# The Development and Psychometric Properties Evaluation of the Indonesian version of the *Wechsler* Adult Intelligence Scale – Fourth Edition (WAIS–IV)

Christiany Suwartono

The studies described in this thesis financially supported by research grants from Indonesian Ministry of Education and Culture specifically the International Research Collaboration and Scientific Publication number: 085/SP2H/PL/DIT.LITABNAS/V/2013, 207/K3/KM/2014, and 094/K3/KM/2015. 

ISBN: 978-94-9289-673-5
Cover Artist: Aa Sugriwa
Cover Design/Layout: Dennis Hendriks || ProefschriftMaken.nl
Print: ProefschriftMaken.nl

# The Development and Psychometric Properties Evaluation of the Indonesian version of the Wechsler Adult Intelligence Scale – Fourth Edition (WAIS–IV)

Proefschrift

ter verkrijging van de graad van doctor
aan de Radboud Universiteit Nijmegen
op gezag van de rector magnificus prof. dr. J.H.J.M. van Krieken,
volgens besluit van het college van decanen
in het openbaar te verdedigen op maandag 26 november 2018
om 10.30 uur precies

door
Christiany Suwartono
geboren op 29 oktober 1982
te Purwokerto, Indonesië

**Promotor:**
prof. dr. R.P.C. Kessels

**Copromotoren:**
dr. M.P.H. Hendriks
dr. M.S. Halim (Universitas Katolik Indonesia Atma Jaya, Indonesië)

**Manuscriptcommissie:**
prof. dr. E.L.J.M. van Luijtelaar
prof. dr. J.J.L. Derksen (Vrije Universiteit Brussel, België)
prof. dr. A.J.R. van de Vijver (Tilburg University)

## CONTENTS

# 1.
# Introduction

Psychological assessment comprises a series of many tests carried out by psychologists and psychological services for various purposes in a community. Unfortunately, the psychometric properties of available assessments in Indonesia, such as the validity, reliability, and even the standardization, are often based on outdated data or are entirely omitted in the administration manuals of these tests (Suwartono & Santoso, 2016). These issues can contribute to psychological measurement errors when assessing clients or patients.

Indonesia is the fourth most populous country in the world, with around 258.7 million inhabitants in 2016 (Badan Pusat Statistik Indonesia (BPS), 2017a). Indonesia is the largest archipelago country in the world, comprising 17,504 islands, although the majority (57.45%) of its population lives on Java. There are six large islands in Indonesia; Kalimantan, Maluku-Papua, Sumatera, Sulawesi, Java, and Bali-Nusa Tenggara.

The multicultural Indonesian society includes about 700 regional languages (BPS, 2009) and 1,340 ethnic groups (BPS, 2017b), 15 of whom are composed of more than one million people (Suryadinata, Arifin, & Ananta, 2003). Its multicultural population is the subject of Indonesia’s national motto, ‘Bhinneka Tunggal Ika’, which means ‘Unity in Diversity’. To this end, the Youth Pledge declared in 1928, wherein the young generation agreed to have one motherland, one nation, and one language, despite their different cultures and backgrounds (Latuconsina & Rafidi, 1996). Consequently, people in Indonesia communicate using the Indonesian language, although they are also fluent in their diverse regional languages, such as Javanese, Sundanese, and others. Regional languages are commonly used at home or in local communities, while the Indonesian language is applied in formal education and national media (Lewis, Simons, & Fennig, 2013). The official Indonesian language is a modernized form of Malay (Brown, 2003), an Austronesian language (Lewis et al., 2013), with influences from a range of other languages, including Arabic, Portuguese, Dutch, Chinese, and Javanese (Errington, 1986).

The population of Indonesia increased by approximately 3.5 million people each year between 2010 and 2016 (BPS, 2017a). This growing population requires the development of competitive human resources in the employment sector. Employers must be able to identify individuals who have exceptional skills to achieve a competitive advantage; therefore, developing tools to measure aptitude, ability, and other characteristics is crucial for selectively expanding their human resources. The role of psychological testing in identifying desirable candidates is an important consideration (Thomas & Scroggins, 2006). Psychological testing can also be used to improve education and even mental health treatment and can help decision-makers to avoid stressful situations and achieve their intended goal efficiently and objectively. The availability of valid and reliable standardized psychological tests is vital; however, Indonesia has some issues in the field of psychological measurement which leads to inadequacies in the psychological tools available (for details, see Suwartono & Santoso, 2016 for detail).

Standardized intelligence testing is one of the most significant accomplishments in Psychology and is one of the most persistent and widely used inventions (Benson, 2003). Intelligence testing is commonly used as part of a comprehensive psychological assessment in Indonesia, especially in school settings (Rohmah, 2011). Both group and individual intelligence testing scales are available in Indonesia. The Wechsler intelligence scales are individual intelligence assessments, which are well known among psychologists in Indonesia (LPSP3, 2015; PERSONA, 2015). The first of these scales, the Wechsler-Bellevue Intelligence Scale (WBIS), was developed in 1939, and it is still taught and used in Indonesia today (Suwartono, Hidajat, Halim, Hendriks, & Kessels, 2014). Internationally, however, an updated version of Wechsler's intelligence scale has been implemented; the Wechsler Adult Intelligence Scale – Fourth Edition (WAIS-IV; Wechsler, 2008a,b).

In the following section, the theoretical concepts and intelligence assessment will be discussed. The Wechsler intelligence scales used both worldwide and in Indonesia are described, and the structure of the Indonesian WAIS-IV will be described in detail. Finally, the studies covered in subsequent chapters of this thesis will also be outlined.

## INTELLIGENCE

In 1904, Charles Spearman proposed the first theory of intelligence, suggesting that separate abilities are relatively highly correlated and that the resulting pattern of intercorrelations could be explained by a single underlying factor term, g, or general cognitive ability, which determines a person's performance in all intellectual tasks (Deary, 2001; Raven, 2008). The g factor comprises two analytically distinct factors; eductive ability (the ability to determine the meaning) and reproductive ability (the ability to reproduce accurate information and learned skills). However, Spearman did not claim that these were separated factors; rather than he argued that there were analytically distinct components of "g" (Raven, 2008). Within academic psychology, Spearman's theory of general intelligence remains the dominant concept of intelligence (Deary, Strand, Smith, & Fernandes, 2007; Jensen, 1993).

After Spearman's theory, Thurstone identified seven independent primary mental abilities (PMAs) based on factor analysis; word fluency, verbal comprehension, spatial visualization, number facility, associative memory, reasoning, and perceptual speed; however, Thurstone later acknowledged that these PMAs were not wholly independent (Mackintosh, 2011). In the mid-1960s, Cattell proposed that Spearman's g should be divided into two distinct but related factors: fluid (Gf) and crystallized intelligence (Gc). Gf involves concept formation and attainment, reasoning, and abstracting, while Gc is defined as the ability to use education, and is measured by tests of academic knowledge, such as vocabulary (Horn & Cattell, 1966; 1967).

Investigations into the theories of intelligence are still ongoing. Vernon proposed a verbal-perceptual model, acknowledging the contribution of general intelligence, but stating that ability tests to determine g result in only two main groups of abilities, the v:ed and k:m factors (Johnson & Bouchard, 2005; Vernon, 1965). The v:ed factors are the verbal and educational abilities arising from academic experiences, which consist of verbal fluency, divergent thinking, and numeracy. The k:m factors are spatial, practical, and mechanical abilities, developed during non-educational experiences. They include perceptual speed, psychomotor abilities, and physical abilities such as proprioception, spatial, and mechanical abilities.



Carroll (1993) performed a systematic exploratory meta-analysis of more than 460 data sets of intelligence test battery. Carroll's theory was called Three-Stratum model. He proposed three hierarchical strata of abilities (narrow, broad, general) divided by their breadth of generality, which had similarities with the Cattell-Horn Gf-Gc model. Carroll used more than a single test battery to distinguish the structure of intelligence; he thought that different instruments drawn from different orientations offer a better opportunity to get the adequate structure. The cross-battery confirmatory factor analysis (CB-CFA) of such data provides stronger evidence of the structure of intelligence by explicitly testing models drawn from one test or theory against each other. Some researchers (Daniel, 1997; McGrew, 1997) have therefore proposed the use of a single term, the Cattell-Horn-Carroll (CHC) theory of intelligence. McGrew (1997; 2009) synthesized these two theories due to the need for a single taxonomy to cover broad and narrow abilities, which would enable the classification of narrow abilities measured using an individually administered intelligence test. CHC theory currently consists of nine broad abilities, Gf (fluid reasoning), Gc (comprehension-knowledge), Gv (visual processing), Ga (auditory processing), Gsm (short-term memory), Glr (long-term storage and retrieval), Gs (cognitive processing speed), Gq (quantitative knowledge), and Grw (reading and writing) (McGrew, 2004; 2009). The CHC theory could serve to bridge the gap between the development and use of intelligence tests also their interpretation. Recently, the revised Wechsler scales have become more consistent with CHC theory, although they are not explicitly based on it (Keith & Reynolds, 2010).

The most recently developed intelligence theory is Johnson and Bouchard's g-Verbal Perceptual Image Rotation Model (g-VPR; Johnson & Bouchard, 2005), which is used to compare the major psychometric models of human intelligence, specifically the Cattell-Horn fluid-crystallized model, Vernon's verbal-perceptual model, and Carroll's three-strata model (Johnson & Bouchard, 2005). The g-VPR model is based on a 42-test battery assessing cognitive ability. The first factor, Verbal, is defined as fluency and the ability to perform verbal tasks, such as comparing and contrasting patterns of words and solving problems expressed verbally. The second factor, Perceptual, is defined as the ability to reason, either by induction or deduction, with numbers or mathematical relationships, operations, and algorithms. The third factor, Image Rotation, is defined as the ability to determine the shape of an object after either a direct or reverse rotation (mirror image). The g-VPR model supports the existence of a general intelligence factor contributing to all aspects of intelligence, but also distinguishes between verbal and perceptual abilities, as well as considering visualization processes involving mental image rotation tasks as an independent contributor to the manifestation of human intelligence (Hunt, 2011; Johnson & Bouchard, 2005).

## ASSESSING INTELLIGENCE

Ideally, the development of theories about human cognitive abilities would be followed by improvements in the tools used to measure them. Raven developed the Progressive Matrices test and the Mill Hill Vocabulary Scale based on Spearman's g theory (Raven, 2008); however, the development of psychological testing is not always based on theoretical considerations. At the beginning of the 20th century, the Binet-Simon intelligence scale was developed to enable the French government to identify children who did not seem to profit from a regular classroom environment and who might instead benefit from being placed in a special education program (Mackintosh, 2011). The Binet-Simon intelligence scale designed to measure native ability to estimate the child's capacity for adapting him or herself to the social environment, also evaluating their judgment (Ayres, 1911). The scale included tests developed by others as well as ones developed by Binet and his collaborators (Boake, 2002). The Binet-Simon scale was widely used in Europe and North America and became a model and source of content for later intelligence tests (Boake, 2002). Two major revisions of the Binet-Simon scale were made by American psychologists, one by Robert Yerkes and James Bridges of the Boston Psychopathic Hospital and the other by Lewis Terman, Stanford University, before the United States entered the First World War (Boake, 2002). Terman's revision, which is known as the Stanford-Binet Intelligence Scale, quickly dominated American intelligence testing.

The First World War increased the demand for large-group intelligence testing because only a relatively few tests were available to evaluate whether candidates were suited for military recruitment and service. Two main Army intelligence tests were developed; Group Examinations Alpha and Beta. The Alpha test was designed to assess literate English speakers and the Beta test was for those who were illiterate or not proficient in English. These Army group examinations, particularly the Alpha verbal subtests and the Beta performance subtests, were a primary source of the subtests and items used in the WBIS, as explicitly stated by Wechsler (1939). Three of the Army Alpha subtests correspond to the WBIS; the Arithmetical Problems Test corresponds to Arithmetic in WBIS, the Alpha Practical Judgement Test corresponds to WBIS Comprehension, and the Alpha Information Test corresponds to Information in WBIS. The Army Beta subtests also become a source of WBIS tests; the Beta Digit Symbol Test corresponds to Digit Symbol in WBIS, the Picture Completion Test corresponds to WBIS Picture Completion, and the Picture Arrangement Test corresponds to Picture Arrangement in WBIS.

Before the development of the Alpha and Beta tests, the army relied on tests such as the Stanford-Binet scale, which heavily emphasized verbal ability (Kaplan & Saccuzzo, 2018). Based on his experience as a psychological examiner in war time, Wechsler attributed the misdiagnoses that resulted from the use of this test to its emphasis on verbal skills acquired through formal education (Boake, 2002). Wechsler worked at an Army camp where he scored Alpha examination protocols, gaining insights that led to him to combine verbal and nonverbal tests to overcome the shortcomings of existing intelligence tests, developing the first Wechsler scale of intelligence, WBIS.

## WECHSLER'S SCALES OF INTELLIGENCE

Wechsler's experience in supervising the tests of patients in mental hygiene clinics and psychiatric wards of Bellevue Hospital reinforced the need of an alternative to the Binet tests, particularly the need for different intelligence scales for adults (Wechsler, 1981). Wechsler adapted his first intelligence scale, WBIS, to assess the intelligence of children in the Wechsler Intelligence Scale for Children (WISC; 1949). The latest version of this test is the Wechsler Intelligence Scale for Children – Fifth Edition (WISC-V; 2014). WBIS has also been adapted into a variety of other tests. Those available on the Pearson Clinical website are the Wechsler Test of Adult Reading (WTAR; 2001), the Wechsler Nonverbal Scale of Ability (WNV; 2006), the Wechsler Individual Achievement Test – Third Edition (WIAT-III; 2009), and the Wechsler Preschool and Primary Scale of Intelligence – Fourth Edition (WPPSI-IV; 2012).

Wechsler (1939) defined intelligence as:

*"The aggregate or global capacity of the individual to act purposefully, to think rationally, and to deal effectively with his [or her] environment. It is global because it characterizes the individual's behavior as a whole; it is an aggregate because it is composed of elements or abilities which, though not entirely independent, are qualitatively differentiable" (Wechsler, 1939: 3).*

Based on his clinical expertise, Wechsler also highlighted several cognitive aspects of intelligence: verbal comprehension, abstract reasoning, perceptual organization, quantitative reasoning, memory, and processing speed. He developed his scale from the available resources and combined the cognitive tests in such way that it met the requirements of an adult scale of intelligence (Wechsler, 1939). Therefore, the factor structure of Wechsler's scales are always studied in every revision (see for example Atkinson, Cyr, Doxey, & Vigna, 1989; Benson, Hulac, & Kranzler, 2010; Cohen, 1952; Van der Heijden, Van den Bos, Mol, and Kessels, 2013; Weiss, Keith, Zhu, & Chen, 2013), as is the scale's reliability (see for example Carvajal, Schrader, & Holmes, 1996; Gignac, & Watkins, 2013; Griffith, & Yamahiro, 1958; Ryan & Schnakenberg-Ott, 2003).

## THE WECHSLER ADULT INTELLIGENCE SCALE

The first intelligence scale developed by Wechsler in 1939 was the WBIS. Some of his subtests were derived from portions of the 1937 revision of the Stanford-Binet test (Comprehension, Arithmetic, Digit Span, Similarity, and Vocabulary), while the rest were adapted from the Army Group Examinations (Picture Arrangement), Koh's Block Design (Block Design), the Army Alpha test (Information, Comprehension), the Army Beta test (Digit Symbol – Coding), Healy's Picture Completion (Picture Completion), and the Pinther-Paterson Test (Object Assembly). These tests were combined as subtests in Wechsler's initial battery of tests (Groth-Marnat, 2009).

Over the years, the WBIS has been revised four times, yielding the WAIS (1950), WAIS-R (1981), WAIS-III (1997), and WAIS-IV (2008) assessment scales. Revisions were usually made when a need is identified for changes to major item's content or for the addition of subtests to measure other specific abilities; however, the revision from WAIS (1950) to WAIS-R (1981) was relatively minor, with around 87% of the original WAIS items being retained. For WAIS-III (1997), 36% of the items were newly developed, with the most changes being made to the Performance subtests. In the WAIS-IV, the revision integrated advances in the field of intellectual assessment and updated the norms to reflect changes in the population over time (Wechsler, 2008b). Significant changes were made as a result of research advances in neuropsychology, cognitive neuroscience, and contemporary intelligence theory, as well as the increasing sophistication of psychological measurement (Weiss, Saklofske, Coalson, & Raiford, 2010). One of the major changes that occurred as Wechsler's scale was developed was the elimination of the verbal and performance Intelligent Quotient (IQ). Moreover, the WAIS-IV includes a measure of Full Scale IQ (FSIQ) and four indexes: Verbal Comprehension Index (VCI), Perceptual Reasoning Index (PRI), Working Memory Index (WMI), and Processing Speed Index (PSI). The importance of fluid intelligence was emphasized by developing new subtests such as the Visual Puzzle and Figure Weights. Further, the concepts of working memory and processing speed were integrated and improved (Groth-Marnat, 2009; Wechsler, 2008b).

The WAIS-IV can measure different areas of intellectual abilities, and can therefore be used to identify intellectual disability, intellectual giftedness, and describe cognitive functioning (Wechsler, 2008a). These distinctions can serve as a guide for diagnosis, treatment, and the development of specific programs or strategies for education. WAIS-IV can also be used as part of a neuropsychological assessment in medical settings (Fitrikasari, Jaeri, & Bintoro, 2013). Moreover, Wechsler's scales of intelligence have been adapted for use in various countries, including China (Lynn & Dai, 1993), Spain (Melendez, 1994), Canada (Lange, 2007), South Africa (Grieve & Van Eeden, 2010), France (Lecerf, Golay, & Reverte, 2012), the Netherlands (Wechsler, 2012), Finland (Roivainen, 2013), and Japan (Murayama et al., 2013). No adaptations of Wechsler's intelligence scales have been made for use in Indonesia however, so the newest revision, the WAIS-IV, was adapted for use in the work presented in this thesis.



## THE STRUCTURE OF THE INDONESIAN WECHSLER ADULT INTELLIGENCE SCALE – FOURTH EDITION

In this research project, the WAIS-IV, which was initially developed in the USA (WAIS-IV-US), was adapted for an Indonesian population. Adaptation is a way of minimizing bias and maximizing the cultural appropriateness of an instrument (Malda, Van de Vijver, Srinivasan, Transler, Sukumar, & Rao, 2008). The adaptation process carefully considered the linguistic and cultural differences of the Indonesian and US populations (Reckase, 1989). The scale was first translated from English to Indonesian, then re-translated back into English, and the meanings of each item were compared. The face validity of the scale was tested with Indonesian people and items were revised as needed. The content validity was discussed with clinical and educational psychologists. This process led to the development of the first Indonesian version of the WAIS-IV (WAIS-IV-ID).

The WAIS-IV is an individually administered measures of cognitive ability for individuals aged 16:0 to 90:11. It consists of 15 subtests, ten of which are core subtests (Block Design (BD), Similarity (SI), Digit Span (DS), Matrix Reasoning (MR), Vocabulary (VC), Arithmetic (AR), Symbol Search (SS), Visual Puzzle (VP), Information (IN), Coding (CD)), with five supplemental subtests (Letter-Number Sequencing (LN), Figure Weights (FW), Comprehension (CO), Cancellation (CA), and Picture Completion (PC)) (Wechsler, 2008). The WAIS-IV provides a measurement of general intellectual functioning (FSIQ) and four index scores: VCI, PRI, WMI, and PSI (Wechsler, 2008b). The FSIQ score is based on ten core subtests, the VCI and PRI scores are each based on three core subtests, and the WMI and PSI scores are each based on two core subtests.

