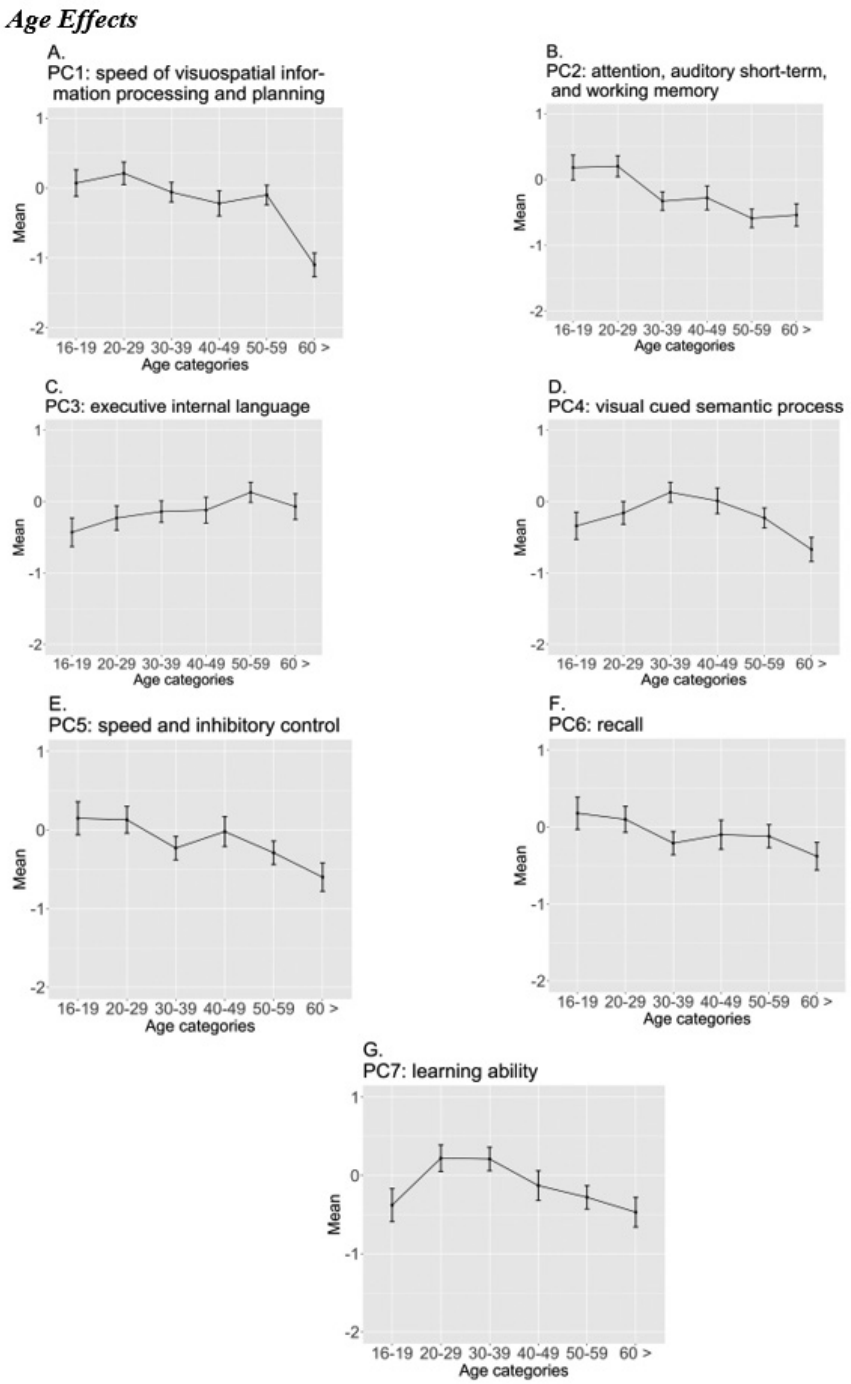

Fig. 1. The mean of the Z-score and standard error for age effects on seven PCs. Four PCs (A, B, E, F) have a significant age effect. Linear trend declines were found for PC 2 (B) and PC 5 (E), and a significant quadratic trend was found for PC 4 (D) and PC 7 (G). Both linear and quadratic trends were found for PC 1 (A)