## VERBAL COMPREHENSION INDEX (VCI)

VCI measures the extent to which an individual understands the meaning of words and possesses factual knowledge related to the verbal material, as well as their ability to conceptualize verbal information and adequately express the test's answers in words (Groth-Marnat, 2009). A higher score in this index means that the individual can work with abstract semantic information and possesses the ability to communicate fluently verbally. Higher scoring individuals can understand spoken language, construct sentences easily, and are interested in educational activities; therefore, the VCI score is affected by educational background.

VCI consists of four subtests; SI, VC, IN, and CO. SI measures logical abstract inductive reasoning and verbal fluency. VC was designed to assess an individual's word knowledge, concept formation, and language development. The IN subtest determines a person's ability to acquire, retain, and retrieve general factual knowledge. CO is a supplemental subtest, intended to measure the examinee's knowledge of conventional standards of behavior and their ability to evaluate the experience, as well as demonstrating their practical knowledge, social maturity, abstract thinking, and generalization (Groth-Marnat, 2009; Wechsler, 2008a, 2008b).

### PERCEPTUAL REASONING INDEX (PRI)

PRI measures perceptual abilities and is less affected by educational background than VCI. PRI measures an individual's ability to integrate perceptual stimuli with the appropriate motor responses, attention to detail, and the evaluation of visual-spatial information.

PRI consists of five subtests; BD, MR, VP, FW, and PC. BD measures nonverbal concept formation and the ability to analyze and synthesize abstract visual stimuli. MR was designed to measure visual-spatial reasoning, abstract reasoning, the simultaneous processing of visual-spatial information, and visual organization. VP measures visual-spatial reasoning, visual recognition, and the identification of parts in relation to the whole. FW is a supplemental subtest, which was designed to measure nonverbal mathematical reasoning, analogical reasoning, and the capacity for sustained effort when visually organizing test material. Another supplemental subtest, PC, measures visual alertness, recognition, and the identification of essential details in objects (Groth-Marnat, 2009; Wechsler, 2008a, 2008b).

### WORKING MEMORY INDEX (WMI)

WMI measures mental flexibility, the ability to hold and manipulate information in short-term memory, concentration, and attention. WMI is a complex and controversial construct compared with the other indexes. Due to the inclusion of the term "memory" in its title, WMI tends to be interpreted as a measure of memory; however, it is actually a narrow measure of the ability to hold and manipulate information for short periods (Groth-Marnat, 2009).

WMI consists of three subtests; DS, AR, and LN. DS consists of three tasks; DS Forward, DS Backward, and DS Sequencing. The shift from one DS task to another requires cognitive flexibility and mental alertness. AR involves mental manipulation, concentration, attention, short- and long-term memory, numerical reasoning ability, and mental alertness. LN is a supplemental subtest, which is only administered to examinees aged 16:0–69:11. LN involves sequential processing, mental manipulation, attention, concentration, memory span, and short-term auditory memory. The LN subtest consists of 10 items, each containing three trials (Groth-Marnat, 2009; Wechsler, 2008b).

### PROCESSING SPEED INDEX (PSI)

PSI requires mental and motor speed to enable a person to solve nonverbal problems. It measures a person's ability to plan, organize, and develop relevant strategies. Speed and concentration are critical factors supporting an individual's performance in PSI, as well as being assessed in WMI (Groth-Marnat, 2009).

PSI consists of three subtests; SS, CD, and CA. SS assesses short-term visual memory, visual motor coordination, cognitive flexibility, visual discrimination, psychomotor speed, the rate of mental operation, attention, and concentration. CD assesses similar areas to SS, including psychomotor speed, the ability to follow directions, accuracy, short-term visual memory, learning ability, visual perception, visual motor coordination, visual scanning ability, and cognitive flexibility. CA is a supplemental subtest only administered to examinees aged 16:0–69:11, and measures perceptual recognition, discrimination, and scanning ability. CA also measures visual neglect, response inhibition, and motor perseveration (Groth-Marnat, 2009; Wechsler, 2008b).

## THESIS OUTLINE

This thesis focuses on the development and psychometric evaluation of the Indonesian version of the WAIS-IV (WAIS-IV-ID). Its adaptation, reliability, and validation are described.

In **Chapter 2**, the first phase of adapting the WAIS-IV for use in Indonesia is described, including its translation, item analysis, and the analysis of subtest reliability in a limited sample. The rearrangement of item sequences in the WAIS-IV-ID was found to be necessary. In **Chapter 3**, the structural validity of WAIS-IV-ID is examined. A confirmatory factor analysis was used to assess the validity of WAIS-IV-ID structure, which was identical to that of the WAIS-IV-US. In **Chapter 4**, the reliability of the WAIS-IV-ID was assessed using test-retest and internal consistency reliability methods. In addition, the inter-scorer reliability for all verbal subtests is reported for a non-clinical sample. In **Chapter 5**, thirteen possible short forms of the WAIS-IV-ID were investigated, based on several previous studies and our research. It was hypothesized that short forms containing no more than four or five subtests might have the highest predictive value, classification accuracy, and reliability coefficients. In **Chapter 6**, the reliability and clinical utility of the WAIS-IV-ID were investigated in three clinical groups: patients with Alzheimer's disease, schizophrenia, and intellectual disability. Using a receiver-operating characteristic (ROC) analysis, the diagnostic validity was tested, and the optimal cut-off for all index and subtest scores of the WAIS-IV-ID was identified. Despite its strong internal structure, the need for further evaluation of the validity of WAIS-IV-ID using external criteria is described in **Chapter 7**. The external validity of the WAIS-IV-ID was investigated using other intelligence tests and the educational achievement of a university grade point average (GPA) as external validation criteria. Finally, the main results of the reported studies are summarized and discussed in **Chapter 8**, in which the future directions of the application of WAIS-IV-ID in Indonesia are also described.

# 2.

## The Development of the Indonesian Wechsler Adult Intelligence Scale–Fourth Edition (WAIS–IV)

**Published as**

Suwartono, C., Halim, M. S., Hidajat, L. L., Hendriks, M. P. H., and 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

## ABSTRACT

Through the years, several translated versions of Wechsler's intelligence test have been used in Indonesia, in clinical, educational or industrial settings. However, instruments such as Wechsler- Bellevue Intelligence Scale are outdated, have not been validated and lack proper normative data, resulting in measurement errors and invalid decisions made on the intellectual potential of individuals. The primary aim of this study was to adapt and validate the Wechsler Adult Intelligence Test - Fourth edition (WAIS-IV) for use in Indonesia. We describe the first phase in the adaptation of the WAIS-IV in the Indonesian language, including translation, item analysis, and reliability of the subtests. The sample of this research consisted of 148 healthy participants who are representative for the Indonesian population with respect to gender, age groups (ages 16 to 83), educational levels, and ethnic background. Results showed that the sequence of the US WAIS-IV could not be applied in Indonesia due to differences in index difficulties. Cronbach's coefficient alphas for the WAIS-IV subtests ranged from .74 - .92. For the subtests from the Verbal Comprehension Index, the inter-rater agreement ranged between .91 - .97. In all, the adaptation of the WAIS-IV for Indonesia is psychometrically promising.



**KEYWORDS**

Assessment, intelligence, item analysis, reliability

Assessment of intelligence is one of the most important topics in psychological testing. Worldwide, Wechsler's intelligence tests are best known and the most widely used for testing intelligence. The original version of the Wechsler intelligence test, the Wechsler-Bellevue Intelligence Scale (WBIS) for adolescents and adults was developed in 1939. Internationally, the WBIS has been extended and modified several times to WAIS, WAIS-R, WAIS-III, and the recently published WAIS-IV (Wechsler, 2008a, 2008b). Adaptations from the original US versions of these tests have been published in many countries, such as the Netherlands (see Van der Heijden, Van den Bos, Mol, & Kessels, 2013), Japan (see Murayama, Iseki, Tagaya, Ota, Kanasuki, Fujishiro, Arai, & Sato, 2013), Finland (see Roivainen, 2013), France (see Lecerf, Golay, & Reverte, 2012), South Africa (see Grieve & Van Eeden, 2010), Canada (see Lange, 2007), Spain (see Melendez, 1994), and China (see Lynn & Dai, 1993). In the latest WAIS-IV items and subtests have been substantially revised, new subtests have been added and norms have been updated to take cohort effects into account (Wechsler, 2008b). The WAIS-IV updates also incorporated theoretical advances in neuropsychology, cognitive neuroscience, and contemporary intelligence theory, as well as increasing sophistication in psychological measurement (Weiss, Saklofske, Coalson, & Rai- ford, 2010).

In Indonesia, the WBIS is still being used for intelligence testing. However, details about its translation (presumably done in the 1960s or 1970s), standardization, psychometric properties and the development of the written manual are unknown, and this version is probably unauthorized (Lembaga Pengembangan Sarana Pengukuran dan Pendidikan Psikologi Fakultas Psikologi Universitas Indonesia, n.d.) and certainly outdated. Despite these shortcomings, the WBIS is still being taught and the most widely administered intelligence test in Indonesia. A translated version of the later Wechsler Adult Intelligence Scale is also available (Seksi Psikodiagnostik, 1992), but suffers from the same shortcomings, is also probably unauthorized and has never gained full acceptance in the Indonesian psychology community. As an example, in Bandung, several faculties of psychology preferred the WBIS over the WAIS as the translated WAIS items were culturally biased (Polhaupessy, n.d). The lack of valid intelligence tests results in measurement error, misleading decisions made about the potentials of individuals (i.e., in vocational training or job selection), as well as a clinical misdiagnosis. This stresses the urgent need for the development and adaptation of a valid intelligence test for the Indonesian population. The WAIS-IV can be used to identify the intellectual disability, intellectual giftedness, and cognitive functioning in examinees (Wechsler, 2008a). The WAIS-IV also can be used as part of the neuropsychological assessment in clinical settings (Fitrikasari, Jaeri, Bintoro, 2013). Therefore, the focus of this study is the development of the Indonesian version of the WAIS-IV and to psychometrically evaluate the Indonesian WAIS-IV.

The WAIS-IV is an individually administered test battery for individuals aged 16 to 90. It consists of 15 subtests, namely Block Design (BD), Similarity (SI), Digit Span (DS), Matrix Reasoning (MR), Vocabulary (VC), Arithmetic (AR), Symbol Search (SS), Visual Puzzle (VP), Information (IN), Coding (CD), Letter Number Sequencing (LN), Figure Weights (FW), Comprehension (CO), Cancellation (CA), and Picture Completion (PC). The WAIS-IV subtests are grouped into the core and supplemental subtests. The first ten subtests are the core subtests and the last five (i.e. LN, FW, CO, CA, and PC) are the supplemental subtests (Wechsler, 2008). The total battery provides an assessment of general intellectual functioning (FSIQ) and four index scores. The four index scales include Verbal Comprehension, Perceptual Reasoning, Working Memory, and Processing Speed (Wechsler, 2008b). The Verbal Comprehension scale includes three core subtests (SI, VC and IN) and one supplemental subtest (CO). The Perceptual Reasoning scale includes three core subtests (BD, MR, and VP) and two supplemental subtests (FW and PC). The Working Memory scale includes two core subtests (DS and AR) and one supplemental subtest (LN). The Processing Speed scale includes two core subtests (SS and CD) and one supplemental subtest (CA).

## METHODS

### PARTICIPANTS

The Indonesian population size was estimated at 237.6 million in 2010. The majority of people live on the island of Java, and the country's six biggest islands are Kalimantan, Maluku-Papua, Sumatera, Sulawesi, Java, and Bali-Nusa Tenggara (Badan Pusat Statistik Indonesia, 2012). The official language is Indonesian, a standardized register of Malay, an Austronesian language (Lewis, Simons, & Fennig, 2013). Indonesia is a multi-ethnic society, with more than 1000 ethnic/sub-ethnic groups. However, the size of most ethnic groups is small, and only 15 groups consist of more than one million people (Suryadinata, Arifin, & Ananta, 2003). Formal education and national media are in the Indonesian language (Lewis, Simons, & Fennig, 2013). In addition to being fluent in the Indonesian language, many Indonesians are also fluent in another regional language, such as Javanese, Sundanese, or others. These regional languages are commonly used at home or in the local community.

The participants for this study were 176 Indonesian individuals aged 16 - 75 years old. Of these, 148 participants completed all 15 subtests. Half of the participants (53.4%) were men and 46.6% were women. Age ranged from 16.2 to 83.9 years ($M = 37.34$, $SD = 16.75$). 17.6% of the participants had less than eight years of education, 12.8% completed junior high school, 33.8% senior high school, 7.4% completed academy, 20.3% had an under- graduate degree, 6.8% a master degree and 1.4% a doctoral degree. With respect to ethnicity, participants in our study belonged to the ethics groups of Tionghoa (37.2%), Java (25%), and Sunda (14.9%). For validity purposes we also recruited participants originally from Ambon, Batak, Betawi, Lampung, Manado, Minang, Nusa Tenggara, Palembang, Papua, Serang, Toraja, and other islands, who lived on Java in the regions mentioned above. Most of the participants were employed (64.86%) in various sectors from education, sanitation, security, transportation, and business sectors.

**SAMPLING METHOD**

Recruitment of participants was performed in accordance with the WAIS-IV Technical and Interpretive Manual (Wechsler, 2008b), taking the Indonesian life expectancy into account. Based on 1980-2010 Population Census (Badan Pusat Statistik Indonesia, 2012), the life expectancy for Indonesian was 70.7 years. We used a quota sampling method. A stratified sampling plan ensured that the normative sample included representative proportions of individuals according to selected demographic variables. The allocation of samples will be based on the following variables: age, sex, education level, and geographic region. In this study, the geographical area covered only Java island, in and around cities of Jakarta, Tangerang, Bogor, and Bandung.

The involvement of the participants in this research is based on several considerations, as stated in the technical manual of the US WAIS-IV (2008b). Participants were excluded when their primary language was not Indonesian, when they primarily used nonverbal language or were uncommunicative, when they were unable to understand instructions and participate fully in testing, when they were inadequate to comply with testing to ensure a valid assessment, when they were already tested with any intelligence test in the past six months, when they had an uncorrected visual impairment, uncorrected hearing loss. Upper-extremity disability that would affect motor performance, currently admitted to psychiatric facility, currently taking medication that might impact cognitive test performance (e.g., antidepressants, antipsychotics, etc), recent and significant change related to cognitive status change, currently receiving chemotherapy in the past two months, and previously diagnosed with any physical condition or illness that might affect to lower the test performance, were excluded from participation.

**INSTRUMENTS**

The original US version of the WAIS-IV was translated into Indonesian by two translators, with bilingual (English-Indonesian) ability. Both translators were educational psychologists and lecturers in the Faculty of Psychology of Atma Jaya Catholic University of Indonesia. They performed the translation in the same period of time. The first two authors (CS, MSH) compared their translations and met with the translators to review the translations, identified differences in Indonesian-English meaning, and adapted the Indonesian-language version to achieve the most accurate culturally equivalent meaning (International Test Commission, 2010).

For the back-translation process, we asked two new independent translators who had the bilingual ability and a background in psychology. The first translator was a clinical psychologist and lecturer in the Faculty of Psychology of Atma Jaya Catholic University of Indonesia. The second translator was an organizational psychologist who works as a consultant. Both independently translated the Indonesian WAIS-IV back into US English. Subsequently, the authors and the translators met to review the back translations, identified differences in English-Indonesian meaning and adapted the Indonesian-language version to achieve the most accurate culturally equivalent meaning. This final translated WAIS-IV was authorized by Pearson Assessment.

The scoring of the WAIS-IV followed the discontinue rules that have been incorporated into the scale to avoid frustration by the examinee because it includes too difficult items. Subtest items were ordered according to increasing difficulty. The discontinue rules were applied after the examinee had several numbers of consecutive scores of zero. The discontinue rules are different in each subtest (Wechsler, 2008b). In agreement with Wechsler (2008a), the administration of BD was discontinued when the examinee obtained two consecutive response score of zero. The administration of SI, MR, VC, AR, VP, IN, FW, CO was discontinued when the examinee obtained three consecutive response score of zero. The administration of DS was discontinued when the examinee's response on both trials of one item was zero. The administration of LN was discontinued after the examinee obtained a score of zero on all trials. The administration of PC was discontinued after the examinee obtained four consecutive zero responses. The SS and CD were discontinued after 120 seconds. For CA, each item discontinued after 45 seconds.

**PROCEDURE**

Testing took place in a counseling room at the university, at a company, school, or another place which met the requirements of a psychological testing environment. That is, no external distracting, quiet or soundproof, without interruptions during test administration.

We contacted potential participants and gave information about this study. If a potential participant agreed to participate, we set an appointment to administer the test. Before the test was administered, all participants provided written informed consents. The WAIS-IV was administered in its prescribed order, and demographic data were collected. All items in all subtests were administered, and discontinue rules were applied during scoring.

**ANALYSES**

Index difficulty was computed for each item in 12 subtests of the WAIS-IV. SS, CD, and CA were excluded because these are timed tests, for which the index difficulty analyses cannot be performed. In the timed test category, items are so easy that all participants will have all items correct if they have enough time to finish them (Anastasi & Urbina, 1997). As, index difficulty can be different for participants from specific age groups, the next analysis explored whether differences existed between two age groups. The first group consisted of people aged 16 to 69 years and the other of people over 70. This division was based on the administration guidelines of the WAIS-IV subtests, as all subtests are only administered to people from 16 - 69 and only 12 subtests administered to people over 70. We performed this analysis with the Wilcoxon Signed Rank Test.

The next step was to reorder the item sequence, apply the discontinue rules, and calculate the total score for each subtest. We compared these to the original sequence's raw score if we applied the discontinue rules using a dependent sample t-test. Also, we calculated Cohen's d (Cohen, 1988). This analysis determined whether the new sequence was the best solution for a subtest.

Reliability coefficients were obtained using Cronbach's coefficient alpha. The alpha coefficient takes the variance of both the total subtest score and item scores into account and provides the reliability that is the average of all possible split-half reliabilities (Anastasi & Urbina, 1997). Cronbach's coefficient alpha cannot be determined for the subtests of the Processing Speed index (SS, CD, and CA). For the Verbal Comprehension subtests (SI, VC, and CO), we calculated the inter-rater reliability. These three subtests have a different scoring system (Wechsler, 2008a), the criteria for which were also translated into Indonesian. We asked three independent raters to score the participant's verbal responses based on the criteria in the Indonesian manual. Two of three raters were educational psychologists and one a psychometrician. None of the raters had any previous experience with the WAIS-IV scoring rules. Reliability coefficients were obtained using intraclass correlation coefficients (two-way mixed model with absolute agreement).



## RESULTS

### ITEM ANALYSIS

For most subtests of the Indonesian translated version of the WAIS-IV, the item index difficulties did not follow the US sequences from easy to difficult (Table 1). Most changes in item sequences had to be made for the subtests in the Verbal Comprehension and Perceptual Reasoning scale (except for BD). Based on this analysis, we re-sequenced the item order of those subtests. Table 1 shows the range and means of index difficulties across subtests and age categories.

**Table 1: WAIS-IV subtest index difficulties.**

| Subtest | Range of p for all ages | Mean p for all ages | Mean p for age 16 - 69 | Mean p for age over 70 | Z over age 70 and age 16 - 69 |
|---|---|---|---|---|---|
| BD | .16 – 1 | .74 | .75 | .70 | −1.14 |
| SI | .05 - .92 | .53 | .53 | .54 | −0.08 |
| DS | .02 – 1 | .55 | .55 | .47 | −4.43** |
| MR | .12 - .99 | .57 | .58 | .46 | −3.39 |
| VC | .12 - 1 | .58 | .58 | .58 | −0.08 |
| AR | .14 - .99 | .56 | .56 | .56 | −0.26 |
| VP | .11 - 1 | .53 | .53 | .45 | −2.98** |
| IN | .02 - .99 | .33 | .33 | .33 | −0.24 |
| LN | .04 - .99 | .62 | .62 | n/a | n/a |
| FW | .12 - .97 | .50 | .50 | n/a | n/a |
| CO | .00 - .82 | .34 | .34 | .37 | −1.05 |
| PC | .02 - .96 | .37 | .38 | .28 | −3.30** |

**Wilcoxon signed rank¬ test (Z test) is significant at the .01 level (2-tailed).

Subsequently, we analyzed the differences in index difficulties between the two age groups. Using the Wilcoxon Signed Rank Test, Z-test, we found that there were significant results on four subtests (DS: Z = −4.43, $p < .01$; PM: Z = −3.39, $p < .01$; VP: Z = −2.98, $p < .01$; PC: Z = −3.30,
$p < .01$). All these subtests were more difficult for participants over age 70 years.

After rearranging the item order, we scored the performance using the discontinue rules (Table 2). Significant differences in performances between the original and rearranged sequences were found for VC ($t_{(147)} = -5.01$, $p < .01$), VP ($t_{(147)} = -3.60$, $p < .01$), IN ($t_{(147)} = -6.50$, $p < .01$), and CO ($t_{(147)} = -12.81$, $p < .01$). Of these, CO had the largest effect size. No significant differences were found for BD, SI, MR, AR, FW, and PC.

## RELIABILITY ANALYSIS

Table 3 shows the reliability coefficients of each WAIS-IV subtests. Cronbach's coefficient alpha of the WAIS-IV subtest ranged from .74 - .92, indicating an acceptable (BD, SI, and CO) to excellent (LN and FW) reliability. For the subtests of the Verbal Comprehension index, we also computed inter-rater reliabilities. The inter-rater agreements were high for SI (r = .97), VC (r = .96), and CO (r = .91).

## DEMOGRAPHIC DATA ANALYSIS

We found no significant differences in the scores of men and women in any of the subtests. Significant differences between participants of different educational backgrounds were found. Table 4 shows the differences between participants that had completed senior high school (N = 50) and those who had a university undergraduate degree (N = 30).

Significant differences were found on most subtests, except CO ($t_{(78)}$ = .22, $p > .05$) and CA ($t_{(78)}$ = 1.47, $p > .05$). The group that had an undergraduate degree scored higher on all subtests, in agreement with previous research (Grieve & Van Eeden, 2010; Matarazzo & Herman, 1984).

Table 5 shows the Pearson product-moment correlation coefficients between age and all subtests in WAIS-IV. We found significant correlations for all subtests of the Processing Speed index, most of the subtests of the Perceptual Reasoning index, and most of the subtest of the Working Memory index. No significant correlations were found between age and the subtests of the Verbal Comprehension index.



For the subtests of the Processing Speed index, the highest correlation was between age and SS ($r_{(146)}$ = −.40, $p < .01$). These results were consistent with numerous studies about the aging-related decline in the speed of information processing (see Verhaeghen & Salthouse, 1997). FW was the only subtest from the Perceptual Reasoning index that did not significantly correlated with age. The highest correlation was found between age and VP, $r_{(146)}$ = −.25, $p < .01$. These results were consistent with studies about age and reasoning (see Verhaeghen & Salthouse, 1997).

**Table 2: The differences between original and rearrange the sequence of subtests' items, scored with discontinuing rules (*df* = 147).**

| Subtest | Mean score of | | t-value | Sig. (2 tailed) | Cohen's d |
|---|---|---|---|---|---|
| | Original | Rearranged | | | |
| BD - BD_Reorder | 34.35 | 34.40 | −1.42 | .16 | 0.12 |
| SI - SI_Reorder | 18.57 | 18.61 | −0.29 | .77 | 0.02 |
| MR - MR_Reorder | 13.13 | 13.38 | −1.67 | .10 | 0.14 |
| VC - VC_Reorder | 30.20 | 33.09 | −5.01 | .00 | 0.41 |
| AR - AR_Reorder | 11.90 | 11.92 | −0.24 | .81 | 0.02 |
| VP - VP_Reorder | 11.66 | 12.24 | −3.60 | .00 | 0.30 |
| IN - IN_Reorder | 7.02 | 7.91 | −6.50 | .00 | 0.53 |
| FW - FW_Reorder | 11.74 | 11.86 | −0.98 | .33 | 0.08 |
| CO - CO_Reorder | 4.88 | 11.84 | −12.81 | .00 | 1.05 |
| PC - PC_Reorder | 8.22 | 8.32 | −1.22 | .22 | 0.10 |

**Table 3: The subtests' reliability coefficients.**

| Name of the subtest | Alpha | SEM | N items | Mean | SD |
|---|---|---|---|---|---|
| Block Design (BD) | .77 | 5.80 | 14 | 35.29 | 12.12 |
| Similarity (SI) | .79 | 2.56 | 18 | 19.19 | 5.53 |
| Digit Span (DS) | .89 | 2.19 | 48 | 26.16 | 6.67 |
| Matrix Reasoning (MR) | .87 | 1.96 | 26 | 14.93 | 5.40 |
| Vocabulary (VC) | .87 | 3.56 | 30 | 34.99 | 9.84 |
| Arithmetic (AR) | .86 | 1.58 | 22 | 12.34 | 4.30 |
| Symbol Search (SS) | n/a | n/a | n/a | 28.14 | 1.37 |
| Visual Puzzle (VP) | .80 | 1.91 | 26 | 13.69 | 4.24 |
| Information (IN) | .88 | 1.59 | 26 | 8.43 | 4.68 |
| Coding (CD) | n/a | n/a | n/a | 59.80 | 23.15 |
| Letter Number Sequencing (LN) | .92 | 1.66 | 30 | 17.43 | 6.04 |
| Figure Weights (FW) | .90 | 1.97 | 27 | 13.48 | 6.17 |
| Comprehension (CO) | .74 | 2.61 | 18 | 12.39 | 5.16 |
| Cancellation (CA) | n/a | n/a | n/a | 3.67 | 12.15 |
| Picture Completion (PC) | .88 | 1.74 | 24 | 8.95 | 4.92 |

For the Working Memory index, significant correlations were found between DS ($r_{(146)}$ = -.30, p < .01) and LN ($r_{(137)}$ = -.39, p < .01). These results are consistent with studies about age and working memory (see Verhaeghen & Salthouse, 1997).

## DISCUSSION

In this study, we described the translation process of the WAIS-IV into the Indonesian language. We performed psychometric analysis, including the order of the item sequences, item analyses, and reliability of each subtest of the WAIS-IV-ID.

As expected, the item sequences of most of the translated WAIS-IV subtests had to undergo major changes. In all subtests from the Verbal Comprehension index, we had to reorder the item sequence, although stayed close to the original items for content purposes. For example, in the subtest Information of the original US test, most of the items that refer to science, geography, world history, world figures, and literature are related to Western culture knowledge. Indonesian participants were more likely to answer correctly in science and geography, but most of them had difficulties in answering items about specific historical persons (like Sacagawea) and literature (Alice in the Wonderland). The index difficulties ranged from .41 - .02. On those items, participants either gave up or said they remembered that these figures had appeared in a movie. In one item about historical figures in the US, their answers were not fully correct although they gave correct information to some extent. Another example is the VC subtest, which also had to undergo major reordering, because the items may have a different meaning in daily conversations compared to their formal meaning in Indonesian. For instance, item number 4 ("bed", in Indonesian "ranjang") of VC, became item 11 after reordering.



**Table 4: Subtests differences between groups with different education backgrounds (*df* = 78).**

| Subtest | M for high school | SD for high school | M for university | SD for university | t-test | Sig (two-tailed) | Cohen's d |
|---|---|---|---|---|---|---|---|
| BD | 34.46 | 11.50 | 44.20 | 11.12 | 3.71 | .00 | 3.71 |
| SI | 19.42 | 4.41 | 22.57 | 3.77 | 3.26 | .00 | 3.26 |
| DS | 26.30 | 6.41 | 30.27 | 5.63 | 2.80 | .01 | 2.80 |
| MR | 14.72 | 5.21 | 18.80 | 4.25 | 3.62 | .00 | 3.62 |
| VC | 34.12 | 8.14 | 41.57 | 7.00 | 4.17 | .00 | 4.17 |
| AR | 12.14 | 4.66 | 15.07 | 3.47 | 2.98 | .00 | 2.98 |
| SS | 28.98 | 9.79 | 34.30 | 10.45 | 2.29 | .02 | 2.29 |
| VP | 13.76 | 3.55 | 17.07 | 4.06 | 3.82 | .00 | 3.82 |
| IN | 8.00 | 3.61 | 11.73 | 4.16 | 4.23 | .00 | 4.23 |
| CD | 61.18 | 19.81 | 76.33 | 17.39 | 3.46 | .00 | 3.46 |
| LN | 17.60 | 5.83 | 20.47 | 5.04 | 2.24 | .03 | 2.24 |
| FW | 13.88 | 6.20 | 17.00 | 5.61 | 2.26 | .03 | 2.26 |
| CO | 13.68 | 4.43 | 13.90 | 4.16 | 0.22 | .83 | 0.22 |
| CA | 31.76 | 11.40 | 35.90 | 13.53 | 1.47 | .15 | 1.47 |
| PC | 8.70 | 4.74 | 12.63 | 4.31 | 3.71 | .00 | 3.71 |

**Table 5: The subtests' correlation with participant's age.**

| | BD | SI | DS | MR | VC | AR | SS | VP | IN | CD | LN | FW | CO | CA | PC |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Age | -.24** | .08 | -.30** | -.18* | .06 | -.02 | -.40** | -.25** | -.01 | -.39** | -.39** | -.16 | .05 | -.36** | -.22** |
| $r^2$ | .06 | .01 | .09 | .03 | .00 | .00 | .16 | .06 | .00 | .16 | .15 | .03 | .00 | .13 | .05 |
| Sig. | .00 | .31 | .00 | .03 | .48 | .83 | .00 | .00 | .88 | .00 | .00 | .06 | .53 | .00 | .01 |

Note: *Correlation is significant at the .05 level (2-tailed); **Correlation is significant at the .01 level (*2-tailed*).

Specific syllables in some words in Indonesia, like the "ke" can be "erased" when spoken. In item number four, the syllable "ke" is not present. The participants who gave incorrect answers in this item, tended to listen to the word but not read it. Even if the examiner had pointed to the written word, many participants still not changed their answer. As a result, many zero responses were scored for this item and its index difficulty was .76. Another example is item number 17 (i.e. "plagiarism", in Indonesian "plagiat"), after reordering, it became item 30. The subtest Arithmetic (AR) from Working Memory index and all subtests from Perceptual Reasoning index, except for Block Design, also needed major reordering of the items sequence. Many of the Picture Completion items resulted in incorrect responses from the participants, because they did not indicate the important missing detail of the picture, but indicated other absent objects in the picture instead. For instance, for the item in which a picture with trees is shown, participants answered that the river lacked fish, or that people or vehicles were missing in the picture. Other incorrect responses were found in several pictures with which Indonesian people are unfamiliar. For example, items such as snow or a stove are less common in Indonesia (many participants responded that the stove is a kind of washing machine). Reordering the items and applying the discontinue rules improved the participants' scores.

Some other recommendations can be made based on our experiences in the data collection process. First, in the Similarity subtest, the example item (what is the similarity between "two and seven") was difficult to answer for some participants. They answered in a concrete way, e.g., "both have edges". If participants do not understand the instruction, then continuation is a problem. To overcome this, we repeated the instruction twice and explained the correct answer of the example item. In case they still did not understand the purpose of the subtest, the examiner gave an example using two concrete objects, e.g., what is the similarity of two pieces of fruit (mango and banana), that are more part of the participants' daily lives. We used this procedure especially in participants with a lower educational background and for senior citizens. Secondly, in the Comprehension subtest, we found no significant differences between the performances of participants who completed high school and university graduates. This result was not consistent with other studies (Grieve & Van Eeden, 2010 , Matarazzo & Herman, 1984). By checking the responses of those participants, we found that many of the answers were very short. From the data collection evaluation, the participants tended to say, "Ok, that is it" or "I do not know anything else" or "It is enough". Their responses usually only covered one general concept, for which they did not obtain the maximum two points. For future research, recommend to rephrase the item appropriately to obtain another response as suggest in the Wechsler (2008a).

We found age effects that are in agreement with previous research. For instance, Verhaeghen and Salthouse (1997) performed a meta-analyses on 50 studies about aging and speed of processing, in which they reported correlations between −.23 and −.68. In our study, we found correlations between −.36 and −.40. Verhaeghen and Salthouse (1997) also performed a meta-analysis on 34 studies in which the relation between age and the Working Memory index was studied. They found that correlations varied between .03 and −.48. In our study, we found a range between −.02 until −.39. These negative correlation coefficients are consistent with previous studies and indicate that the number of items that could be remembered immediately decreases with age and affects the total score. Overall, most of the correlations between WAIS-IV subtests with age showed significant and negative correlations. Verhaeghen and Salthouse (1997) suggest that almost all aspects and type of information processing are affected by age, causing a broad decline in many facets of cognitive functioning when people get older. They concluded that there are two types of general factors that explain the cognition differences in the adult. The first type refers to basic and relatively pervasive loss in processing speed. The second type refers to the ability to preserve information in a temporary working memory store while processing is carried out. These two types are not independent because the proportion of age-related variance in cognitive performance that is related to working memory capacity is also shared to a large extent with processing speed measures. Like in their study our results also show that the subtests from the Processing Speed index and most of Working Memory index are correlated with age. We did not find this correlation for AR from the Working Memory index because this subtest also depends on the more crystallized verbal and quantitative capacities.

Comparing the reliability coefficients from the Indonesian adaptation of the WAIS-IV to the US version (Wechsler, 2008b), all reliability coefficients are acceptable to good (although in general somewhat lower than the US coefficients). Moreover, the subtests in the Verbal Comprehension index have excellent inter-rater agreement reliability. A possible weakness is the probing used by the examiner. We tested participants from different educational and ethnic backgrounds. Some of the participants needed explanations in their regional language, such as Javanese, Sundanese, or Mandarin. Sometimes the participants gave mixed responses in regional language and Indonesian language, but the examiner sometimes was less familiar with the regional language. For further research, we recommend recruiting examiners from different ethnic backgrounds who understand the regional language. Moreover, recruitment from this study was limited to Jakarta and its surroundings. For further research, the geographical area has to be expanding to other areas of Indonesia. Also, a larger sample is recommended to replicate the findings of this study and to analyze the construct validity of the translated version.

Overall, the results of this first study provide a strong foundation for developing the Indonesian version of WAIS-IV further. It is psychometrically promising; the reliabilities of the subtests are good and have a high inter-rater agreement. The next step is to explore the factorial structure of the Indonesian WAIS-IV using a larger sample.

# 3.

## Structural Validity of the Indonesian Wechsler Adult Intelligence Scale – Fourth Edition (WAIS–IV–ID)

**Revised version submitted:**
Suwartono, C., Hendriks, M. P. H., Hidajat, L. L., Halim, M. S., and Kessels, R. P. C. Structural Validity of the Indonesian Wechsler Adult Intelligence Scale – Fourth Edition (WAIS-IV-ID). (2017).

## ABSTRACT

This study examines the structural validity of the Indonesian version of the Wechsler Adult Intelligence Scale – Fourth Edition (WAIS-IV-ID). Based on previous research using the US version, we tested 11 theoretical models with 1,401 Indonesian participants. We found that four- and five-factor models provided an adequate fit to the data, with the four-factor model performing slightly better. Verbal comprehension, perceptual reasoning, working memory, and processing speed emerged as first-order factors, resembling the original factor structure of the US version of the WAIS-IV. Our findings contribute to the comprehension and interpretation of adult intellectual functioning, as measured by the WAIS-IV. Future research is needed to examine the external validity and clinical utility of the WAIS-IV-ID.



**KEYWORDS**

Factor analysis, Indonesian WAIS-IV, intelligence, internal validity

From the original Wechsler-Bellevue Intelligence Scale to the third revision of the Wechsler Adult Intelligence Scale (WAIS-III), a structure consisting of verbal and performance factors was used. The four-factor structure which was found on WAIS-III (Garcia, Ruiz, & Abad, 2003; Ward, Ryan, & Axelrod, 2000), formed a basis of different index scores.

**Table 1: Model specifications.**

| Model | Factor | Subtests | Note |
|---|---|---|---|
| 1 | All 15 subtests on a general factor | SI, VC, IN, CO*, DS, AR, LN*, BD, MR, VP, FW*, PC*, SS, CD, CA* | Single-factor model (Canivez & Watkins, 2010). |
| 2 | Verbal | SI, VC, IN, CO*, DS, AR, LN* | Two-factor model: verbal vs. performance IQ (Boake, 2002; Wechsler, 2008b). |
| | Performance | BD, MR, VP, FW*, PC*, SS, CD, CA* | |
| 3 | Verbal | SI, VC, IN, CO* | Represents the three-factor model (Wechsler, 2008b). |
| | Reasoning | BD, MR, VP, FW*, PC* | |
| | Combined WM and PS | DS, AR, LN*, SS, CD, CA* | |
| 4 | VC | SI, VC, IN, CO* | Four-factor model (Bowden et al., 2011a, 2011b; Garcia et al., 2003; Ward et al., 2000; Wechsler, 2008b). |
| | PR | BD, MR, VP, FW*, PC* | |
| | WM | DS, AR, LN* | |
| | PR | SS, CD, CA* | |
| 5 | as Model 4 | with a joint loading of AR on PR as well as WM. | Models 5–7 are variations of model 4 (Benson et al., 2010; Golay et al., 2011; Wechsler, 2008b; Weiss et al., 2013). |
| 6 | as Model 4 | with a joint loading of AR on VC as well as WM. | |
| 7 | as Model 4 | with a joint loading of AR on VC and PR as well as WM. | |
| 8 | Gc | SI, VC, IN, CO* | Five-factor model of WAIS-IV based on CHC broad abilities (see Benson et al., 2010; Niileksela et al., 2013; Ward et al., 2012). |
| | Gs | SS, CD, CA* | |
| | Gv | BD, VP, PC* | |
| | Gf | MR, FW*, AR | |
| | Gsm | DS, LN* | |
| 9 | as Model 8 | with AR loading on Gc and Gf. | Models 9–11 are variations of model 8 (Benson et al., 2010; Ward et al., 2012; Weiss et al., 2013). |
| 10 | as Model 8 | with AR loading on Gsm and Gf. | |
| 11 | as Model 8 | with AR loading on Gc, Gsm, and Gf. | |

* Supplemental subtest.