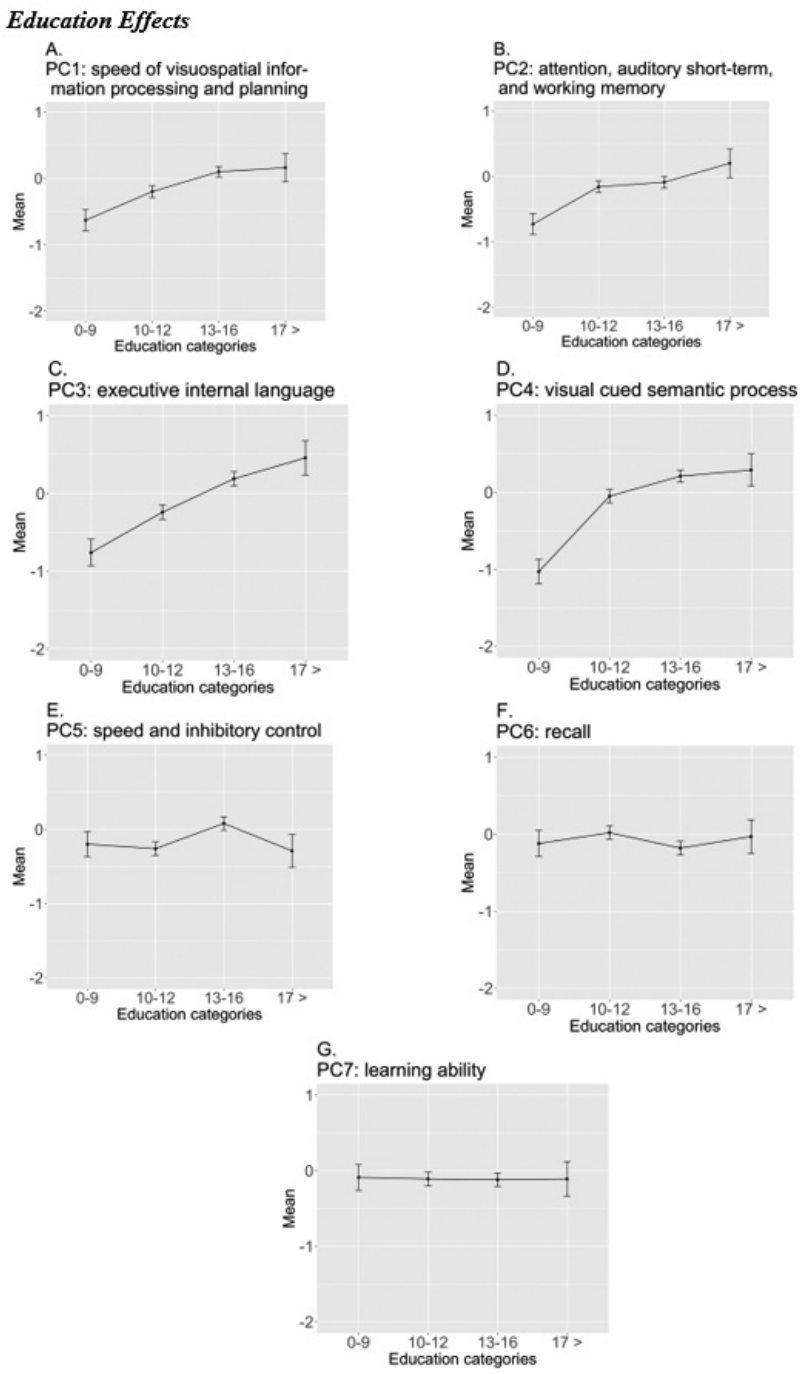

Fig. 2. The mean of the Z-score and standard error of the mean for education affect the seven PCs. Significant education effects were found for PC 1 (A) to PC 5 (E) and showed an increased performance along with the higher education level. PC 6 (F) and PC 7 (G) were not sensitive to education.

As illustrated in Fig. 1, the general tendency wasan age-dependent decline for PC1, PC2, and PC5 as witnessed by a significant linear orthogonal trend for these three PCs. PC1 also showed a significant quadratic trend, representing an accelerated decline after 60. Striking isthat η^2 of PC1 (.10, indicating a medium effect size) was among the largest of all PC, demonstrating that this PC, representing the **speed of visuospatial information processing**, is one of the most age-sensitive cognitive constructs with declining starting at 30 and accelerating around 60. For PC2 (**attention, auditory short-term, and working memory),** a sudden decrease was found after 30 years since the 30+ groups scored less good than the 16-29 year persons. The second difference was between 20-29 and age 40+. The linear decline for PC5 was gradual and lower scores were found for the 60+ group compared to 16-29. For two other PCs, PC4 and PC7, a quadratic trend was found, first showing an age-dependent increase with a maximum at 30-39 for PC4 and 20-39 for PC7, followed by an age-dependent decrease with the lowest scores for the oldest group. PC4 represents **visual cued semantic processes,**and PC7 represents**learning ability.** Orthogonal trend analyses confirmed the lack of significant age effects for PC3 (**executive internal language).** PC6 **(recallability)**showed sudden early decline at age 30 and a further decline at 50+.

Education Effects

Five of the seven PCs were significantly influenced by education with p<.01, PC 6, and PC 7, variables of the RAVLT that represented **recall** and **verbal learning ability** did not show an effect of education. The education-related effects on the PCs are illustrated in Fig. 2. With some variations, all five significant PCs tend to climb along with the educational level. Medium effect sizes were obtained for PC1, PC3, and PC4, while small effect sizes for PC2 and PC5.

Regarding PC 1 (**Speed of visuo-spatial information processing and planning**) and PC 2 **(attention**, **auditory short-term, and working memory)**, all education groups outperformed the group with the least years of education (0-9, junior high) since significant higher scores were obtained for the 10+ (senior high) versus the 0-9 years group. Furthermore, for PC 3 (**executive internal language**), the two lowest (junior and senior high) and two highest (undergraduate and postgraduate) groups differed. While PC4, had the largest effect size(**visual cued semantic process**), there was an increase between 0-9 and 10+ and a further increase between 10-12 and 13-16 years of education. PC 5 (**speed and inhibitory control**), with small effect size, showed a better performance for the 13-16 years compared to those with fewer years of education.

Post-hoc tests following the interaction between age and education for PC2 showed that the senior high school group (*N*=159) showed a significant decline at 30 years, while there was no significant decline for those with an undergraduate degree (*N*=267).

DISCUSSION

Preliminary normative scores

Our first aim was to provide preliminary normative scores for ten NPTs adapted for the Indonesian population (termed as INTB). So far, these scores are lacking, and if these tests are used in clinical practice, normative data from other countries, cultures, and originating from other epochs is insufficient and invalid. This study is the first in which data from a coherent battery of ten internationally well accepted cognitive tests were collected in one test session, allowing us to analyse the coherence among the test scores, and their underlying constructs of the test scores in the battery. The availability of normative scores is crucial for neuropsychologists in Indonesia to help them interpret the scores of their patients or other clients. Nevertheless, mean, median, dispersion measures, and percentile scores are available for 33 variables from 490 healthy participants, representing the urban Indonesian population living in Java Island. We can conclude that the tests in this battery have good reliability, and the results showed that they represent seven independent cognitive constructs. Most of these constructs were sensitive to aging and education, contributing to the validity of the tests and the constructs derived from the tests.