Chapter 3

This four-factor structure was also confirmed for the WAIS-IV (Bowden, Saklofske, & Weiss, 2011a, 2011b; Golay, Reverte, & Lecerf, 2011; Wechsler, 2008a, 2008b; Weiss, Saklofske, Coalson, & Raiford, 2010). The four index scores are Verbal Comprehension (VC), Perceptual Reasoning (PR), Working Memory (WM), and Processing Speed (PS).

Recently, a synthesis of the Carroll three-stratum theory and the Cattell-Horn-Carroll *Gf-Gc* theory (CHC theory; see McGrew, 2009) was proposed as a framework to be used when making hypotheses formulating various aspects of human cognitive abilities. From the perspective of CHC theory, the VC, WM, and PS factors can be argued to measure crystallized ability (*Gc*), short-term memory (*Gsm*), and processing speed (*Gs*), respectively; however, within the theoretical CHC framework, the PR factor may be separated into *Gv* (visual-spatial processing) and *Gf* (fluid reasoning).

The results of Weiss, Keith, Zhu, and Chen's (2013) study suggested a five-factor solution, which is the integration of CHC theory with data from the WAIS-IV. Other studies have also demonstrated that five-factor solutions, consisting of *Gf, Gv, Gc, Gsm*, and *Gs,* are optimum for the WAIS-IV (see Benson, Hulac, & Kranzler, 2010; Niileksela, Reynolds, & Kaufman, 2013; Ward, Bergman, & Hebert, 2012). The PRI (one of the four indexes of the WAIS-IV) was split into *Gf* (MR and FW) and *Gv* (BD, VP, and PC). In this study, we examined 11 models utilizing one to five factors (with g as the highest factor) to investigate the structural validity of an authorized Indonesian-language version of WAIS-IV (WAIS-IV-ID; Suwartono, Halim, Hidajat, Hendriks, & Kessels, 2014; see Table 1).

## METHODS

### PARTICIPANTS

Participants were recruited from an Indonesian population using the standardized protocol described in the WAIS-IV manual (Wechsler, 2008b). A quota sampling technique was used to attempt to represent the populations of the six major islands in Indonesia. The quota was based on the percentage of the Indonesian population inhabiting each island (BPS, 2012).

A total of 1,401 participants completed the WAIS-IV-ID, including 613 men (43.75%) and 788 women (56.25%). Participants ranged in age from 16.20 to 69.90 years (*M* = 32.56; *SD* = 14.60). Their educational attainment was also varied: 6.50% completed junior high school, 48.11% completed senior high school, 38.69% obtained an undergraduate degree (BA or BSc), and 6.63% completed the post-graduate study (MA, MSc, or Ph.D.). The participants were from Java (52.97%), Sumatra (20.70%), Sulawesi (13.06%), Borneo (6.35%), Bali (3.71%), and Nusa Tenggara (3.21%).

### INSTRUMENTS

The WAIS-IV-ID consists of ten core subtests: Block Design (BD), Similarities (SI), Digit Span (DS), Matrix Reasoning (MR), Vocabulary (VC), Arithmetic (AR), Symbol Search (SS), Visual Puzzle (VP), Information (IN), and Coding (CD). It also contains five supplemental subtests: Letter-Number Sequencing (LN), Figure Weights (FW), Comprehension (CO), Cancellation (CA), and Picture Completion (PC). The items on the WAIS-IV-ID subtests are identical or equivalent to those on the US version of the WAIS-IV (WAIS-IV-US) (Suwartono et al., 2014; Wechsler, 2008a). The WAIS-IV-ID was administered individually, following the WAIS-IV guidelines (Wechsler, 2008a).

### DATA ANALYSES

The standardized subtest scores were analyzed using LISREL 8.80 (Jöreskog & Sörbom, 2006). Several goodness-of-fit indexes were computed to determine the acceptable model fit (Akaike, 1987; Bentler, 1995; Hu & Bentler, 1999; Jöreskog & Sörbom, 1996; Kline, 2005). The Root Mean Square Error of Approximation (RMSEA), the Comparative Fit Index (CFI), and the Standardized Root Mean Residual (SRMR) were the primary indexes used to evaluate the model fit, in addition to Akaike's Information Criterion (AIC), the Tucker-Lewis Index (TLI), and the Adjusted Goodness-of-Fit Index (AGFI).

## RESULTS AND DISCUSSION

The four- and five-factor models (Table 2) met all the criteria for an adequate fit, represented by lower AIC values and higher AGFI scores (Akaike, 1987; Kline, 2005). The four-factor structure model (model 5) had the best fit of all models. PR had the highest contribution of *g* loading (.83), then WM (.78), PS (.74), and VC (.69). Of the five-factor structure models, model 10 fit the data best. *Gf* had the highest *g* loading (.99), then *Gv* (.91), *Gsm* (.70), *Gs* (.66), and *Gc* (.65).

**Table 2: Goodness-of-fit indexes for the individual models.**

| Model | $\chi^2$ | df | $\chi^2$/df | RMSEA | CFI | AIC | SRMR | TLI | AGFI |
|---|---|---|---|---|---|---|---|---|---|
| 1 | 3218.28 | 90 | 35.76 | .16 | .86 | 3278.28 | .10 | .83 | .69 |
| 2 | 2881.42 | 88 | 32.74 | .15 | .87 | 2945.42 | .10 | .85 | .71 |
| 3 | 1643.34 | 87 | 18.89 | .11 | .90 | 1709.34 | .09 | .88 | .81 |
| 4 | 1320.91 | 86 | 15.36 | .10 | .92 | 1388.91 | .09 | .90 | .84 |
| 5 | 1061.12 | 85 | 12.48 | .09 | .93 | 1131.12 | .09 | .91 | .87 |
| 6 | 1326.38 | 85 | 15.60 | .10 | .92 | 1396.38 | .09 | .90 | .84 |
| 7 | 1064.23 | 84 | 12.67 | .09 | .93 | 1136.23 | .09 | .91 | .87 |
| 8 | 1147.93 | 85 | 13.51 | .10 | .92 | 1217.93 | .09 | .91 | .86 |
| 9 | 1149.27 | 84 | 13.68 | .10 | .93 | 1221.27 | .09 | .91 | .86 |
| 10 | 1108.37 | 84 | 13.19 | .09 | .93 | 1180.37 | .09 | .91 | .86 |
| 11 | 1111.12 | 83 | 13.39 | .09 | .93 | 1185.12 | .09 | .91 | .86 |

Note: All p-values for $\chi^2$ were significant at the .01 level.

In model 5 (see Figure 1), AR was found to correlate with the measures of WM (.29) and PR (.52). The highest loading for a subtest for each index was CO for VC (.77), MR for PR (.78), DS for WM (.82), and CD for PS (.82).

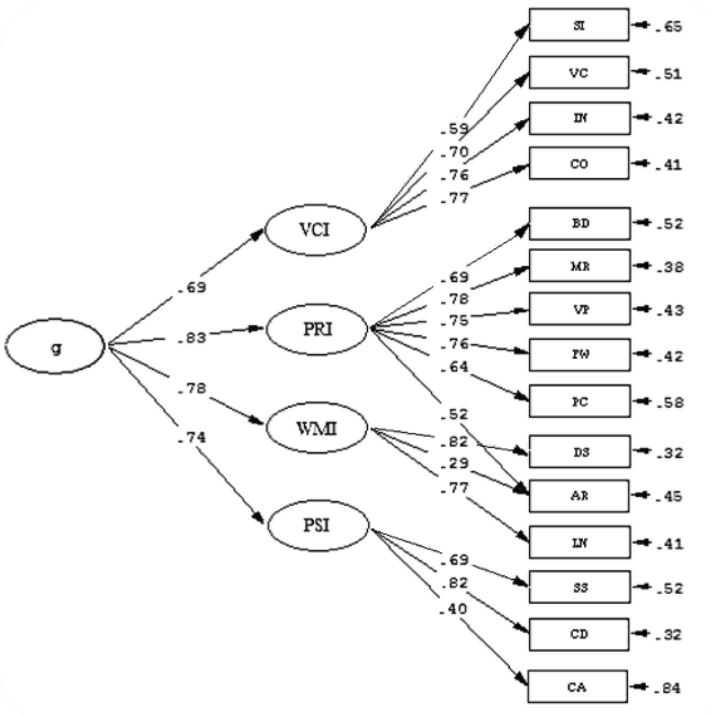


**Figure 1: The best-fit four-factor model (Model 5).**

In our Indonesian sample, we found a slight advantage for the four-factor models over the five-factor models, consistent with the factor solutions reported by Bowden et al. (2011a, 2011b), Golay et al. (2011), and Wechsler (2008a, 2008b); however, our data still also support the use of a five-factor model (model 10; Figure 2), in line with the findings of Benson et al. (2010), with fluid reasoning (*Gf*) included as the fifth factor.
In model 10 (Figure 2), AR was found to correlate with the measures of *Gsm* (.21) and *Gf* (.59), despite the long-debated use of this subtest in the Wechsler scales (Georgas, Weiss, van de Vijver, & Saklofske, 2003).

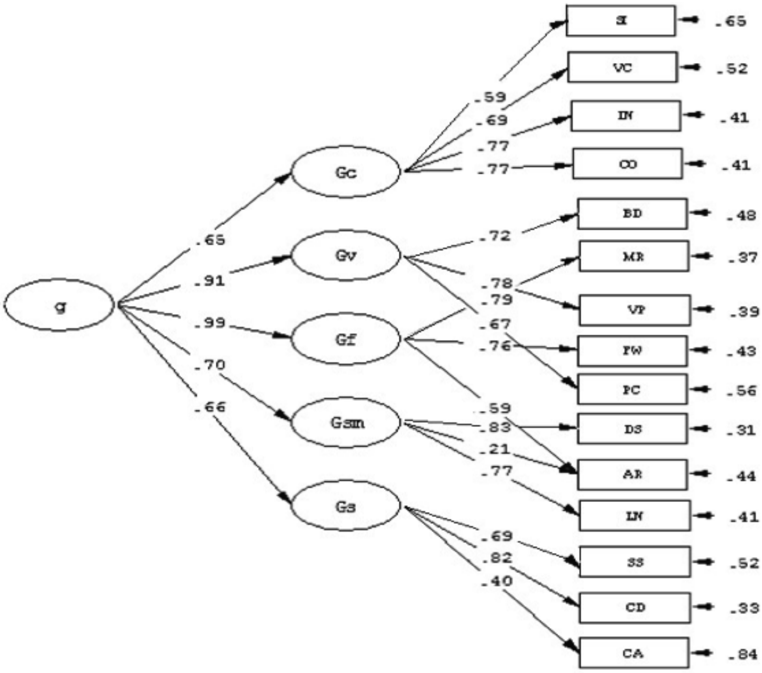


**Figure 2: The best-fit five-factor model (Model 10).**

AR assesses cognitive processes that overlap with several other factors, such as the ability to remember inputs for computation (WM or *Gsm*) and knowledge about how to solve problems systematically (PR or *Gf*). Our results did not support the findings of Weiss et al. (2013) as AR was weak to nonexistent VC or *Gc* factor loading for all four- and five-factor models.

The sample in the present study did not include individuals over 70 years old, and the inclusion of these older individuals in future research is therefore required to determine whether these models fit the data for that age category. Moreover, the confirmatory factor analysis does not capture all aspects of a test's score validities (Canivez, Konold, Collins, & Wilson, 2009), suggesting that additional validation should be performed using both external criteria and clinical populations.

# 4.

# Reliability Analysis of the Indonesian Wechsler Adult Intelligence Scale – Fourth Edition (WAIS–IV–ID) in a Non–Clinical Sample

**In preparation:**
Suwartono, C., Hendriks, M. P. H., Halim, M. S., and Kessels, R. P. C. Reliability Analysis of the Indonesian Wechsler Adult Intelligence Scale – Fourth Edition (WAIS-IV-ID) in a Non-Clinical Sample.

## ABSTRACT

Previous factor analysis studies have examined the structural validity of the Indonesian version of the Wechsler Adult Intelligence Scale – Fourth Edition (WAIS-IV-ID). The reliability of test scores is one of the most important psychometric properties that can be evaluated. Here, we report that the WAIS-IV-ID has internal consistency, inter-scorer reliability, and test-retest reliability. This reliability was assessed by administering the test to a standardized sample of 1,500 Indonesian participants, 77 of whom were included in a test-retest study. We estimated the internal consistency by calculating the coefficient α for all subtests of the WAIS-IV-ID and used coefficient ω to analyze the internal consistency reliability of the FSIQ and index scores. We analyzed the inter-scorer reliability for all verbal subtests, using three examiners, which was found to be excellent. The FSIQ score was found to reliable based on the coefficients α and ω, as well as the test-retest reliability analysis. The reliability coefficients were higher for FSIQ than for the index scores. The α estimates and test-retest reliability coefficients for the indexes were higher than the ω estimates. It was concluded that the WAIS-IV-ID is reliable based on internal consistency methods, inter-scorer reliability, and a test-retest method.



**KEYWORDS**

Internal consistency, non-clinical sample, reliability, test-retest, WAIS-IV, Indonesia

Psychological testing of cognitive constructs, especially intellectual functioning, is an important practice issue in Psychology. Decisions about employee hire, referral for special education, and neurological or psychiatric diagnoses are often based on cognitive measurements such as intelligence tests (Kessels & Hendriks, 2016). One of the most widely used intelligence scales are the Wechsler Scales (Van Scoyoc, 2017).

The latest Wechsler Scale for adults, the Wechsler Adult Intelligence Scale – Fourth Edition (WAIS-IV) has been adapted into the Indonesian language (WAIS-IV-ID) (Suwartono, Halim, Hidajat, Hendriks, & Kessels, 2014). Since information on the reliability of this assessment is limited, we previously examined the structural validity of WAIS-IV-ID (Suwartono, Hendriks, Hidajat, Halim, & Kessels, submitted). A test score on a scale such as WAIS-IV-ID is only meaningful if it reliably and accurately reflects a psychological construct; therefore, evidence should be collected to support the validity of this assessment (Cook & Beckman, 2006). Wechsler (2008b) reported the inter-scorer reliability, internal consistency, and test-retest stability of the WAIS-IV-US. Here, we investigate the reliability of the WAIS-IV-ID (FSIQ, indexes, and subtests) by assessing its inter-scorer agreement, internal consistency, and test-retest reliability.

Verbal subtests are a target of inter-scorer reliability studies, which check whether the scoring is objective and consistent across scorers based on the criteria provided in the Indonesian verbal test scoring manual (Hallgren, 2012). Test-retest reliability is also used to determine the stability of one's score across multiple assessments using the same test (Cook & Beckman, 2006; Miller, 1995).

According to the Classical Test Theory (CTT), test score reliability is defined as a variance ratio equal to the true variance of the test scores divided by the total variance of the test scores (Crocker & Algina, 2008). The internal consistency usually obtained using the coefficient alpha (α) methods; coefficient α assumes that the test is entirely homogeneous, where all components of the test are a single common true-score factor and all unique variance arises from measurement error (Miller, 1995). The WAIS-IV-ID has a multidimensional structure, and the coefficient α may combine multiple sources of true score variance (Gignac, 2014). Although using coefficient α is a common method for estimating the internal consistency reliability of test scores, an alternative approach for estimating this reliability is a model-based internal consistency reliability score (Miller, 1995; Gignac & Watkins, 2013), including the and omega (ω) (McDonald, 1985), OmegaH (ωh) (Zinbarg, Revelle, Yovel, & Li, 2005), and OmegaS ($\omega_s$) (Reise, Bonifay, & Haviland, 2012) scores. Coefficient ω represents the internal consistency reliability associated with the total score, independent of the additional factor variance or true score variance (Gignac, 2014). Coefficient $\omega_h$ represents the unique internal consistency reliability associated with total scale composite scores (Zinbarg et al., 2005), while $\omega_s$ represents the unique internal consistency reliability associated with the subscale scores (Reise et al., 2012).

This study focuses on the internal consistency reliability estimates of the WAIS-IV-ID, as represented by the coefficients α, $\omega_h$, and $\omega_s$. The coefficient ω values were hypothesized to be lower than the α values due to the multidimensionality of g. The inter-scorer reliability was predicted to be high due to the concise and unambiguous scoring guidelines, while the indexes were predicted to be stable over a period of time, as assessed using test-retesting.

## METHODS

### PARTICIPANTS

To study the internal consistency of the WAIS-IV-ID, 1,500 participants were recruited as a sample population. The participants were divided into four age categories (16–19, 20–34, 35–54, and 55–69 years old; Table 1), following the method of Gignac and Watkins (2013). Most participants were from Java (58.80%), Sumatra (18.10%), Sulawesi (11.90%), Borneo (5.30%), Nusa Tenggara (3.00%), and Bali (2.90%). The participants also belonged to the following ethnic groups: Javanese (23.50%), Tionghoa (20.90%), Batak (6.10%), Minahasa (2.90%), Balinese (2.80%), Minang (2.60%), Kei (1.00%), and Other (40.20%).

**Table 1: Demographic data for internal consistency reliability analysis.**

| | | Age | | Sex (% of participants) | | Highest education level attained (% of participants) | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Age (years) | N | M | SD | Male | Female | Junior high school | Senior high school | Under-graduate degree | Masters degree | Doctoral degree |
| 16–69 | 1500 | 31.43 | 14.08 | 44.90 | 55.10 | 7.70 | 55.80 | 29.10 | 6.90 | 0.50 |
| 16–19 | 227 | 18.19 | 0.93 | 30.30 | 69.70 | 27.30 | 72.20 | 0.40 | - | - |
| 20–34 | 808 | 24.43 | 3.83 | 48.10 | 51.90 | 1.50 | 60.00 | 31.10 | 7.40 | - |
| 35–54 | 303 | 44.06 | 5.78 | 50.50 | 49.50 | 6.90 | 37.60 | 44.20 | 10.20 | 1.00 |
| 55–69 | 162 | 61.29 | 4.13 | 38.30 | 61.70 | 13.00 | 45.70 | 30.90 | 7.40 | 3.10 |

To study the test-retest reliability, a total of 77 young adults (35.10% males, 16–33 years old ($M$ = 21.68, $SD$ = 2.4)) were recruited. Most of them had completed senior high school (89.60%), the rest were undergraduates (9.10%), or had finished junior high school (1.30%). They all came from Java (Jakarta and Surabaya).

### INSTRUMENTS

The WAIS-IV-ID consists of 15 subtests; Block Design (BD), Similarity (SI), Digit Span (DS), Matrix Reasoning (MR), Vocabulary (VC), Arithmetic (AR), Symbol Search (SS), Visual Puzzle (VP), Information (IN), Coding (CD), Letter Number Sequencing (LN), Figure Weights (FW), Comprehension (CO), Cancellation (CA), and Picture Completion (PC). The first ten subtests are considered core subtests, while the latter five are supplemental subtests (Wechsler, 2008a). The WAIS-IV-ID provides a measure of general intellectual functioning, in terms of a FSIQ and four index scores: Verbal Comprehension Index (VCI), Perceptual Reasoning Index (PRI), Working Memory Index (WMI), and Processing Speed Index (PSI).

## PROCEDURE

The WAIS-IV-ID is an individual intelligence test. It was administered according to the administrative rules indicated in the WAIS-IV manual (Wechsler, 2008a). Institution or participants were contacted and given an overview of this study. If they agreed to participate, they signed an informed consent form and were scheduled to take the test. The participants were administered the WAIS-IV-ID on two separate occasions, 10 to 65 days apart ($M$ = 26.66, $SD$ = 13.54).

The inter-scorer reliability was assessed using the results of 256 participants for the four subtests of the VCI. The comparison involved assessments performed by three well-trained scorers, a clinical psychologist, an educational psychologist, and a psychometrician. Each scorer scored the participants' answers independently, based on the criteria outlined in the WAIS-IV-ID scoring guidelines. One scorer only scored the participant answers for SI, VC, and CO, while the others scored the answers for all subtests of verbal comprehension. None of the scorers had any previous experience with the WAIS-IV scoring rules.



## ANALYSES

For internal consistency reliability, coefficients $\alpha$, $\omega_h$, and $\omega_s$ were calculated. The coefficient $\alpha$ values were calculated for the subtests with IBM SPSS 23. The coefficient $\alpha$ values for the indexes were calculated as functions of the reliabilities of the components and their intercorrelations using the following formula (Nunnally & Bernstein, 1994, p.269):

$$r_{yy} = 1 - \frac{k - \Sigma rxx}{\bar{R}y}$$

where $k$ represents the number of components, $rxx$ is the reliabilities of the components, and $\bar{R}y$ is the sum of elements in the correlation matrix of the components (including the unities in the diagonal). The $\omega_h$ and $\omega_s$ scores were calculated as described by Gignac (2014), using IBM SPSS Amos 22.0.0. A phantom composite variable and its corresponding latent variable were created, and the implied correlation between them was squared to get the internal consistency reliability associated with the composite scores. For the comparison between the internal consistency reliability estimates, a difference of .06 or greater between coefficients $\alpha$ and $\omega$ was considered significant (Gignac, Bates, & Jang, 2007).

Three independent scorers were asked to score the participants' verbal responses based on the criteria outlined in the Indonesian WAIS-IV Indonesian verbal test scoring manual. Reliability coefficients were obtained using intra-class correlation coefficients (ICC), calculated using a two-way mixed model with an absolute agreement method (McGraw & Wong, 1996; Shrout & Fleiss, 1979). An ICC close to 1 indicates a high similarity between the scores determined by each of the raters.

The means and standard deviations of each subtest and index were calculated, and a Pearson product-moment correlation was used to perform the reliability analysis. The 95% confidence intervals (CI) of the test-retest reliability coefficients were also calculated. The standard differences were calculated using the mean score differences between the tests taken by the same participant on two different occasions divided by the pooled standard deviation (Cohen, 1988). The relationship between the test-retest interval and the stability of the WAIS-IV subtests and indexes was determined by correlating the time intervals between testing with the absolute score differences between the test and retest examination scores.

## RESULTS

**Table 2: Internal consistency reliabilities of the WAIS-IV-ID subtests, estimated with coefficient alpha (α).**

| Scale | 16–69 year olds (N = 1500) | 16–19 year olds (n = 227) | 20–34 year olds (n = 808) | 35–54 year olds (n = 303) | 55–69 year olds (n = 162) |
|---|---|---|---|---|---|
| BD | .75 | .72 | .75 | .76 | .72 |
| MR | .89 | .86 | .88 | .89 | .85 |
| VP | .84 | .81 | .84 | .85 | .83 |
| FW | .88 | .88 | .89 | .86 | .83 |
| PC | .82 | .78 | .80 | .82 | .72 |
| SI | .63 | .57 | .57 | .68 | .74 |
| VC | .79 | .70 | .80 | .81 | .73 |
| IN | .82 | .77 | .82 | .84 | .81 |
| CO | .73 | .70 | .70 | .76 | .75 |
| DS | .84 | .82 | .83 | .82 | .76 |
| AR | .83 | .82 | .83 | .85 | .79 |
| LN | .72 | .66 | .68 | .73 | .76 |

Table 2 displays the coefficient α values across all subtests and all five age groups. For the age categories 16–69, 35–54, and 55–69, the MR subtest had the highest coefficient α values (.89, .89, and .85, respectively). For the age categories 16–19 and 20–34, the FW scores had the highest coefficient α (.88 and .89, respectively). The SI subtest scores had the lowest coefficient α for the age categories 16–69, 16–19, 20–34, and 35–54 (.63, .57, .57, and .68, respectively). For the age category 55–69, the PC subtest had the lowest coefficient α (.72).

Table 3 displays the internal consistency coefficients α, $\omega_h$, and $\omega_s$ estimated across the WAIS-IV-ID scales and for all five age categories. For the age category 16–69, the coefficient α estimate for FSIQ was .97, and for the indexes, it ranged from .86 (PSI) to .94 (PRI). The $\omega_h$ estimate for FSIQ was .82, and the $\omega_s$ values ranged from .18 (PRI) to .37 (PSI). For the age category 16–19, the coefficient α estimate for FSIQ was .96, and for the indexes, it ranged from .85 (PSI) to .94 (PRI). The $\omega_h$ estimate for FSIQ was .79, and the $\omega_s$ for the index scores ranged from .22 (PRI) to .38 (PSI). For the age category 20–34, the coefficient α estimate for FSIQ was .96, and for the index scores, it ranged from .85 (PSI) to .94 (PRI). The $\omega_h$ estimate for FSIQ was .82, and the $\omega_s$ ranged from .18 (PRI) to .39 (PSI). For the age category 35–54, the coefficient α estimate for FSIQ was .97, and for the index scores, it ranged from .87 (PSI) to .95 (PRI). The $\omega_h$ estimate for FSIQ was .87, and the $\omega_s$ for the index scores ranged from .13 (PRI) to .32 (PSI). For the age category 55–69, the coefficient α estimate for FSIQ was .96, and for the index scores, it ranged from .86 (PSI) to .93 (PRI). Finally, the $\omega_h$ estimate for FSIQ was .77, and the $\omega_s$ for the index scores ranged from .17 (WMI) to .46 (PSI).



**Table 3: Comparison of internal consistency reliabilities of WAIS-IV-ID FSIQ and index scores, estimated by the coefficients α, $\omega_h$, and $\omega_s$.**

| Age categories (years) | N | α | | | | | $\omega_h$ | $\omega_s$ | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | FSIQ | PRI | VCI | WMI | PSI | FSIQ | PRI | VCI | WMI | PSI |
| 16–69 | 1500 | .97 | .94 | .90 | .90 | .86 | .82 | .18 | .28 | .21 | .37 |
| 16–19 | 227 | .96 | .94 | .86 | .89 | .85 | .79 | .22 | .27 | .31 | .38 |
| 20–34 | 808 | .96 | .94 | .88 | .89 | .85 | .82 | .18 | .29 | .22 | .39 |
| 35–54 | 303 | .97 | .95 | .91 | .90 | .87 | .87 | .13 | .23 | .15 | .32 |
| 55–69 | 162 | .96 | .93 | .90 | .87 | .86 | .77 | .22 | .39 | .17 | .46 |

* The PSI subtest reliability score came from the test-retest coefficient.

The reliability of the FSIQ score as determined using the coefficient α estimate ranged from .96 to .97. The corresponding coefficient $\omega_h$ estimate ranged from .77 to .87, with an average difference between the coefficient α and ω reliability estimates of .15, which may be considered significant (Gignac et al., 2007). The reliability of the VCI scores estimated using coefficient α was .86–.91, whereas the $\omega_s$ for all four index scores ranged from .23 to .39. The average difference between the coefficient α and ω reliability estimates was .60, which is considered significant (Gignac et al., 2007). The reliability of the PRI scores estimated with the coefficient α ranged between .93 and .95, whereas the $\omega_s$ for all four index scores ranged between .13 and .22. The average difference between the coefficient α and ω reliability estimates amounted to .75, which can be considered significant (Gignac et al., 2007). The reliability of the WMI scores estimated with the coefficient α ranged between .87 and .90, while the ωs for all four index scores ranged from .15 to 31. The average difference between the coefficient α and ω reliability estimates was .68, which is considered significant (Gignac et al., 2007). The reliability of the PSI scores estimated with the coefficient α ranged between .85 and 87, whereas the $\omega_s$ for all four index scores ranged from .32 to .46. The average difference between the coefficient α and ω reliability estimates amounted to .47, which may be considered significant (Gignac et al., 2007).

The inter-scorer reliability for the four subtests of verbal comprehension was high, ranging from .94 (CO) to .98 (IN) (Table 4).

**Table 4: The inter-scorer reliability coefficient.**

| Subtests | *n* scorer | Intraclass correlation | 95% CI | |
|---|---|---|---|---|
| | | | Lower bound | Upper bound |
| SI | 3 | .97 | .96 | .98 |
| VC | 3 | .95 | .93 | .96 |
| IN | 2 | .98 | .97 | .98 |
| CO | 3 | .94 | .93 | .96 |

The test-retest reliability coefficient of the FSIQ was .91 (see Table 5). The test-retest for the index scores PRI (.80) and PSI (.85) were good, while for VCI (.75) and WMI (.77) they were adequate. For the subtests, the test-retest correlation coefficients ranged from .50 (SI) to .86 (CD). Among the subtests, only BD, IN, DS, SS, and CD had good test-retest reliabilities ($r \geq .70$). All correlation coefficients were significant at the .01 level (two-tailed). We found no significant correlation between the time intervals and the absolute score differences between the test-retest coefficient values for all subtests and indexes of WAIS-IV-ID ($r_{(75)}$ = –0.14–0.24, $p > .01$).

The largest effect sizes revealed by the standard differences were for FSIQ, PRI, BD, PC, and SS, while moderate effect sizes were found for WMI, PSI, MR, FW, and AR, and small effect sizes were identified for VCI, VP, DS, LN, CD, and CA, as well as all verbal subtests (SI, VC, IN, and CO).

**Table 5: The test-retest reliability coefficients.**

| | Test | | Retest | | r | 95% CI | | Standard difference |
|---|---|---|---|---|---|---|---|---|
| | $M_1$ | $SD_1$ | $M_2$ | $SD_2$ | | Lower | Upper | |
| FSIQ | 106.01 | 8.86 | 112.16 | 10.03 | .91 | .86 | .94 | 1.48 |
| PRI | 102.35 | 9.93 | 109.00 | 10.34 | .80 | .71 | .87 | 1.05 |
| VCI | 101.32 | 5.63 | 102.87 | 6.60 | .75 | .63 | .83 | 0.35 |
| WMI | 107.30 | 11.31 | 112.06 | 12.70 | .77 | .66 | .85 | 0.58 |
| PSI | 111.69 | 17.03 | 119.21 | 17.44 | .85 | .77 | .90 | 0.78 |
| BD | 9.73 | 2.33 | 11.40 | 2.65 | .78 | .67 | .85 | 1.00 |
| MR | 10.87 | 2.23 | 11.87 | 1.73 | .52 | .33 | .66 | 0.50 |
| VP | 10.77 | 2.10 | 11.56 | 2.17 | .67 | .53 | .78 | 0.46 |
| FW | 11.94 | 2.29 | 13.03 | 1.90 | .55 | .37 | .69 | 0.54 |
| PC | 8.64 | 2.24 | 10.49 | 2.45 | .68 | .54 | .78 | 0.98 |
| SI | 9.13 | 1.25 | 9.58 | 1.45 | .50 | .31 | .65 | 0.33 |
| VC | 12.38 | 1.43 | 12.25 | 1.44 | .54 | .36 | .68 | 0.09 |
| IN | 9.40 | 1.87 | 9.90 | 2.12 | .83 | .75 | .89 | 0.42 |
| CO | 9.14 | 1.37 | 9.06 | 1.26 | .53 | .35 | .68 | 0.06 |
| DS | 11.66 | 2.75 | 12.26 | 2.90 | .70 | .57 | .80 | 0.27 |
| AR | 11.03 | 2.28 | 12.13 | 2.63 | .67 | .53 | .78 | 0.55 |
| LN | 12.04 | 3.56 | 13.10 | 3.85 | .66 | .52 | .77 | 0.35 |
| SS | 11.65 | 3.50 | 13.64 | 3.43 | .75 | .63 | .83 | 0.81 |
| CD | 12.66 | 3.49 | 13.42 | 3.53 | .86 | .78 | .91 | 0.40 |
| CA | 9.03 | 2.64 | 9.52 | 2.69 | .60 | .43 | .73 | 0.21 |

* The PSI subtest reliability score came from the test-retest coefficient.

## DISCUSSION

In the present study, we examined the internal consistency, inter-scorer reliability, and test-retest reliability of the Indonesian version of the WAIS-IV. Based on the coefficient α method, we concluded that the internal consistency of FSIQ and PRI revealed by the WAIS-IV-ID is quite similar to the previously reported consistency of the WAIS-IV-US, while the coefficient α of the index scores VCI and WMI were slightly lower than those of the WAIS-IV-US (Wechsler, 2008b). For the subtests, CD had the same reliability coefficient as the WAIS-IV-US; however, other subtests of the WAIS-IV-ID had lower, but still reliable, coefficient α values than the US version of the test (Wechsler, 2008b).

The model-based reliability, indicated by coefficient $\omega_h$, of the FSIQ scores was lower than the coefficient α estimates, but still sufficient to support the reliability of the result. Coefficient α was found to overestimate the internal consistency reliability of the

Chapter 4

FSIQ by an average of .15; however, the level of reliability for FSIQ indicated by $\omega_h$ was still sufficient for interpretation. This result is in line with the findings reported by Gignac and Watkins (2013) and Watkins (2017). Moreover, the results also indicate that FSIQ is a multidimensional construct.

In contrast to the FSIQ, the estimated internal consistencies associated with the indexes ($\omega_s$) were all very low across all age groups; for example, in the 20–34 age group, the coefficient $\omega_s$ scores for PRI, VCI, WMI, and PSI were estimated to be .18, .29, .22, and .39, respectively, while the coefficient α values were estimated at .94, .88, .89, and .85, respectively. These results were similar to those reported by Gignac and Watkins (2013) and Watkins (2017) and indicated that the indexes were more heavily influenced by general factors than group-specific factors. The coefficient α values of the index scores seem to be misleading because most of the exploratory power in each index score is due to general factors (Watkins, 2017). The lower ω coefficients indicate that the WAIS-IV-ID subtests do not define the latent variables very well, and consequentially, the construct replicability tends to be challenging. The indexes should therefore be considered questionable for a meaningful interpretation. For index scores with $\omega_s$ estimates less than 0.50, Gignac and Watkins (2013) argue that meaningful interpretations are impossible. These results also support our previous research about the structural validation of the WAIS-IV-ID, in which we argued that the FSIQ score should be considered the first step in the interpretation process (Suwartono et al., submitted). Moreover, when aggregating across multiple measures (subtests and indexes), errors will cancel each other out, leading to a better estimation of the true score. The principle of aggregation indicates that multiple measures of the same trait will yield a more stable and representative estimate than a single measure (Rushton, Brainerd, & Pressley, 1983), which explains why the test-retest and coefficient α of internal consistency reliability values for the subtests were lower than those of the FSIQ.

The strength of the current study was the use of ω coefficients as the internal consistency method, rather than relying on the α coefficients. Coefficient ω values are appropriate for a multidimensional test battery, especially one with a hierarchical structure such as the WAIS-IV-ID (Watkins, 2017). The use of the ω coefficient is limited by the lack of agreement on the optimal way to determine an adequate level of ω reliability, and the lack of consensus on the best computation of standard errors for the ω coefficients (Watkins, 2017).

The inter-scorer reliability scores were excellent, indicating a high degree of consistency between scorers in their interpretation of a participant's answers. All verbal subtests could be reliably scored by scorers with different backgrounds, using the scoring guideline provided in the verbal test scoring manual. The guidelines for scoring the verbal subtests of the WAIS-IV-ID can, therefore, be considered reliable and ready to be used in clinical practice.

Overall, the retest scores were better than those of the first test, although they were slightly decreased for the VC and CO subtests. This may due to the inclusion of open questions in the VC and CO subtests, for which the subtests require more elaborative answers. Participants usually did not elaborate on their answer, despite being probed by the examiner. Therefore, in the retest, for VC and CO, participants even gave shorter answers. We used a relatively short time interval between test administrations (generally around three weeks); therefore, participants might recall their first test, potentially resulting in the enhanced retest scores. This meant that the retest administration was not fully independent of the first assessment. Furthermore, some participants reported that they were curious about the correct responses, so after the first test they may have searched for the answers in books or other resources, such as the internet. Learning or practice effects might also have occurred, since the same test was used on both occasions, as has been suggested previously (Farahat, Rohlman, Storzbach, Ammerman, & Anger, 2003; Goldstein & Watson, 1989). Future research should use a longer time interval to minimize these learning or practice effects.

Moreover, reflecting on the process of WAIS-IV-ID administration, it might be useful to consider a short form of the WAIS-IV-ID, rather than asking participants to complete the full test twice. Many Indonesian people are not used to completing a long psychological assessment. The WAIS-IV-ID took about 1.5 to 2.0 hours; therefore, it is possible that a decline in motivation may have interfered in the testing of some participants, especially for the subtests at the end of the battery. Freund and Holling (2011) also stressed the importance of motivation. The inclusion of a new task considered attractive by the participants may increase their motivation.

In conclusion, we determined that the WAIS-IV-ID is reliable using the internal consistency, inter-scorer reliability, and test-retest reliability. The estimated reliability scores of the indexes determined using the test-retest and coefficient α internal consistency reliability methods were lower but still acceptably reliable for clinical use.

# 5.

# The Development of a Short Form of the Indonesian Version of the Wechsler Adult Intelligence Scale – Fourth Edition (WAIS–IV–ID)

**In preparation:**
Suwartono, C., Hendriks, M. P. H., Hidajat, L. L., Halim, M. S., and Kessels, R. P. C.
The Development of a Short Form of the Indonesian Version of the Wechsler Adult Intelligence Scale – Fourth Edition (WAIS-IV-ID).

## ABSTRACT

The Wechsler intelligence scales are very popular in clinical practice and for research purposes. However, because they are time consuming researchers and psychologists to explore the possibility of shorter test battery compositions. In this study, we investigated 13 potential short forms of the Indonesian version of the Wechsler Adult Intelligence Scale – Fourth Edition (WAIS-IV-ID). The data were collected from a standardized sample of 1,745 Indonesian participants. None of the estimated Full Scale IQ (FSIQ) values from the short forms (FSIQEst) were significantly different from the FSIQ scores estimated using the full WAIS-IV-ID, and a strong correlation was observed between each of these values. The classification accuracies of the short forms were between 56.79% and 81.03%. Based on the results presented here, we recommend the use of a WAIS-IV-ID short form combining four subtests: Matrix Reasoning, Information, Arithmetic, and Coding. The validity of this short form was demonstrated in a second study with a separate sample population; however, future research should investigate its validity with a larger sample size.



**KEYWORDS**

Classification accuracy, Indonesian WAIS-IV, intelligence, reliability, short forms, validity

The frequent and widespread use of the comprehensive and time-consuming Wechsler intelligence scales for psychological assessments highlight the need for psychologists to explore the possibility of developing shorter batteries of tests. The aim of generating these short forms is to reduce the time required for their administration while maintaining the estimated validity of their Full Scale IQ (FSIQ) scores. Attempts to develop these short forms have been conducted since the publication of the first version of Wechsler's intelligence test, the Wechsler-Bellevue Intelligence Scale (McNemar, 1950; Rabin, 1943), with attempts made to develop short forms of the subsequent tests; the Wechsler Adult Intelligence Scale (WAIS) (Doppelt, 1956; Maxwell, 1957), the Wechsler Adult Intelligence Scale – Revised (Cyr & Brooker, 1984; Kaufman, Ishikuma, & Kaufman-Packer, 1991; Silverstein, 1982; Ward, 1990), the Wechsler Adult Intelligence Scale – Third Edition (Donnell, Pliskin, Holdnack, Axelrod, & Randolph, 2007; Jeyakumar, Warriner, Raval, & Ahmad, 2004; Lange & Iverson, 2008; Ringe, Saine, Lacritz, Hynan, & Cullum, 2002; Ryan & Ward, 1999), and the latest version, Wechsler Adult Intelligence Scale – Fourth Edition (WAIS-IV) (Denney, Ringe, & Lacritz, 2015; Girard, Axelrod, Patel, & Crawford, 2015; Meyers, Zellinger, Kockler, Wagner, & Miller, 2013; Ryan, Kreiner, Gontkovsky, & Umfleet, 2015).