We chose to collect only data in Java because in Java Island lives about 56% of the total population of Indonesia and the increase in urbanization in Java has reached 80% (Sub-directorate of Statistical Demographic, 2013). We assume that the urbanized population in Java Island can be used as a pilot for the development prior to more complete normative scores representing a more comprehensive geographical range of the Indonesian population. Currently, data are collected on three other islands to increase the representation Indonesia's norms and to analyse possible local differences. Normative scores in the form of mean, median, and SD are the parameters for the success of the adaptation of the various test tools.

The normative data from the tests can be compared with data obtained elsewhere: First, DS, both forward and backward, with a similar wide age range were close to previous studies reported in South Africa, Brazilian, and seven European countries (Ostrosky-Solís & Lozano, 2006; Zimmermann, de Cardoso, Trentini, Grassi-Oliveira, & Foncesa, 2015). Our verbal memory scores of the RAVLT were lower for the variable STPR, LTPR, and LOT compared to a study by de Sousa Magalhaes et al. (2019). However, their participants were younger (age ranging between 17-40 years) than ours, and considering the well-known age-dependent decrease in memory, this can be expected. Our average score for trials one to five came close to what was reported by de Sousa Magalhaes et al.(2019).

The Five Point test was similar in the number of participants and age range as studied by Catellani et al. (2011) were associated with producing the unique design. For three other tests, the Token test, Boston Naming Test, and Stroop Color Word Test, the means showed comparable and close to other countries' reports (Ktaiche, Fares, & Abou-Abbas, 2021; Troyer, Leach, & Strauss, 2006). The

performances of our sample on the TMT were a bit slower on trial A but almost the same for trial B compared to the results from a Scandinavian sample of 170 participants of 41-84 years (Espenes et al., 2020). We used words starting with the letters K, S, and T for the adapted phonemic word fluency test based on Hendrawan and Hatta (2010). Although the number of correct words is somewhat dependent on the number of words starting with these letters in Bahasa, there were no large differences between the three subtests and the international scores. Also, the FR reproduction was adapted, and this is the first report of its normative score. Our normative scores were rather similar to an earlier study (de Brito-Marques et al., 2012). The test was conducted in Brazil and reported the correlation between the score of visual reproduction for normal older adults and their education level. The mean row time in seconds of the Bourdon Wiersma test was also comparable with normative scores in a Dutch population with the Indonesian sample was about 2 seconds per row faster. In contrast, Indonesians made more errors than the Dutch (https://andi.nl/tests/aandacht-en-werkgeheugen/bw/) and this might reflect a speed-accuracy trade-off. Finally, our test score from our sample comes close to what is internationally reported. However, it should be kept in mind that a detailed comparison between our outcomes and what is internationally reported is less meaningful considering differences in the assessment's circumstances and demographic factors. Moreover, language performance in Bahasa Indonesia (for many Indonesians, Bahasa Indonesia is their second language) and cultural and socioeconomic factors might also contribute to differences between scores of other countries and our sample.

Reliability of the test battery

A good test-retest reliability and internal consistency coefficient, and standard error measurement of almost all variables of the ten tests of the INTB were found. However, the variables learning over trials and short and long-term retention from the RAVLT showed an ICC value under 0.5 (de Sousa Magalhaeset al., 2012). It might be because we have used the same words in the retest session, instead of a parallel version that is commonly. People cannot learn over trials so much in case their initial performance (A1) is already high in the second assessment. They remembered the items from the previous session. In support, it was found that the amount of learning over trials was the only variable in which a significant decrease was found in the second session compared to the first assessment (16.86 to 9.20). Next, their recall scores on the second assessment were almost perfect, and above 96%. The test-retest reliability of the mean score over the first five trials of the RAVLT was close to 0.8. Therefore, and in agreement with the international literature, we have no reason to doubt that the ICC and test-retest reliability of the RAVLT test is more than sufficient.

Cognitive constructs and age and education effects

Principal Component Analysis was used as an exploratory factor analysis capable of inferring fundamental cognitive construct functions, while the result does

not depend on certain assumptions. We identified seven constructs, some contained a mixture of different variables from different tests, and others were determined by the outcomes of a single test. We identified, among others, speed of visual-spatial information processing, visual cued semantic process, verbal recall, verbal learning, attention/working memory, executive internal language, and inhibitory control, most of which are commonly acknowledged cognitive constructs. The outcomes of confirmatory factor analysis demonstrated the goodness of fit of the seven-factor model. These outcomes contribute to the construct validity of the INTB as a group of tests measuring different aspects of cognition. Earlier, two studies of the Montreal Cognitive Assessment (MoCA) test, covering eight different cognitive domains, explored the underlying factor from a different subset of MoCA items. The earlier study yielded five factors (memory, attention/processing speed, visuospatial, language, and executive function), and the later study yielded four factors (visuospatial/executive function, memory, attention, and language) (Moafmashhadi & Koski, 2013; Vogel et al., 2015). Some of our results resemble those reported by these authors with one obvious difference regarded the language tests. In INTB, we used three language tests that measure different aspects. The PCA showed that the three language tests loaded on two different constructs: VF loaded on the executive internal language function. In contrast, the BNT and Token loaded on a construct named visual cued semantic process together with FR. Interestingly, also the TMT-time B loaded in this factor. In all these four tests, the presentation of visual stimuli is followed by a motor act (drawing, speaking, and pointing). The distinction between executive internal language and the visual cued semantic process is more often found in the psycho-linguistic literature; differences between semantic and executive aspects and deficits of language function have been found in different patient categories (Dick et al., 2001).