From both research and clinical perspectives, the use of full psychological measurements is often optimum; however, participants have limited time available and tend not to cooperate well if the tests are too long. The duration of the test is therefore critical. If researchers need only an estimation of a participant's overall cognitive function, a shorter but equally valid short-form test to estimate the FSIQ is an excellent practical choice.



McNemar (1950) suggested that the sample population should not be too homogeneous; therefore, considering the anticipated use of the FSIQ score for screening purposes in the future, this study involved healthy adult participants. Three possible strategies can be used to develop a valid short-form WAIS assessment. The first strategy is to reduce the number of items in all subtests, while the second and third strategies rely on deleting subtests to develop a representative set of subtests to yield an optimal approximation of FSIQ. The second strategy uses a prorated calculation of FSIQ based on a selection of subtests, while the third strategy relies on a regression-based equation to select the subtests that should be used to estimate FSIQ (for reviews, see King & King, 1982; Levy, 1968; Silverstein, 1990). From a validity perspective, short forms that reduce the number of items rather than the number of subtests are less reliable than those containing identifying the optimal combinations of four or five subtests (Silverstein, 1990); therefore, in this study, we compared several sets of subtests to estimate the FSIQ scores given by several short forms of the WAIS-IV-ID.

In Table 1, we present an overview of the short forms studied here. Since Wechsler's intelligence scales were designed to measure verbal and nonverbal (performance) factors of cognitive ability, we compared four two-subtest short forms, each of which included a verbal factor and a performance factor (SF1–SF4). These two-subtest short forms were all based on previous studies, except for SF4, which was based on the results of our stepwise regression providing a good fit to predict FSIQ. Furthermore, we compared short forms consisting of four, five, six, or seven WAIS-IV subtests. The WAIS-IV has a four-factor structure (Wechsler, 2008b), comprising indexes of Verbal Comprehension (VC), Perceptual Reasoning (PR), Working Memory (WM), and Processing Speed (PS). Our previous study showed that this four-factor model is a better fit than a five-factor model (Suwartono, Hendriks, Hidajat, Halim, & Kessels, submitted). In order to reflect the structure of the Indonesian-language version of the WAIS-IV (WAIS-IV-ID; Suwartono, Halim, Hidajat, Hendriks, and Kessels, 2014), we selected the most representative subtest from each factor based on our best-fit four- (SF6) and five-factor models (SF13). We also added two further short forms based on our data collection: SF5 was developed using the highest independent correlation coefficients between each subtest and the FSIQ, while SF12 combines SF4 and SF5. In addition, we included several short forms based on previous studies of the WAIS short forms, provided that those subtests were still available in the WAIS-IV (SF7–SF11). In total, we compared the psychometric properties of 13 short forms with the WAIS-IV-ID.

**Table 1: Model specifications.**

| Short forms | Subtests | Basis of the model |
|---|---|---|
| SF1 | IN and BD | Maxwell (1957); Ringe, Saine, Lacritz, Hynan, and Cullum (2002) |
| SF2 | VC and BD | Cyr and Brooker (1984); Denney, Ringe, and Lacritz (2015); Maxwell (1957); Ringe, Saine, Lacritz, Hynan, and Cullum (2002); Silverstein (1982) |
| SF3 | IN and PC | Kaufman, Ishikuma, and Kaufman-Packer (1991) |
| SF4 | MR and CD | Stepwise regression method |
| SF5 | IN, MR, AR, and CD | Correlations with FSIQ |
| SF6 | IN, MR, DS, and CD | The best-fit four-factor model (Suwartono et al., submitted) |
| SF7 | CO, BD, AR, and CD | McNemar (1950) |
| SF8 | SI, PC, AR, and CD | Kaufman, Ishikuma, and Kaufman-Packer (1991) |
| SF9 | DS, AR, SS, and CD | Wechsler (CPI; 2008) |
| SF10 | SI, VC, IN, BD, MR, and VP | Wechsler (GAI; 2008) |
| SF11 | SI, IN, BD, PC, DS, AR, and CD | Ward (1990) |
| SF12 | IN, MR, DS, AR, and CD | Combination of SF4 and SF5 |
| SF13 | IN, MR, VP, DS, and CD | The best-fit five-factor model (Suwartono et al., 2017) |

Notes. IN = Information; BD = Block Design; VC = Vocabulary; PC = Picture Completion; MR = Matrix Reasoning; CD = Coding; AR = Arithmetic; DS = Digit Span; CO = Comprehension; SI = Similarity; SS = Symbol Search; VP = Visual Puzzle.

## METHODS

### SAMPLING METHOD

In order to identify the best short form for the WAIS-IV-ID, participants were recruited from the Indonesian population using the standardized protocol described in the WAIS-IV manual (Wechsler, 2008b). A quota sampling method was used to represent the population census data from the six largest islands in Indonesia; 57.49% (136.6 million individuals) live on Java, 21.31% on Sumatra, 7.31% on Sulawesi, 5.8% on Kalimantan, and 5.50% live on Nusa Tenggara and Bali (BPS, 2012b). A second study was conducted to assess the validity of the chosen short form. This validation study was conducted in Jakarta using a convenience sampling method.

### PARTICIPANTS

Data were collected from 1,745 participants, including 736 men (42.20%) and 1,009 women (57.80%), whose ages ranged from 16.0 to 69.9 years old (*M* = 31.74, *SD* = 14.12). Their education levels were as follows: 8% completed only junior high school, 48.60% completed senior high school, 37.0% had obtained an undergraduate degree (BA or BSc), and 6.40% had completed a postgraduate degree (MA, MSc, or Ph.D.). Most of the participants were from Java (58.20%), and the rest were recruited from Sumatra (16.60%), Sulawesi (10.50%), Borneo (8%), and Nusa Tenggara and Bali (6.70%).

A total of 20 participants were involved in the validation study, including three men (15%) and 17 women (85%). Their ages ranged from 17 to 60 years old (*M* = 29.80, *SD* = 13.80). About 60% of these participants completed an undergraduate degree (BA, BSc), 25% completed senior high school, 5% completed junior high school, and the remaining 5% completed a Master degree. All participants were from Jakarta.



### INSTRUMENTS

The WAIS-IV-ID consists of ten core subtests, Block Design (BD), Similarity (SI), Digit Span (DS), Matrix Reasoning (MR), Vocabulary (VC), Arithmetic (AR), Symbol Search (SS), Visual Puzzle (VP), Information (IN), Coding (CD); as well as five supplemental subtests, Letter-Number Sequencing (LN), Figure Weights (FW), Comprehension (CO), Cancellation (CA), and Picture Completion (PC). The items of the WAIS-IV-ID subtests are same as or equivalent to those of the WAIS-IV-US (Suwartono et al., 2014; Wechsler, 2008a).

### PROCEDURE

The WAIS-IV-ID was administered individually following the guidelines in the test manual (Wechsler, 2008a). All examiners (*N* = 98) were undergraduate Psychology students in the last year of their study or recent Psychology graduates. They were all extensively trained and had passed the test administration course for the WAIS-IV-ID. All participants gave their written informed consent before participating in the study. The research proposal and informed consent forms were approved by all institutions who agreed to participate. After the best short form was selected, a second study was performed. New participants were recruited to participate in two test sessions. In the first session they took the short-form test, while in the second they took the full WAIS-IV-ID. The interval between the first and second sessions ranged from 27–50 days (*M* = 36.7, *SD* = 7.06).

**STATISTICAL ANALYSIS**

The procedure for determining the subtests used in the short forms was based either on those used in previous research, those indicated by a stepwise regression, or those indicated by independent correlations of subtests with the FSIQ. A regression analysis was performed after selecting the short forms. The goodness of fit for the prediction model was determined using a modified version of $R^2$ adjusted for the number of predictors in the model (Field, 2013). Next, a regression equation was formulated for transforming the standardized subtest scores into an estimated FSIQ score ($FSIQ_{Est}$). The mean discrepancies between the FSIQ and FSIQEst scores were tested using a dependent t-test. The classification accuracy was calculated, which is the percentage agreement of intelligence classification between the $FSIQ_{Est}$ estimated by the short form and the FSIQ determined using the full WAIS-IV-ID (Jones, 1967; Levy, 1968; Mumpower, 1964; Silverstein, 1990).

The reliability of each short form was calculated using a composite reliability formula (Crawford, Allum, & Kinion, 2008; Nunally & Bernstein, 1994):

$$r_{YY} = 1 - \frac{k - \Sigma r_{xx}}{Ry}$$

where $r_{YY}$ is the reliability coefficient of the subtest combination, $k$ is the number of component subtests, $r_{xx}$ is the reliability coefficient of the short form's components, $R_y$ is the sum of coefficient correlations in the component correlation matrix.

The following formula was used to calculate the standard error of measurement (Crawford, Allum, & Kinion, 2008; Ley, 1972):

$$\mathrm{SEM}x = \mathrm{S_x} \ \sqrt{1 - \mathrm{rxx}}$$

where Sx is the standard deviation of the short form and rxx is its reliability coefficient. The validation of the short forms was determined by the correlation between their FSIQEst and the FSIQ (Silverstein, 1990).

The unidimensionality of nine models (SF5–SF13) was also tested. The ω coefficient of reliability considers the factor loadings from a factor analysis and remains unbiased for uncorrelated errors (Padilla & Divers, 2016). The ω reliabilities of the short forms were calculated in R following the method outlined by Peters (2014). The ω reliability is based on hierarchical factor model and can be used with multidimensional scales. The standardized subtest scores were analyzed with LISREL 8.80 (Jöreskog & Sörbom, 2006). The chi-square ratio ($\chi^2$/df), p-value, Root Mean Square Error of Approximation (RMSEA), and Akaike's Information Criterion (AIC) were applied to assess the goodness of fit of the theoretical models with the sample data. Details of these fit indexes can be found in various sources (see Hu & Bentler,1999; Kline, 2005). A good model would provide nonsignificant goodness-of-fit results at a .05 threshold (Barrett, 2007). RMSEA determines the deviation from a perfect fit; Hu and Bentler (1999) suggest that RMSEA values less than or equal to .06 indicate a good fit. The AIC compares different models; smaller AIC values indicate a better fit after accounting for model complexity (Akaike, 1987).

The following criteria were applied to the evaluation of the short forms (Levy, 1968; Silverstein, 1990): the magnitude of correlation between FSIQ and FSIQEst, the mean differences between FSIQ and FSIQEst revealed by paired sample t-tests, and the accuracy of the classification agreement between FSIQ and FSIQEst. The basic requirement for any short form is a minimum correlation of .90 with the score of the full assessment (Groth-Marnat, 2009).

In the second study, Wilcoxon Signed Ranks Test were used to determine whether the FSIQ and FSIQEst were significantly different. Spearman's rho described the FSIQEst and FSIQ correlation, in which each subtest in the short forms was represented. To ascertain the reliability of each short form, a composited reliability formula was also used. The standard error of each measurement was calculated.



## RESULTS

Table 2 shows the results of all criteria used to evaluate the short forms. All of the regression analyses on the various short forms indicated that they significantly predicted FSIQ. The goodness of fit of our prediction models to estimate the FSIQ was calculated with an adjusted R2, which ranged from .60 (SF3) to .94 (SF11). Short forms that could serve as a predictor model should reflect more than 90% of the variance in FSIQEst. Based on the adjusted R2, we found that SF5, SF11, SF12, and SF13 were qualified to predict the FSIQ.

The first short form evaluation was the determination of the correlation coefficients between the FSIQ and the FSIQEst values. For all of the short forms, the correlations between FSIQ and FSIQEst were significant ($p < .01$), ranging from .77 (SF3) to .97 (SF11); however, only SF5–SF13 had correlation coefficients higher than .90, and could contribute to the FSIQ around 81%.

Second, we tested whether the FSIQ and FSIQEst values (based on the regression models) were significantly different using a paired t-test analysis. None of the FSIQEst scores from the short forms were significantly different to the FSIQ value.

The third evaluation was a determination of classification accuracy. Classification accuracy is the number of correct predictions made, divided by the total number of predictions made, multiplied by 100 to turn it into a percentage. We classified the FSIQ and FSIQEst for each participant using a manual, then compared whether they were classified into the same classifying IQ scores category. The classification accuracy of the FSIQEst values of each short form compared with the FSIQ ranged from 56.79% (SF3) to 81.03% (SF11). We found that SF5, SF11, SF12, and SF13 were the four short forms with the highest classification accuracies.

**Table 2: Results on the criteria to evaluate each short form.**

| Short forms | Adjusted $R^2$ | Correlation between FSIQ and $FSIQ_{Est}$ (*r*-value) | Difference between FSIQ and $FSIQ_{Est}$ (*t* value) | Significance (*p* value) | Classification accuracy (%) |
|---|---|---|---|---|---|
| SF1 | .65 | .81** | 1.59 | .11 | 58.45 |
| SF2 | .66 | .81** | 0.31 | .76 | 59.26 |
| SF3 | .60 | .77** | 0.03 | .97 | 56.79 |
| SF4 | .78 | .88** | 0.03 | .97 | 67.62 |
| SF5 | .91 | .95** | -0.25 | .81 | 77.08 |
| SF6 | .90 | .95** | 0.28 | .78 | 75.76 |
| SF7 | .87 | .93** | 0.13 | .90 | 71.75 |
| SF8 | .85 | .92** | -0.59 | .55 | 72.49 |
| SF9 | .83 | .91** | -0.50 | .62 | 69.40 |
| SF10 | .87 | .93** | -0.95 | .34 | 75.30 |
| SF11 | .94 | .97** | 0.97 | .33 | 81.03 |
| SF12 | .92 | .96** | -0.07 | .94 | 79.08 |
| SF13 | .93 | .96** | -0.74 | .46 | 79.48 |

**. Correlation is significant at the .01 level (two-tailed).

Table 3 shows the reliability parameters for all short forms. The composite reliability coefficients ranged from .82 (SF2) to .95 (SF12 and SF13). The short forms comprised of four or more subtests (SF5–SF13) had composite reliability coefficients above .90, although this was also achieved by SF4, which consists of two subtests (MR and CD) and had a coefficient of .91. The ω reliability coefficients of the short forms with four or more subtests ranged from .69 (SF10) to .81 (SF12).

SF5, SF7, and SF8 had insignificant chi-square test values ($p > .01$) for their structural validities, implying that the subtests included these short forms were sufficiently able to measure IQ as a unidimensional construct. Moreover, these short forms fulfilled the criteria for RMSEA (RMSEA < .06), for which smaller values indicate a better model fit and predicted values close to the observed data values.

Considering the evaluation criteria from Table 2, we concluded that SF5 is the best short form test for the WAIS-IV-ID. SF5 produced the following values: $\chi^2$/df = 2.91, p > .01, RMSEA = .03, and AIC model = 21.82. This choice was further supported by combining the results of the evaluation criterion, test reliability, and structural validity. SF5 comprises the Verbal Comprehension factor IN, the Perceptual Reasoning factor MR, the Working Memory factor AR, and the Processing Speed factor CD

**Table 3: Reliability and validity of the individual models.**

| Short form | Reliability | | SEM[3] | Goodness of fit model | | | |
|---|---|---|---|---|---|---|---|
| | Composite[1] | Omega[2] | | $\chi^2$/df | p | RMSEA | AIC model |
| SF1 | .85 | not applicable | 4.02 | not applicable | | - | - |
| SF2 | .82 | not applicable | 4.40 | not applicable | | - | - |
| SF3 | .88 | not applicable | 3.52 | not applicable | | - | - |
| SF4 | .91 | not applicable | 3.37 | not applicable | | - | - |
| SF5 | .94 | .77 | 3.03 | 2.91 | .05 | .03 | 21.82 |
| SF6 | .94 | .72 | 3.08 | 20.51 | .00 | .11 | 57.01 |
| SF7 | .91 | .71 | 3.69 | 3.54 | .03 | .04 | 23.07 |
| SF8 | .91 | .72 | 3.68 | 1.61 | .20 | .02 | 19.22 |
| SF9 | .92 | .73 | 3.25 | 111.39 | .00 | .25 | 238.78 |
| SF10 | .94 | .69 | 3.05 | 60.16 | .00 | .18 | 565.43 |
| SF11 | .94 | .72 | 3.02 | 20.70 | .00 | .11 | 317.74 |
| SF12 | .95 | .81 | 2.83 | 11.82 | .00 | .08 | 79.11 |
| SF13 | .95 | .73 | 2.90 | 25.18 | .00 | .12 | 145.92 |

[1] Based on the composite reliability (Crawford, Allum, & Kinion, 2008; Nunally & Bernstein, 1994).
[2] Based on the Omega-hierarchical value (Peters, 2014).
[3] Based on the composite reliability coefficient.

A second study was performed to assess the validity of using SF5 as a short form of the WAIS-IV-ID. A Wilcoxon Signed Ranks test indicated no significant difference between the FSIQ calculated by the full test and the $FSIQ_{Est}$ values predicted using SF5 ($Z = -1.68$, $p = .09$). The time taken to administer the four SF5 subtests was 15–38 minutes ($M = 25.45$, $SD = 5.36$). The SF5 classification accuracy was 70%, while its composite reliability coefficient was .94 (SEM = 2.73). The Spearman's rho correlation between FSIQ and the FSIQEst predicted with SF5 was $r_{(18)} = .89$, p < .01. Each subtest in the short form had a significant correlation with FSIQ, ranging from .59 (CD) to .76 (AR). The largest coefficient correlations for each index were IN with VCI ($r_{(18)} = .83$, p < .01), MR with PRI ($r_{(18)} = .71$, p < .01), AR with WMI ($r_{(18)} = .74$, p < .01), and CD with PSI ($r$ (18) = .83, p < .01).

Chapter 5

More details about the correlations between the short-form and Full Scale indexes can be found in Table 4.

**Table 4: Correlation between the short form and Full Scale IQ of WAIS-IV-ID.**

| SF5 | FSIQ | VCI | PRI | WMI | PSI |
|---|---|---|---|---|---|
| $FSIQ_{Est}$ | .89** | .74** | .68** | .78** | .58** |
| IN | .69** | .83** | .41 | .60** | .49* |
| MR | .67** | .49* | .71** | .49* | .29 |
| AR | .76** | .58** | .61** | .74** | .20 |
| CD | .59** | .50* | .31 | .47* | .83** |

Note: ** Correlation is significant at the .01 level (two-tailed); * Correlation is significant at the .05 level (two-tailed).

## DISCUSSION

In the current study, we examined the reliability and validity of 13 short forms of WAIS-IV-ID, which consisted of two, four, five, six, or seven subtests. Overall, no significant differences were observed between the FSIQ value determined using the full test and the FSIQEst values predicted using the short forms. For the short forms comprising two subtests, we found that SF4 (subtests MR and CD) yielded the best estimate of the Full Scale IQ. This result does not support the findings of Denney, Ringe, and Lacritz (2015), who reported that SF2, consisting of VC and BD, was the two-subtest short form with the best fit. In our study, SF4 had a higher classification accuracy, correlation with the FSIQ, and reliability than SF2.