Although the outcomes of the PCA also showed that there were no clear superfluous tests that showed a large overlap with other tests in our battery, except that the single letter category might replace the three VFT tests, several measures demonstrated multiple associations. We found five variables that also loaded in more than one factor with a not large (smaller or close to .40) coefficient. Variable TMT time B, errors of the Token Test, and RAVLT (mean A1-A5) also loaded on PC2 (attention, short-term and working memory), acknowledging the attentional and working memory aspects of these three variables (Strauss &Spreen, 1991). DS forward also shares the coefficient loading with the VFT, and it might represent an attentional aspect of the fluency task. The moderate loading of the number of errors of the Bourdon Wiersma on the learning ability suggests that an attentional component is also involved here.

Our constructs also fitted partly in the Cattel-Horn-Carrol (CHC) model: a model trying to encompass the theoretical structure of intelligence as broad domains of cognition (van Rentergem et al., 2020). The current CHC taxonomy distinguishes four conceptual groupings (i.e., motor abilities, perceptual processing, controlled attention, and acquired knowledge) on two levels (i.e., speed and level). Our PC1 belongs to the theoretical CHC constructs “perceptual processing and motor

abilities". Our PC2 and PC5 to CHC construct "controlled attention'' with PC2 narrower to working memory. Three of our PCs, PC3, PC6, and PC7, might have fit in the broad cognitive abilities associated with "acquired knowledge" and seem also close to long-term storage and retrieval fluency, while PC7 may reflect also "learning efficiently". PC4 is associated with the three broad abilities "acquired knowledge, perceptual processing, and controlled attention". However, more quantitative analyses are necessary to see how well the data from the I-NTB fit into the CHC model.

The performance patterns convincingly showed heterogeneity regarding whether there is an age-dependent decline or not, the age of the maximum performance, and the age of the beginning of the decline (Hartshorne & Germine, 2015; Lezak et al., 2012). We confirmed their outcomes since six of the seven factors showed an age-dependent decline – three different ages at which a maximum of the cognitive abilities was found. Two patterns regarding the onset of the decline were revealed: one started as early as age 30, and another started much later, at age 60.

Six of the seven constructs showed age-dependent changes, commonly found for most cognitive abilities measured by neuropsychological tests (Lezak et al., 2012). An exception occurred for the factor which measures the executive internal language function. This construct (PC3) is based on the outcome of a word production task, mainly a phonemic verbal fluency task. The executive elements of this test are self-monitoring, inhibition, and working memory. The language function is access to one's lexicon to switch to different semantic categories and evaluate words' spontaneous production. Word production seems well preserved and remains intact for healthy adults (Cohen, Marsiske, & Smith, 2019; Glisky, 2007). The executive internal language might represent a cumulation of language knowledge acquired and well preserved throughout life (crystalized knowledge). Crystalized cognitive ability can be contrasted with a fluid ability (Cohen et al., 2019). Previous studies reported that fluid ability tends to decline from age 20 to 80 while crystalized ability there is an improvement until approximately age 60 (Murman, 2015). We noticed that PC3 has crystallized and fluid abilities elements and found an improvement until age 50. PC 4 also has mixed abilities with BNT as crystalized and FR and Token as fluid, and this might be the reason for the improvement from 16 until age 39 and a plateau until age 49 followed by a decline.

We have six age categories: the youngest group was 16-19 years, followed by decade groups, and ended with the elderly at age 60+. We found three different peak performances across the outcomes of the seven constructs. A first pattern showed a peak in the early adulthood (the twenties),which regarded PC1, PC2 (Hester et al., 2014), PC5, and PC6. A second pattern was found for PC7 with peak performance at age 20-39 (Messinis, Tsakona, Malefaki, & Papathanasopoulos, 2007). The last pattern was typical for only PC4, which peaked at age 30-49. The first pattern, the highest values in the youngest group followed by a monotonic decline, was typical for our PC1 and PC5, and in both tasks, speed was a crucial factor. Tucker-Drob (2019) reported that cognitive abilities such as visuospatial ability and processing speed peak in the early adulthood (the twenties) and decline afterward.

Furthermore, there were two general patterns regarding age at the start of the decline. One onset decline started as early as age 30 for cognitive performance related to attention, auditory short-term and working memory, and recall ability (Lezak et al., 2012). A subsequent second cognitive performance decline started much later at 60. The latter decline was related to inhibitory control, speed of visuospatial information processing and planning, visual cued semantic process, recall, and learning ability. The decline across five different cognitive constructs might partly represent a decline in sensory perception, health, and socioeconomic status (Murman, 2015). It is not uncommon in cross-sectional research to find a steep decline from age 60 onwards (Tucker-drob, 2019). Cohort effects, in our case, fewer years of education in our 60+ group, were not the reason for this decline. All six significant age effects remain present in an ANCOVA, with education as a covariate.

Compared to the age effect, educational experience has, in general, larger effect sizes than age, and on five of the seven constructs, significant education effects were found. Only recall and learning ability were without significant effects. Both constructs come from RAVLT variables. Bolla-Wilson & Bleecker (1986) found that verbal intelligence was associated with RAVLT performance more than years of education.

The significant education effects on the other five PCs align with a large amount of literature showing that education affects almost all cognitive tests (Weber&Skirbekk, 2014; Jansen et al., 2021; Guerra-Carrillo, Katovich, & Bunge, 2017). Of note is that the most significant differences in education were between primary education and senior high for the speed of visuospatial information processing and planning, attention, auditory short term and working memory, and visual cued semantic process. For executive internal language, visual cued semantic process and speed and inhibitory control, having an undergraduate made the difference or a further difference with lower educated groups (Guerra-Carrillo et al., 2017). Further higher education, from undergraduate to postgraduate, did not matter a great deal considering that no significant increases were found between 13-16 versus 17+. This lack of other differences might be ascribed to our sample's not-so-large number of persons with postgraduate education. Two constructs, short and long-term verbal ability and verbal learning ability, are not education-dependent. The ability to learn and remember is not dependent on the years of education after obtaining junior high.

In addition to age and education as main factors, an interaction of age x education effects on the PC measuring "attention, auditory short term, and working memory" was found. As revealed by post-hoc tests, it was found that those with a senior high education (also for those with only junior high) showed a sharp decrease in attention, auditory short-term, and working memory from 30 years onward, while this was not the case for the undergraduates; their decline is non-significant, and if it happened it started later. This emphasizes the relevance of education in the prevention of cognitive decline.

Limitations

A limitation of the current dataset is that the sample is relatively small for elderly people and also the less well-educated groups are poorly represented, while the number of subjects between 20 and 40, and well-educated was rather large. Further, our sample is mostly from urban areas and the population from rural areas was underrepresented.

The preliminary scores are also not adjusted as yet for the commonly used demographic factors age, education, and sex, while it is not clear whether adaptations for the language spoken in public as well as at home and for ethnicity, Indonesian's population consists of many different ethnic groups, are imperative. This awaits the collection of a larger dataset and analyses of the putative role of all these factors on the performance of these tests.

Being aware of the limitations of our sample regarding an uneven distribution of both age and education levels, and that mostly the urban population was assessed, we are convinced that these results can be used as basic references for the cognitive performance of healthy adults in Java. It fills in the lack of normative scores on these ten tests.

Finally, we have to collect data from different clinical populations, such as patients with neurological and psychiatric diseases, to get more insights in the clinical validation of our test battery.

We are currently expanding the neuropsychological dataset with the aid of university partners unified in a consortium. This consortium represents six areas from four islands in Indonesia and will expand in the future. Data storage and calculation of normative scores of the ten tests are accommodated in I-ANDI, a dynamic database, and an online platform (Wahyuningrum et al., 2021).

CONCLUSIONS

We conclude that all test scores were in good agreement with what is internationally reported. The psychometric analyses showed objectives concerning reliability and validity of the tests in the INTB are considered promising. Interestingly, as expected, not all constructs showed the same age-dependent decline, and a somewhat unique age-affected pattern for each of the cognitive constructs was found. Education effects were more significant than age effects. The interaction between education and age underlines the relevance of education in preventing early aging. It is hoped that the INTB can be used for the Indonesian population and that the preliminary normative data reported here may enhance the use of the tests in the future. Indonesia's large ethnic and linguistic diversity is a challenge for more definite normative scores for all Indonesians. Therefore, reluctance in its use on a large scale is still advised until more insight into the role of ethnicity and spoken languages is obtained.

Acknowledgment

This work was supported by the Directorate of Higher Education General of Indonesia with number 0317/AK.04/2022. We thank Prof. Dr. R.P.C. Kessels, Dr.

Vitoria Piai, Dr.Loes van Aken, and Dr. J.M. Oosterman for the valuable feedback and discussion.

REFERENCES

Adioetomo, S. M. and Ghazy M. (2014). Indonesia on the Threshold of Population Ageing. *UNFPA Indonesia Monograph Series*: No.1. https://doi.org/10.1299/kikaic.65.1319

Ananta A, Arifin EN, SairiHasbullah M, Handayani NB, Pramono A (2015) Demography of Indonesia's Ethnicity. Institute of Southeast Asian Studies, Singapore.

Agelink van Rentergem, J. A., de Vent, N. R., Schmand, B. A., Murre, J. M. J., Staaks, J. P. C., Huizenga, H. M., & ANDI Consortium. (2020). The factor structure of cognitive functioning in cognitively healthy participants: A meta-analysis and meta-analysis of individual participant data. Neuropsychology Review, 30(1), 51–96. https://doi.org/10.1007/s11065-019-09423-6

Beavers, A. S., Lounsbury, J. W., Richards, J. K., Huck, S. W., Skolits, G. J., & Esquivel, S. L. (2013). Practical considerations for using exploratory factor analysis in educational research. *Practical Assessment, Research and Evaluation*, 18(6), 1–13.

Bialystok, E., Craik, F. I. M., Binns, M. A., Ossher, L., & Freedman, M. (2014). Effects of bilingualism on the age of onset and progression of MCI and AD: Evidence from executive function tests. *Neuropsychology*, 28(2), 290–304. https://doi.org/10.1037/neu0000023

Bridges, A. J., & Holler, K. A. (2007). How many is enough? Determining optimal sample sizes for normative studies in pediatric neuropsychology. *Child Neuropsychology*, 13(6), 528–538. https://doi.org/10.1080/09297040701233875

Cattelani, R., Dal Sasso, F., Corsini, D., &Posteraro, L. (2011). The Modified Five-Point Test: normative data for a sample of Italian healthy adults aged 16-60. *Neurological sciences: official journal of the Italian Neurological Society and of the Italian Society of Clinical Neurophysiology*, 32(4), 595–601. https://doi.org/10.1007/s10072-011-0489-4

Chapman, R. M., Mapstone, M., McCrary, J. W., Gardner, M. N., Porsteinsson, A., Sandoval, T. C., Guillily, M. D., Degrush, E., & Reilly, L. A. (2011a). Predicting conversion from mild cognitive impairment to Alzheimer's disease using neuropsychological tests and multivariate methods. *Journal of Clinical and Experimental Neuropsychology*, 33(2), 187–199. https://doi.org/10.1080/13803395.2010.499356

Chapman, R. M., Mapstone, M., Porsteinsson, A. P., Gardner, M. N., John, W., Degrush, E., Reilly, L. A., Sandoval, T. C., &Guillily, M. D. (2011b). Diagnosis of Alzheimer's Disease Using Neuropsychological Testing Improved by Multivariate Analyses Robert. *J Clin Exp Neuropsychol*., 32(8), 793–808. https://doi.org/10.1080/13803390903540315.