All short forms consisting of four subtests yielded reliable FSIQEst values; however, only SF5, SF7 (McNemar, 1950), and SF8 (Kaufman et al., 1991) had satisfactory goodness of fit index results ($\chi^2$/df, p > .01, RMSEA < .06, AIC model). Our findings suggest that SF5 (subtests IN, MR, AR, and CD) had the highest predictive value (based on the adjusted R2 value), percentage of classification accuracy, and coefficient of reliability among the four-subtest short forms. For the short forms that consist of five or more subtests (SF10–SF13), the goodness of fit index results were not satisfactory ($\chi^2$/df, p < .01, RMSEA > .06, AIC model).

Decisions on which subtests to include may depend on the type of information required (Groth-Marnat, 2009). When time limitation is the primary factor for reducing the evaluation, short forms containing more than four subtests may not be an ideal solution (Denney et al., 2015). Our second study showed that time required to administer SF5 was approximately 25 minutes. We, therefore, recommend the use of SF5 in place of the full WAIS-IV-ID to reduce the time required while retaining a maximum validity.

The strength of the current study is the use of a normal standardized sample. The development of short forms are usually based on a clinical sample, as was the case for the short forms of the original US version of the WAIS-IV (Denney et al., 2015; Girard et al., 2015; Meyers et al., 2013; Ryan et al., 2015). We also included several recently proposed short forms based on our findings in previous research (Suwartono et al., submitted). Of these, SF5 is recommended for use in both research or clinical settings to estimate FSIQ. Our second study investigated the psychometric properties of SF5, and despite the small sample size, the results were promising. Further research is needed to replicate these findings in a larger sample population. Furthermore, future studies should investigate the psychometric properties of the short forms in clinical populations.

Nevertheless, the current study does have some limitations. We developed our short forms based on data obtained from a sample who completed the full WAIS-IV battery of tests. If the motivation and attention of the participants varied during the administration of the full test, these scores may have affected the selection of subtests used to estimate FSIQEst (Thompson, 1987). Further research should investigate the validity of the short forms in an independent sample, and examine whether the classification agreement rates remain high, indicating the best trade-off between a reduced administration time and a potential loss in reliability and validity (Girard et al., 2015; Smith et al., 2000).

It should be noted that the results obtained using any short form should be interpreted with caution, as these only represent an estimate of FSIQ (King & King 1982; Silverstein, 1990). Short forms are best used to obtain a quick indication of intelligence, determining whether an additional (neuro)psychological assessment is required (Groth-Marnat, 2009). In addition, the short forms might be useful for research in which individual classifications or absolute FSIQs do not have diagnostic consequences (Kaufman & Kaufman, 2001).

# 6.

# Reliability and Receiver–Operating Characteristic Analysis for Evaluating the Indonesian Wechsler Adult Intelligence Scale – Fourth Edition (WAIS–IV–ID) in Clinical Groups

**In preparation:**
Suwartono, C., Hendriks, M. P. H., Halim, M. S., and Kessels, R. P. C. Reliability and Receiver-Operating Characteristic Analysis for Evaluating the Indonesian Wechsler Adult Intelligence Scale – Fourth Edition (WAIS-IV-ID) in Clinical Groups.

## ABSTRACT

The Wechsler Adult Intelligence Scale – Fourth Edition (WAIS-IV) has previously been used to examine cognitive functioning abilities both in non-clinical and clinical groups. An Indonesian version of WAIS-IV (WAIS-IV-ID) had been comprehensively psychometrically analyzed in non-clinical groups. In the present study, we investigate the reliability and clinical utility of the WAIS-IV-ID in three clinical groups: patients with Alzheimer's disease, patients with schizophrenia, or adults with an intellectual disability. Using a receiver-operating characteristic (ROC) analysis, we tested the diagnostic validity of the WAIS-IV-ID to find the optimal cut-off scores for all indexes and subtests. The non-clinical standardized sample scores were higher than those of the individuals in each of the clinical groups. In general, all scores from WAIS-IV-ID were sufficiently reliable to distinguish between non-clinical and clinical groups. The ROC analysis indicated that Full Scale IQ could be used as a detection index for non-clinical and clinical groups. Future research should investigate the external validation and long-term stability of using WAIS-IV-ID for a clinical population.



**KEYWORDS**

Alzheimer's disease, Indonesia, intellectual disability, receiver-operating characteristic, schizophrenia, sensitivity, specificity, WAIS-IV

One of the critical tasks for clinicians is diagnosing a patient's clinical condition. An incorrect diagnosis would delay the therapy a patient needs and decrease treatment efficiency. For psychological conditions, clinicians need assessment instruments, such as psychological tests, that can reliably and accurately distinguish between the clinical and non-clinical population.

Wechsler's intelligence scales are the most popular psychological measurements of intelligence (Kaufman & Lictenberger, 2000). The latest revision, the Wechsler Adult Intelligence Scale – Fourth Edition (WAIS-IV), is a comprehensive instrument that assesses adult cognitive functioning abilities, measuring four broad domains; verbal comprehension, perceptual reasoning, working memory, and processing speed (Wechsler, 2008a).

In our previous study, we determined that the reliability of an Indonesian version of WAIS-IV (WAIS-IV-ID) was excellent for a non-clinical sample; however, it has not previously been investigated in clinical samples.

Receiver-operating characteristic (ROC) analyses have most frequently been used in medical research for visualizing and analyzing the performance of diagnostic assessment tools such as psychological tests. Furthermore, ROC analyses can be utilized to evaluate the diagnostic validity of psychological tests (Pintea & Moldovan, 2009). ROC analyses are a valuable tool, not only to select the optimal cut-off for a variety of clinical situations, but also to balance the inherent tradeoffs between sensitivity and specificity (Zou, O'Malley, & Mauri, 2007); nevertheless, the optimal cut-off is dependent on the clinical use of the test (Westin, 2001).

The purpose of the present study is to use a ROC analysis to determine the reliability and utility of the WAIS-IV-ID in individuals diagnosed with three clinical conditions: Alzheimer's disease, schizophrenia, and intellectual disability. Although, there were various research results in schizophrenia, the consensus is that multiple aspects of cognition are impaired, including attention, working memory, encoding acquisition, and executive ability (Blanchard & Neale, 1994; Heinrichs, 1993). Therefore, non-clinical individuals were predicted to achieve higher scores than the clinical individuals.



## METHODS

### PARTICIPANTS

For the clinical groups (N = 136), the participants were recruited with agreement from the patients themselves, the treating physician, and the hospital. A matched non-clinical sample (N = 142) was recruited using the standardized protocol outlined in the WAIS-IV manual (Wechsler, 2008b), based on the characteristics of the clinical sample. The participants were recruited from Java island, since WAIS-IV-ID would be used predominantly on Java, where more than half of the Indonesian population (57.49% or 136.6 million individuals) reside (BPS, 2012).

In total, 278 individuals participated in this study. The non-clinical group comprised 142 participants (58.5% men, 41.5% women), whose ages ranged from 16 to 71 (M = 36.46, SD = 17.11). Of the non-clinical group, 39.40% participants had completed junior high school, 43% had completed senior high school, 12.70% had obtained an undergraduate degree (BA or BSc), and 4.90% had completed a postgraduate degree (MA, MSc, or Ph.D.). The ethnicities of the non-clinical participants were Javanese (40.80%), Tionghoa (38.70%), Sundanese (7.70%), and Other (12.80%).

Three clinical samples were included in this study; patients with Alzheimer's disease (n = 44), patients with schizophrenia (n = 48), and adults with an intellectual disability (n = 44). Details of each clinical sample can be found in Table 1. The patients with Alzheimer's disease were recruited by research assistants using referrals from hospitals in Yogyakarta and Tangerang between January and November 2015. The patients with schizophrenia were recruited by research assistants using referrals from mental hospitals in Malang and Jakarta between May and July 2015. Finally, adults with an intellectual ability were recruited by research assistants from a rehabilitation center in Jakarta between January and March 2015 (see Table 1).

**Table 1: Demographic data.**

| Variables | Total clinical sample | Alzheimer's disease | | Schizophrenia | | Intellectual disability | |
|---|---|---|---|---|---|---|---|
| | | Clinical | Healthy | Clinical | Healthy | Clinical | Healthy |
| Sample size | 136 | 44 | 46 | 48 | 51 | 44 | 45 |
| Sex (%) | | | | | | | |
| Male | 61.8 | 40.91 | 45.65 | 85.42 | 86.27 | 56.81 | 40 |
| Female | 38.2 | 59.09 | 54.35 | 14.58 | 13.73 | 43.19 | 60 |
| Education (%) | | | | | | | |
| Junior high school | 50.70 | 15.91 | 10.87 | 37.50 | 11.76 | 100 | 100 |
| Senior high school | 32.40 | 31.82 | 34.78 | 62.50 | 88.24 | | |
| BA or BSc | 11.80 | 36.36 | 39.13 | | | | |
| MA or MSc | 4.40 | 13.64 | 13.04 | | | | |
| PhD | 0.70 | 2.27 | 2.18 | | | | |
| Ethnicity (%) | | | | | | | |
| Javanese | 25.70 | 34.09 | 21.74 | 37.50 | 45.10 | 4.54 | 53.33 |
| Sundanese | 6.60 | 6.82 | 2.17 | 12.50 | 15.69 | - | 6.67 |
| Tionghoa | 19.10 | 47.73 | 63.04 | 8.33 | 39.21 | 2.27 | 11.11 |
| Other | 48.50 | 11.36 | 13.05 | 41.67 | | 91.13 | 28.89 |
| Age (in years) | | | | | | | |
| *M* | 41.51 | 68.23 | 58.69 | 32.23 | 26.31 | 24.91 | 25.24 |
| *SD* | 20.64 | 8.97 | 3.10 | 7.61 | 6.47 | 9.19 | 10.79 |
| Range | 16–82 | 46–82 | 56–71 | 21–54 | 21–52 | 16–44 | 16–48 |

## INSTRUMENTS

All participants completed the WAIS-IV-ID, which consists of ten core subtests (Block Design (BD), Similarity (SI), Digit Span (DS), Matrix Reasoning (MR), Vocabulary (VC), Arithmetic (AR), Symbol Search (SS), Visual Puzzle (VP), Information (IN), Coding (CD)) and five supplemental subtests (Letter-Number Sequencing (LN), Figure Weights (FW), Comprehension (CO), Cancellation (CA), and Picture Completion (PC)). The items included in the WAIS-IV-ID subtests are identical or equivalent to those of the WAIS-IV-US (Suwartono et al., 2014; Wechsler, 2008a).

## PROCEDURE

The WAIS-IV-ID was administered according the guidelines indicated within the administration manual (Wechsler, 2008a). The WAIS-IV-ID was administered by psychologists, psychometrists, and research assistants (N = 39) who were extensively trained and had passed an internal test for administering and scoring the WAIS-IV-ID. All participants gave their written informed consent before participating in the study. The research proposal and informed consent forms were approved by all institutions who agreed to patient participation.

The patient groups were matched with the non-clinical samples according to their education, sex, and age. An analysis of the non-clinical and clinical samples revealed no significant difference in the education levels ($\chi^2_{(4)} = 4.09$, $p = .39$) or sex ($\chi^2_{(1)} = .32$, $p = .57$) of the participants in each group; however, the variance in age between the non-clinical and clinical samples was not equal ($t_{(262.42)} = 2.21$, $p = .03$).

## STATISTICAL ANALYSES

Coefficient alpha (α) was used for reliability testing, and was calculated as a function of the reliabilities of the components and their intercorrelations using the following formula (Nunnally & Bernstein, 1994, p.269):

$$r_{yy} = 1 - \frac{k - \Sigma rxx}{\bar{R}y}$$

where $k$ is the number of components, rxx is the reliabilities of the components, and $\bar{R}y$ is the sum of the elements of the correlation matrix of the components (including the unities in the diagonal).

Prior to the ROC analysis, the non-clinical and clinical samples were compared to analyze whether the scores between the two groups were significantly different. A Levene's test for the equality of variances was performed to determine whether the two groups had an equal variance, then the *t*-value was determined using a *t*-test for independent samples. The Cohen's d score was also calculated (Lakens, 2013).

For the ROC analysis, the area under the ROC curve (AUC), Youden index, sensitivity, specificity, cut-off, and likelihood ratios (LR) were calculated using MedCalc

Statistical Software version 16.2.1. The AUC combines a measure of sensitivity and specificity for assessing the inherent validity of diagnostic tests (Kumar & Indrayan, 2011). Its value varies between 0 and 1, where 1 reflects a diagnostic test that is perfect at differentiating between a clinical and a non-clinical individual. According to Fawcett (2006), a test should have an AUC value above 0.50. AUC values can be interpreted as low (.50 < AUC ≤ .70), moderate (.70 < AUC ≤ .90), or high (.90 < AUC ≤ 1.00) (Swets, 1988). The Youden index is a global measure of the overall discriminative power of a test's performance. It is calculated from the sum of sensitivity and specificity minus one (Youden, 1950). Youden's index is not sensitive to distinguish between sensitivity and specificity; therefore, sensitivity and specificity data were also provided. Sensitivity is the probability of getting a positive test result from an individual with a clinical condition while specificity is the probability of obtaining a negative test result from an individual without a clinical condition (Šimundić, 2008). The best test overall has a score of .79 for sensitivity and .86 for specificity (Westin, 2001). The LR is a ratio of the probability that a test result is correct to the probability that the test result is incorrect (Šimundić, 2008). The sensitivity and specificity values of the test are needed to calculate the LR, which is determined for both positive and negative test results, respectively expressed as the LR+ (sensitivity / (1 – specificity)) and LR– (1 – sensitivity/specificity). An LR+ above 10 and an LR– below 0.10 is considered strong evidence to rule in or rule out diagnoses, respectively, in most situations (Deeks & Altman, 2004).

## RESULTS

Table 2 shows the reliabilities of the WAIS-IV-ID subtests calculated for the clinical samples (individuals with Alzheimer's disease, schizophrenia, or an intellectual disability).

**Table 2: Reliability of the WAIS-IV-ID in clinical samples.**

| Internal consistency | Total clinical sample | Patients with Alzheimer's disease | Patients with schizophrenia | Adults with an intellectual disability |
|---|---|---|---|---|
| FSIQ | | | | |
| VCI | .95 | .95 | .86 | .91 |
| PRI | .95 | .94 | .92 | .95 |
| WMI | .95 | .91 | .90 | .82 |
| PSI* | N/A | N/A | N/A | N/A |
| SI | .87 | .85 | .76 | .86 |
| VC | .94 | .94 | .84 | .88 |
| IN | .78 | .82 | .63 | .66 |
| CO | .81 | .84 | .61 | .61 |
| BD | .86 | .83 | .78 | .89 |
| MR | .89 | .81 | .87 | .91 |
| VP | .85 | .87 | .83 | .81 |
| FW | .88 | .90 | .82 | .81 |
| PC | .75 | .76 | .74 | .76 |
| DS | .90 | .82 | .84 | .80 |
| AR | .83 | .79 | .67 | .52 |
| LN | .87 | .87 | .81 | .73 |
| SS* | N/A | N/A | N/A | N/A |
| CD* | N/A | N/A | N/A | N/A |
| CA* | N/A | N/A | N/A | N/A |

* SS, CD, and CA are PSI subtests; the coefficient α is not the proper reliability estimate for these groups, and these data are not available as the test-retest was not performed in the clinical populations.

Chapter 6

The reliability coefficients for the patients with Alzheimer's ranged from .91 (WMI) to .95 (VCI). The subtest reliability coefficients ranged from .76 (PC) to .94 (VC). Meanwhile, for patients with schizophrenia, the indexes showed reliability coefficients ranging from .86 (VCI) to .92 (WMI). The subtest reliability coefficients ranged from .61 (CO) to .87 (MR). Finally, for individuals with an intellectual disability, the indexes ranged from .82 (WMI) to .95 (PRI). In this group, the subtest reliability coefficients ranged from .52 (AR) to .91 (MR).

**Table 3: The WAIS-IV-ID performances in clinical samples compared with matched controls.**

| Tests | Clinical samples (*N* = 136) | | Non-clinical sample (*N* = 142) | | Differences between group | | | | | Cohen's d |
|---|---|---|---|---|---|---|---|---|---|---|
| | *M* | *SD* | *M* | *SD* | Levene's Test | Sig. | t | df | Sig. (two-tailed) | |
| FSIQ | 60.32 | 13.08 | 92.36 | 13.93 | 0.65 | .42 | 19.75 | 276 | .00 | 2.37 |
| VCI | 64.79 | 11.99 | 90.84 | 11.13 | 2.04 | .15 | 18.78 | 276 | .00 | 2.25 |
| SI | 3.40 | 2.25 | 7.68 | 2.00 | 3.29 | .07 | 16.76 | 276 | .00 | 2.01 |
| VC | 4.39 | 2.88 | 9.84 | 2.38 | 11.63 | .00 | 17.13 | 262.08 | .00 | 2.06 |
| IN | 3.98 | 1.73 | 7.58 | 2.58 | 37.86 | .00 | 13.75 | 247.32 | .00 | 1.65 |
| CO | 2.79 | 1.71 | 6.36 | 2.51 | 32.60 | .00 | 13.93 | 249.69 | .00 | 1.67 |
| PRI | 69.38 | 14.04 | 95.51 | 14.66 | 0.42 | .52 | 15.16 | 276 | .00 | 1.82 |
| BD | 5.32 | 3.09 | 9.24 | 2.76 | 2.83 | .09 | 11.17 | 276 | .00 | 1.34 |
| MR | 4.04 | 2.59 | 8.82 | 3.07 | 4.58 | .03 | 14.05 | 271.92 | .00 | 1.69 |
| VP | 4.93 | 2.40 | 9.75 | 3.02 | 11.23 | .00 | 14.78 | 266.90 | .00 | 1.77 |
| FW | 4.98 | 2.37 | 10.08 | 3.58 | 18.58 | .00 | 13.69 | 246.17 | .00 | 1.64 |
| PC | 3.09 | 2.15 | 6.76 | 3.00 | 14.14 | .00 | 11.78 | 255.86 | .00 | 1.41 |
| WMI | 66.70 | 14.05 | 95.98 | 15.09 | 0.00 | .96 | 16.73 | 276 | .00 | 2.01 |
| DS | 3.99 | 2.90 | 9.31 | 3.03 | 0.09 | .77 | 14.96 | 276 | .00 | 1.80 |
| AR | 4.18 | 2.58 | 9.31 | 3.18 | 5.60 | .02 | 14.80 | 268.99 | .00 | 1.78 |
| LN | 3.77 | 2.82 | 9.51 | 3.06 | 4.36 | .04 | 15.64 | 252.34 | .00 | 1.88 |
| PSI | 62.36 | 12.93 | 93.75 | 16.31 | 7.58 | .01 | 17.82 | 266.80 | .00 | 2.14 |
| SS | 3.28 | 2.65 | 8.82 | 3.25 | 2.11 | .15 | 15.54 | 276 | .00 | 1.86 |
| CD | 2.97 | 2.28 | 8.85 | 3.51 | 25.97 | .00 | 16.63 | 243.20 | .00 | 2.00 |
| CA | 2.48 | 2.05 | 7.80 | 3.14 | 10.66 | .00 | 16.38 | 244.84 | .00 | 1.97 |

A series of t-tests indicated statistically significant differences between the non-clinical and clinical participants for their FSIQ and all index and subtest scores of the WAIS-IV-ID. The non-clinical sample showed significantly higher scores for all indexes and subtests. For the overall clinical sample, the effect size of the FSIQ was large ($d$ = 2.37). The effect sizes of the index scores were large, ranging from 1.82 (PRI) to 2.25 (VCI). For the subtests, the effect size ranged from 1.34 (BD) to 2.06 (VC).