Cohen, J. (1988). Statistical Power Analysis for the Behavioral Sciences (2nd ed.). Hillsdale, NJ: Lawrence Erlbaum Associates, Publishers.

Cohen, R. A., Marsiske, M. M., & Smith, G. E. (2019). Neuropsychology of aging. *Handbook of Clinical Neurology*, 167, 149–180. https://doi.org/10.1016/B978-0-12-804766-8.00010-8

De Brito-Marques, P. R., Cabral-Filho, J. E., & Miranda, R. M. (2012). Visual reproduction test in normal elderly: Influence of schooling and visual task complexity. *Dementia e Neuropsychologia*, 6(2), 91–96. https://doi.org/10.1590/s1980-57642012dn06020005

De Renzi, E., &Vignolo, L. A. (1962). The token test: A sensitive test to detect receptive disturbances in aphasics. *Brain: a journal of neurology*, 85, 665–678. https://doi.org/10.1093/brain/85.4.665

De Sousa Magalhães, S., Malloy-Diniz, L. F., & Hamdan, A. C. (2012). Validity convergent and reliability test-retest of the Rey Auditory Verbal Learning Test. *Clinical Neuropsychiatry: Journal of Treatment Evaluation*, 9(3), 129–137.

Dick, F., Bates, E., Wulfeck, B., Utman, J. A., Dronkers, N., &Gernsbacher, M. A. (2001). Language deficits, localization, and grammar: evidence for a distributive model of language breakdown in aphasic patients and neurologically intact individuals. *Psychological review*, 108(4), 759-788. https://doi.org/10.1037/0033-295x.108.4.759

Elkana, O., Eisikovits, O. R., Oren, N., Betzale, V., Giladi, N., & Ash, E. L. (2015). Sensitivity of neuropsychological tests to identify cognitive decline in highly educated elderly individuals: 12 months follow up. *Journal of Alzheimer's Disease,* 49(3), 607–616. https://doi.org/10.3233/JAD-150562

Espenes, J., Hessen, E., Eliassen, I. V., Waterloo, K., Eckerström, M., Sando, S. B., Timón, S., Wallin, A., Fladby, T., &Kirsebom, B. E. (2020). Demographically adjusted trail making test norms in a Scandinavian sample from 41 to 84 years. *Clinical Neuropsychologist*, 34(S1), 110–126. https://doi.org/10.1080/13854046.2020.1829068

Fong, M. W. M., van Patten, R., & Fucetola, R. P. (2019). The Factor Structure of the Boston Diagnostic Aphasia Examination, Third Edition. *Journal of the International Neuropsychological Society*, 1–5. https://doi.org/10.1017/S1355617719000237

Friedman, N. P., Miyake, A., Young, S. E., DeFries, J. C., Corley, R. P., & Hewitt, J. K. (2008). Individual Differences in Executive Functions Are Almost Entirely Genetic in Origin. *Journal of Experimental Psychology: General*, 137(2), 201–225. https://doi.org/10.1037/0096-3445.137.2.201

Geerinck A, Alekna V, Beaudart C, Bautmans I, Cooper C, De Souza Orlandi F, et al. (2019). Standard error of measurement and smallest detectable change of the Sarcopenia Quality of Life (SarQoL) questionnaireAn analysis of subjects from 9 validation studies Enhanced Reader. *PLoS ONE*, 14(4), 1–13. https://doi.org/https://doi.org/10.1371/journal.pone.0216065

Glisky, E. (2007). Changes in Cognitive Function in Human Aging. *Brain Aging: Models. Method, and Mechanisms* (Issue April 2007, pp. 3–20). https://doi.org/10.1201/9781420005523.sec1

Geffen, G. M., Butterworth, P., & Geffen, L. B. (1994). Test-retest reliability of a new form of the auditory verbal learning test (AVLT). *Archives of clinical neuropsychology : the official journal of the National Academy of Neuropsychologists*, 9(4), 303–316.

Guerra-Carrillo, B., Katovich, K., & Bunge, S. A. (2017). Does higher education hone cognitive functioning and learning efficacy? Findings from a large and diverse sample. *PLoS ONE* (Vol. 12, Issue 8). https://doi.org/10.1371/journal.pone.0182276

Hartshorne, J. K., &Germine, L. T. (2015). When Does Cognitive Functioning Peak? The Asynchronous Rise and Fall of Different Cognitive Abilities Across the Life Span. *Psychological Science*, 26(4), 433–443. https://doi.org/10.1177/0956797614567339

Hendrawan, D., & Hatta, T. (2010). Evaluation of stimuli for development of the Indonesian version of verbal fluency task using ranking method. *Psychologia*, 53(1), 14–26. https://doi.org/10.2117/psysoc.2010.14

Hermens, D. F., Naismith, S. L., Lagopoulos, J., Lee, R. S. C., Guastella, A. J., Scott, E. M., &Hickie, I. B. (2013). Neuropsychological profile according to the clinical stage of young persons presenting for mental health care. *BMC Psychology*, 1(8), 1–9. https://doi.org/10.1186/2050-7283-1-8

Jansen MG, Geerligs L, Claassen JAHR, Overdorp EJ, Brazil IA, Kessels RPC and Oosterman JM. (2021). Positive Effects of Education on Cognitive Functioning Depend on Clinical Status and Neuropathological Severity. *Front. Hum. Neurosci*. 15:723728. doi: 10.3389/fnhum.2021.723728

Karen Bolla-Wilson & Margit L. Bleecker (1986) Influence of verbal intelligence,sex, age, and education on the rey auditory verbal learning test, Developmental *Neuropsychology*,2:3, 203-211, DOI:10.1080/87565648609540342

Kern, R. S., Nuechterlein, K. H., Green, M. F., Baade, L. E., Fenton, W. S., Gold, J. M., Keefe, R. S. E., Mesholam-Gately, R., Mintz, J., Seidman, L. J., Stover, E., & Marder, S. R. (2008). The MATRICS Consensus Cognitive Battery, part 2: Co-norming and standardization. *American Journal of Psychiatry*, 165(2), 214–220. https://doi.org/10.1176/appi.ajp.2007.07010043

Kessels, R.P.C. & Hendriks, M.P.H. (2021). Neuropsychological Assessment. Encyclopedia of Mental Health. 197-201. https://doi.org/10.1016/B978-0-323-91497-0.00017-5

Koo, T. K., & Li, M. Y. (2016). A Guideline of Selecting and Reporting Intraclass Correlation Coefficients for Reliability Research. *Journal of Chiropractic Medicine*, 15(2), 155–163. https://doi.org/10.1016/j.jcm.2016.02.012

Ktaiche, M., Fares, Y., &Abou-Abbas, L. (2021). Stroop color and word test (SCWT): Normative data for the Lebanese adult population. *Applied Neuropsychology:Adult*, 0(0), 1–9. https://doi.org/10.1080/23279095.2021.1901101

Lamar, M., Swenson, R., Penney, D., Hospital, L., &Libon, D. (2018). Encyclopedia of Clinical Neuropsychology. *Encyclopedia of Clinical Neuropsychology*, January. https://doi.org/10.1007/978-3-319-56782-2

Lezak, M. D., Howieson, D. B., Loring, D. W., Hannay, H. J., & Fischer, J. S. (2004). Neuropsychological Assessment (4th ed.). Oxford University Press.

Lezak, M. D., Howieson, D. B., Bigler, E. D., &Tranel, D. (2012). Neuropsychological Assessment (5th ed.). Oxford University Press.

Lövdén, M., Fratiglioni, L., Glymour, M. M., Lindenberger, U., & Tucker-Drob, E. M. (2020). Education and Cognitive Functioning Across the Life Span. *Psychological Science in the Public Interest*, 21(1), 6–41. https://doi.org/10.1177/1529100620920576

Mengual-Macenlle, N., Marcos, P. J., Golpe, R., & González-Rivas, D. (2015). Multivariate analysis in thoracic research. *Journal of Thoracic Disease*, 7(3), E2–E6. https://doi.org/10.3978/j.issn.2072-1439.2015.01.43

Messinis, L., Tsakona, I., Malefaki, S., &Papathanasopoulos, P. (2007). Normative data and discriminant validity of Rey's Verbal Learning Test for the Greek adult population. *Archives of clinical neuropsychology: the official journal of the National Academy of Neuropsychologists*, 22(6), 739–752. https://doi.org/10.1016/j.acn.2007.06.002

Moafmashhadi, P., & Koski, L. (2013). Limitations for interpreting failure on individual subtests of the Montreal Cognitive Assessment. *Journal of Geriatric Psychiatry and Neurology*, 26(1), 19–28. https://doi.org/10.1177/0891988712473802

Murman, D. L. (2015). The Impact of Age on Cognition. *Seminars in Hearing*, 36(3), 111–121. https://doi.org/10.1055/s-0035-1555115

Nielsen, T. R., Segers, K., Vanderaspoilden, V., Bekkhus-Wetterberg, P., Minthon, L., Pissiota, A., Bjørkløf, G. H., Beinhoff, U., Tsolaki, M., Gkioka, M., & Waldemar, G. (2018). Performance of middle-aged and elderly European minority and majority populations on a Cross-Cultural Neuropsychological Test Battery (CNTB). *Clinical Neuropsychologist*. https://doi.org/10.1080/13854046.2018.1430256

Ostrosky-Solís, F., & Lozano, A. (2006). Digit Span: Effect of education and culture. *International Journal of Psychology*, 41(5), 333–341. https://doi.org/10.1080/00207590500345724

Palta, M., Chen, H. Y., Kaplan, R. M., Feeny, D., Cherepanov, D., &Fryback, D. G. (2011). Standard error of measurement of 5 health utility indexes across the range of health for use in estimating reliability and responsiveness. *Medical Decision Making*, 31(2), 260–269. https://doi.org/10.1177/0272989X10380925

Peña-Casanova, J., Blesa, R., Aguilar, M., Gramunt-Fombuena, N., Gómez-Ansón, B., Oliva, R., Molinuevo, J. L., Robles, A., Barquero, M. S., Antúnez, C., Martínez-Parra, C., Frank-García, A., Fernández, M., Alfonso, V., & Sol, J. M. (2009). Spanish multicenter normative studies (NEURONORMA project): Methods and sample characteristics. *Archives of Clinical Neuropsychology*, 24, 307–319. https://doi.org/10.1093/arclin/acp027

Pesau, H. G., &Luijtelaar, G. van. (2021). Equivalence of Traditional and Internet-Delivered Testing of Word Fluency Tasks. *JurnalPsikologi*, 20(1), 35–49. https://doi.org/10.14710/jp.20.1.35-49

Ravdin, L. D., &Katzen, H. L. (2013). Handbook on the neuropsychology of aging and dementia. *Handbook on the Neuropsychology of Aging and Dementia*. https://doi.org/10.1007/978-1-4614-3106-0

Reitan, R. M., & Wolfson, D. (1995). Category Test and Trail Making Test as measures of frontal lobe functions. *Clinical Neuropsychologist*, 9(1), 50–56. https://doi.org/10.1080/13854049508402057

Kessels, R.P.C.& Hendriks, M.P.H. (2021). Neuropsychological Assessment. Third Edition.

Santos, N. C., Costa, P. S., Amorim, L., Moreira, P. S., Cunha, P., Cotter, J., & Sousa, N. (2015). Exploring the factor structure of neurocognitive measures in older individuals. *PLoS ONE*, 10(4), 1–18. https://doi.org/10.1371/journal.pone.0124229

Siedlecki, K. L., Honig, L. S., & Stern, Y. (2008). Exploring the structure of a neuropsychological battery across healthy elders and those with questionable dementia and Alzheimer's disease. *Neuropsychology*, 22(3), 400–411. https://doi.org/10.1037/0894-4105.22.3.400

Suwartono, C., Halim, M. S., Hidajat, L. L., Hendriks, M. P. H., &Kessels, R. P. C. (2014). Development and Reliability of the Indonesian Wechsler Adult Intelligence Scale—Fourth Edition (WAIS-IV). *Psychology*, 5, 1611-1619. http://dx.doi.org/10.4236/psych.2014.514171.

Sulastri, A., Utami, M. S. S., Jongsma, M., Hendriks, M., & van Luijtelaar, G. (2019). The Indonesian Boston Naming Test: Normative data among healthy adults and effects of age and education on naming ability. *International Journal of Science and Research*, 8(11), 134–139.

Strauss, E., &Spreen, E. M. S. S. O. (1991). A Compendium of Neuropsychological Tests: Administration, Norms, and Commentary. In *OXFORD UNIVERSITY PRESS* (Vol. 41, Issue 11). https://doi.org/10.1212/wnl.41.11.1856-a

Troyer, A. K., Leach, L., & Strauss, E. (2006). Aging and response inhibition: Normative data for the Victoria Stroop Test. *Aging, Neuropsychology, and Cognition*, 13(1), 20–35. https://doi. org/10.1080/138255890968187

Tucha, L., Aschenbrenner, S., Koerts, J., & Lange, K. W. (2012). The Five-Point Test: Reliability, Validity and Normative Data for Children and Adults. *PLoS ONE*, 7(9), 1–11. https://doi.org/10.1371/journal.pone.0046080

Tucker-Drob E. M. (2019). Cognitive Aging and Dementia: A Life Span Perspective. *Annual review of developmental psychology,* 1, 177–196.https://doi.org/10.1146/annurev-devpsych-121318-085204

Utami, M.S.S, Santoso, J.B. Suryani, A.O. Goeritno, H., Widhianingtanti, L.T., Sulastri. A., van Luijtelaar, G.(2022). Adaptation of Rey Auditory Verbal Learning Test for Indonesia Word Order Accuracy, Validity, and Reliability. *In Preparation*.

Vogel, S. J., Banks, S. J., Cummings, J. L., & Miller, J. B. (2015). Concordance of the Montreal cognitive assessment with standard neuropsychological measures. *Alzheimer's and Dementia: Diagnosis, Assessment and Disease Monitoring, 1,* 289–294. https://doi.org/10.1016/j.dadm. 2015.05.002

Wahyuningrum, S. E., van Luijtelaar, G., &Sulastri, A. (2021). An online platform and a dynamic database for neuropsychological assessment in Indonesia. *Applied Neuropsychology:Adult*, 0(0), 1–10. https://doi.org/10.1080/23279095.2021.1943397

Weber, D. &Skirbekk, V. (2014). The Educational Effect on Cognitive Functioning: National versus Individual Educational Attainment. *IIASA Interim Report*. IIASA, Laxenburg, Austria: IR-14-008

Zillmer, E. A, Spiers, M. V, & Culbertson, W. C. (2008). Principles of neuropsychology. Higher Education, 574. Retrieved from http://books.google.com/books?id=wIk1PwAACAAJ&pgis=1

Zimmermann, N., de Cardoso, C. O., Trentini, C. M., Grassi-Oliveira, R., & Fonseca, R. P. (2015). NormasBrasileirasPreliminares E AInvestigação Dos Efeitos De Idade E Escolaridade No Desempenho Dos Testes Wisconsin De Classificação De Cartas Modificado, Stroop De Cores E Palavras E DígitosEmAdultos. *Dementia e Neuropsychologia, 9(2),* 120–127. https://doi.org/10.1590/1980-57642015DN92000006

Zucchella, C., Federico, A., Martini, A., Tinazzi, M., Bartolo, M., &Tamburin, S. (2018). Neuropsychological testing. *Practical Neurology, 18(3)*, 227–237. https://doi.org/10.1136/practneurol-2017-001743

Corresponding author:
Augustina Sulastri
Psychology Faculty, Soegijapranata Catholic University Semarang, Indonesia
Jl. Pawiyatan Luhur IV/1, Bendan Dhuwur, Semarang, Jawa Tengah, Indonesia, 50234
Email: ag.sulastri@unika.ac.id
ORCID:0000-0002-0107-7590