Significant differences were observed between the scores obtained by the non-clinical sample and the patients with Alzheimer's disease for all indexes and subtests of the WAIS-IV-ID (Table 4). The non-clinical sample performed better in all indexes and subtests. The effect size of FSIQ was large (1.99) for this comparison. The effect sizes of the indexes were also large, ranging from 1.54 (VCI) to 1.76 (PRI and PSI). For the subtests, the effect sizes ranged from .92 (VC) to 2.16 (CA).

**Table 4: WAIS-IV-ID performances of the patients with Alzheimer's disease compared with the matched controls.**

| Tests | Patients with Alzheimer's disease (*n* = 44) | | Non-clinical sample (*n* = 46) | | Differences between group | | | | | Cohen's d |
|---|---|---|---|---|---|---|---|---|---|---|
| | *M* | *SD* | *M* | *SD* | Levene's Test | Sig. | t | df | Sig. (two-tailed) | |
| FSIQ | 68.64 | 14.15 | 94.41 | 11.63 | 3.01 | .09 | 9.46 | 88 | .00 | 1.99 |
| VCI | 71.48 | 13.04 | 88.96 | 9.40 | 5.09 | .03 | 7.32 | 88 | .00 | 1.54 |
| SI | 4.32 | 2.53 | 7.57 | 1.87 | 4.58 | .04 | 6.94 | 88 | .00 | 1.46 |
| VC | 6.64 | 3.06 | 9.02 | 1.96 | 2.57 | .00 | 4.38 | 72.72 | .00 | 0.92 |
| IN | 4.20 | 1.76 | 7.48 | 2.32 | 6.76 | .01 | 7.52 | 88 | .00 | 1.58 |
| CO | 3.82 | 2.12 | 6.70 | 2.14 | 0.12 | .73 | 6.41 | 88 | .00 | 1.35 |
| PRI | 76.34 | 14.37 | 102.07 | 14.79 | 0.02 | .90 | 8.37 | 88 | .00 | 1.76 |
| BD | 6.82 | 3.19 | 10.50 | 2.69 | 1.43 | .24 | 5.93 | 88 | .00 | 1.25 |
| MR | 5.27 | 2.40 | 9.98 | 3.26 | 8.30 | .00 | 7.83 | 82.65 | .00 | 1.65 |
| VP | 5.82 | 2.86 | 10.74 | 3.53 | 4.67 | .03 | 7.24 | 88 | .00 | 1.52 |
| FW | 3.55 | 3.61 | 10.61 | 3.77 | 0.57 | .45 | 9.07 | 88 | .00 | 1.90 |
| PC | 3.84 | 2.34 | 6.72 | 2.58 | 0.19 | .67 | 5.53 | 88 | .00 | 1.16 |
| WMI | 76.73 | 13.05 | 97.00 | 12.87 | 0.16 | .69 | 7.42 | 88 | .00 | 1.56 |
| DS | 5.57 | 2.76 | 8.83 | 2.63 | 0.78 | .38 | 5.74 | 88 | .00 | 1.21 |
| AR | 6.23 | 2.50 | 10.15 | 2.58 | 0.35 | .56 | 7.33 | 88 | .00 | 1.54 |
| LN | 2.80 | 3.25 | 9.04 | 2.97 | 12.47 | .00 | 9.51 | 86.38 | .00 | 2.00 |
| PSI | 68.41 | 16.15 | 94.72 | 13.56 | 2.93 | .09 | 8.38 | 88 | .00 | 1.76 |
| SS | 4.70 | 3.36 | 9.22 | 2.58 | 7.26 | .01 | 7.12 | 8.69 | .00 | 1.50 |
| CD | 3.73 | 2.93 | 8.83 | 2.99 | .04 | .84 | 8.17 | 88 | .00 | 1.72 |
| CA | 1.25 | 2.16 | 7.72 | 3.64 | 11.26 | .00 | 1.30 | 73.69 | .00 | 2.16 |

Chapter 6

The scores of the non-clinical control group were significantly different to those achieved by the patients with schizophrenia for all indexes and subtests of the WAIS-IV-ID (Table 5). The non-clinical sample had a higher performance on all indexes and subtests. The effect size of FSIQ was large (2.75) for this comparison. The effect sizes of the indexes were large, ranging from 2.02 (PRI) to 2.63 (VCI). For the subtests, the effect sizes were also large, ranging from 1.45 (BD) to 2.75 (VC).

**Table 5: The WAIS-IV-ID performances of the patients with schizophrenia compared with the matched controls.**

| Tests | Patients with schizophrenia (*n* = 48) | | Non-clinical sample (*n* = 51) | | Differences between group | | | | | Cohen's d |
|---|---|---|---|---|---|---|---|---|---|---|
| | *M* | *SD* | *M* | *SD* | Levene's Test | Sig. | t | df | Sig. (two-tailed) | |
| FSIQ | 63.46 | 7.76 | 96.69 | 15.27 | 20.33 | .00 | 13.76 | 75.19 | .00 | 2.75 |
| VCI | 67.83 | 8.36 | 95.35 | 12.22 | 7.20 | .01 | 13.14 | 88.78 | .00 | 2.63 |
| SI | 3.88 | 1.90 | 8.25 | 2.20 | 1.42 | .24 | 10.58 | 97 | .00 | 2.12 |
| VC | 4.38 | 2.08 | 10.75 | 2.50 | 0.69 | .41 | 13.75 | 97 | .00 | 2.75 |
| IN | 4.98 | 1.26 | 8.59 | 2.83 | 38.61 | .00 | 8.27 | 70.07 | .00 | 1.65 |
| CO | 2.98 | 1.06 | 6.92 | 2.68 | 63.93 | .00 | 9.72 | 66.09 | .00 | 1.94 |
| PRI | 72.58 | 9.80 | 96.82 | 13.88 | 8.43 | .00 | 10.09 | 90.10 | .00 | 2.02 |
| BD | 6.06 | 1.92 | 9.49 | 2.74 | 8.69 | .00 | 7.25 | 89.76 | .00 | 1.45 |
| MR | 4.44 | 2.28 | 9.10 | 2.97 | 5.48 | .02 | 8.71 | 97 | .00 | 1.74 |
| VP | 5.42 | 1.82 | 9.90 | 2.62 | 8.91 | .00 | 9.93 | 89.33 | .00 | 1.99 |
| FW | 5.77 | 2.04 | 10.67 | 3.64 | 15.09 | .00 | 8.32 | 79.66 | .00 | 1.66 |
| PC | 3.33 | 1.98 | 7.49 | 2.98 | 5.20 | .02 | 8.13 | 97 | .00 | 1.63 |
| WMI | 70.33 | 10.35 | 100.35 | 17.95 | 16.02 | .00 | 10.26 | 80.83 | .00 | 2.05 |
| DS | 5.10 | 2.31 | 10.22 | 3.67 | 16.37 | .00 | 8.35 | 84.94 | .00 | 1.67 |
| AR | 4.46 | 1.83 | 9.98 | 3.74 | 25.70 | .00 | 9.40 | 73.63 | .00 | 1.88 |
| LN | 5.46 | 2.27 | 10.59 | 3.55 | 4.35 | .04 | 8.51 | 97 | .00 | 1.70 |
| PSI | 63.42 | 9.52 | 98.24 | 16.97 | 12.25 | .00 | 12.68 | 79.58 | .00 | 2.54 |
| SS | 3.27 | 1.81 | 9.49 | 3.26 | 9.00 | .00 | 11.83 | 79.04 | .00 | 2.37 |
| CD | 3.27 | 1.85 | 9.84 | 3.72 | 20.87 | .00 | 11.23 | 74.39 | .00 | 2.25 |
| CA | 3.58 | 2.01 | 7.96 | 2.88 | 2.12 | .15 | 8.71 | 97 | .00 | 1.74 |

The *t*-test values comparing non-clinical matched controls different scores for all indexes and subtests of the WAIS-IV-ID when compared with those of adults with an intellectual disability (Table 6). The non-clinical sample scored more highly on all indexes and subtests. For the clinical samples, the effect size of FSIQ was large (3.75). The effect sizes of the index scores were also large, ranging from 2.45 (PSI) to 4.31 (WMI). For the subtests, the effect sizes were large, ranging from 1.48 (PC) to 4.64 (LN).

**Table 6: The WAIS-IV-ID performances of the adults with intellectual disability compared with the matched controls.**

| Tests | Adults with an intellectual disability (*n* = 44) | | Non-clinical sample (*n* = 45) | | Differences between group | | | | | Cohen's d |
|---|---|---|---|---|---|---|---|---|---|---|
| | *M* | *SD* | *M* | *SD* | Levene's Test | Sig. | t | df | Sig. (two-tailed) | |
| FSIQ | 48.57 | 6.92 | 85.36 | 11.93 | 16.18 | .00 | 17.84 | 70.87 | .00 | 3.75 |
| VCI | 54.80 | 6.82 | 87.64 | 9.95 | 7.08 | .01 | 18.20 | 78.02 | .00 | 3.82 |
| SI | 1.98 | 1.55 | 7.16 | 1.76 | 0.09 | .77 | 14.74 | 87.00 | .00 | 3.10 |
| VC | 2.16 | 1.33 | 9.64 | 2.35 | 14.84 | .00 | 18.57 | 69.89 | .00 | 3.90 |
| IN | 2.66 | 1.26 | 6.56 | 2.12 | 13.50 | .00 | 10.59 | 71.88 | .00 | 2.22 |
| CO | 1.55 | .87 | 5.38 | 2.41 | 66.60 | .00 | 10.00 | 55.58 | .00 | 2.10 |
| PRI | 58.93 | 11.63 | 87.31 | 11.44 | 0.00 | .98 | 11.61 | 87.00 | .00 | 2.44 |
| BD | 3.00 | 2.72 | 7.67 | 2.07 | 2.80 | .10 | 9.13 | 87.00 | .00 | 1.92 |
| MR | 2.36 | 2.26 | 7.31 | 2.32 | 0.81 | .37 | 10.17 | 87.00 | .00 | 2.14 |
| VP | 3.50 | 1.76 | 8.58 | 2.50 | 2.51 | .12 | 11.06 | 87.00 | .00 | 2.32 |
| FW | 3.30 | 1.80 | 8.87 | 3.03 | 10.07 | .00 | 10.56 | 71.82 | .00 | 2.22 |
| PC | 2.07 | 1.73 | 5.98 | 3.27 | 14.75 | .00 | 7.07 | 67.16 | .00 | 1.48 |
| WMI | 52.70 | 3.81 | 89.98 | 11.57 | 39.97 | .00 | 20.51 | 53.62 | .00 | 4.31 |
| DS | 1.18 | .95 | 8.78 | 2.37 | 44.69 | .00 | 19.91 | 57.92 | .00 | 4.18 |
| AR | 1.82 | .99 | 7.69 | 2.38 | 29.67 | .00 | 15.23 | 59.17 | .00 | 3.20 |
| LN | 1.18 | .81 | 8.76 | 2.14 | 50.54 | .00 | 22.12 | 56.69 | .00 | 4.64 |
| PSI | 55.16 | 8.58 | 87.67 | 16.59 | 13.43 | .00 | 11.65 | 66.27 | .00 | 2.45 |
| SS | 1.86 | 1.76 | 7.67 | 3.61 | 12.49 | .00 | 9.68 | 64.16 | .00 | 2.03 |
| CD | 1.89 | 1.43 | 7.73 | 3.49 | 23.53 | .00 | 10.39 | 58.75 | .00 | 2.18 |
| CA | 1.39 | .95 | 7.71 | 2.91 | 35.24 | .00 | 13.84 | 53.36 | .00 | 2.91 |

The AUC value of all clinical samples was high for the FSIQ (.95; Table 7). For the indexes, the AUC ranged from .90 (PRI) to .94 (VCI), while for the subtests, it ranged from .83 (PC) to .94 (LN). The FSIQ was the best indicator for ruling-in non-clinical participants as clinical patients (LR+ = 13). FSIQ, VCI, VC, and CA were good clinical indicators, with sensitivity values over .79 and specificity values over .86.

**Table 7: The receiver-operating characteristic (ROC) analysis of all clinical samples.**

| Clinical samples | AUC | Youden index | Sensitivity | Specificity | Cut-off | LR+ | LR– |
|---|---|---|---|---|---|---|---|
| FSIQ | .95 | .74 | .80 | .94 | ≤ 71 | 13 | .21 |
| VCI | .94 | .76 | .85 | .91 | ≤ 76 | 9 | .16 |
| SI | .91 | .67 | .82 | .85 | ≤ 5 | 5 | .21 |
| VC | .91 | .70 | .83 | .87 | ≤ 7 | 6 | .20 |
| IN | .89 | .62 | .87 | .75 | ≤ 5 | 3 | .17 |
| CO | .87 | .64 | .88 | .76 | ≤ 4 | 4 | .16 |
| PRI | .90 | .64 | .82 | .82 | ≤ 81 | 5 | .22 |
| BD | .82 | .47 | .76 | .71 | ≤ 7 | 3 | .34 |
| MR | .88 | .60 | .85 | .75 | ≤ 6 | 3 | .20 |
| VP | .91 | .65 | .90 | .75 | ≤ 7 | 4 | .13 |
| FW | .89 | .62 | .86 | .76 | ≤ 7 | 4 | .18 |
| PC | .83 | .46 | .83 | .63 | ≤ 5 | 2 | .27 |
| WMI | .92 | .64 | .84 | .80 | ≤ 80 | 4 | .20 |
| DS | .89 | .54 | .87 | .67 | ≤ 7 | 3 | .19 |
| AR | .89 | .59 | .79 | .80 | ≤ 6 | 4 | .26 |
| LN | .94 | .73 | .91 | .82 | ≤ 7 | 5 | .11 |
| PSI | .93 | .71 | .82 | .89 | ≤ 74 | 7 | .20 |
| SS | .90 | .64 | .85 | .79 | ≤ 6 | 4 | .19 |
| CD | .92 | .66 | .84 | .82 | ≤ 5 | 5 | .20 |
| CA | .92 | .71 | .83 | .88 | ≤ 4 | 7 | .19 |

The AUC value for the patients with Alzheimer's disease was high for the FSIQ (.92; Table 8). The AUC of the indexes ranged from .86 (VCI and WMI) to .89 (PRI). For the subtests, the AUC ranged from .71 (VC) to .96 (LN). The subtest LN was the best predictor for ruling-out an individual as a patient with Alzheimer's disease (LR– = .09). PRI, WMI, SI, and LN were good clinical indicators, with both sensitivity scores over .79 and specificity values over .86.

**Table 8: The receiver-operating characteristic (ROC) analysis of the patients with Alzheimer's disease.**

| Patients with Alzheimer's disease (*n* = 44) | | AUC | Youden index | Sensitivity | Specificity | Cut-off | LR+ | LR– |
|---|---|---|---|---|---|---|---|---|
| FSIQ | | .92 | .65 | .91 | .74 | ≤ 86 | 4 | .12 |
| VCI | | .86 | .46 | .89 | .57 | ≤ 85 | 2 | .19 |
| | SI | .83 | .67 | .81 | .86 | ≤ 5 | 6 | .22 |
| | VC | .71 | .17 | .84 | .33 | ≤ 9 | 1 | .48 |
| | IN | .88 | .61 | .79 | .82 | ≤ 5 | 4 | .26 |
| | CO | .83 | .49 | .82 | .67 | ≤ 5 | 2 | .27 |
| PRI | | .89 | .67 | .81 | .86 | ≤ 79 | 6 | .22 |
| | BD | .81 | .47 | .82 | .65 | ≤ 9 | 2 | .28 |
| | MR | .87 | .52 | .82 | .70 | ≤ 7 | 3 | .26 |
| | VP | .86 | .56 | .80 | .76 | ≤ 7 | 3 | .27 |
| | FW | .88 | .57 | .79 | .78 | ≤ 7 | 4 | .27 |
| | PC | .79 | .32 | .86 | .46 | ≤ 6 | 2 | .30 |
| WMI | | .86 | .67 | .81 | .86 | ≤ 77 | 6 | .22 |
| | DS | .80 | .51 | .80 | .72 | ≤ 7 | 3 | .29 |
| | AR | .85 | .52 | .91 | .61 | ≤ 9 | 2 | .15 |
| | LN | .96 | .83 | .92 | .91 | ≤ 7 | 10 | .09 |
| PSI | | .88 | .58 | .84 | .74 | ≤ 84 | 3 | .22 |
| | SS | .85 | .54 | .82 | .72 | ≤ 7 | 3 | .25 |
| | CD | .88 | .58 | .91 | .67 | ≤ 7 | 3 | .13 |
| | CA | .89 | .72 | .88 | .84 | ≤ 5 | 5 | .15 |

For the individuals with schizophrenia, the AUC value for the FSIQ was high (.98; Table 9). The AUC of the indexes ranged from .93 (WMI) to .97 (VCI and PSI). For the subtests, the AUC ranged from .86 (BD) to .96 (SS). The PSI and VCI were the best indicators for ruling-in an individual as a patient with schizophrenia (LR+ = 11–15). FSIQ, PRI, WMI, VC, FW, and SS were the best indicators for ruling-out non-clinical individuals as patients with schizophrenia (LR– = .03–.09). FSIQ, VCI, PSI, VC, LN, SS, and CD were good clinical indicators, with sensitivity scores over .79 and specificity scores over .86.

**Table 9: The receiver-operating characteristic (ROC) analysis for the patients with schizophrenia.**

| Patients with schizo-phrenia (*n* = 48) | | AUC | Youden index | Sensitivity | Specificity | Cut-off | LR+ | LR– |
|---|---|---|---|---|---|---|---|---|
| FSIQ | | .98 | .84 | .96 | .88 | ≤ 75 | 8 | .05 |
| VCI | | .97 | .84 | .90 | .94 | ≤ 76 | 15 | .11 |
| | SI | .93 | .68 | .92 | .76 | ≤ 6 | 4 | .11 |
| | VC | .98 | .86 | .96 | .90 | ≤ 7 | 10 | .05 |
| | IN | .88 | .57 | .88 | .69 | ≤ 6 | 3 | .17 |
| | CO | .90 | .72 | .92 | .80 | ≤ 4 | 5 | .10 |
| PRI | | .92 | .72 | .98 | .75 | ≤ 88 | 4 | .03 |
| | BD | .86 | .60 | .83 | .76 | ≤ 7 | 4 | .22 |
| | MR | .90 | .64 | .85 | .78 | ≤ 6 | 4 | .19 |
| | VP | .92 | .72 | .92 | .80 | ≤ 7 | 5 | .10 |
| | FW | .87 | .66 | .94 | .73 | ≤ 8 | 3 | .09 |
| | PC | .87 | .64 | .85 | .78 | ≤ 5 | 4 | .19 |
| WMI | | .93 | .69 | .98 | .71 | ≤ 89 | 3 | .03 |
| | DS | .88 | .44 | .81 | .63 | ≤ 7 | 2 | .30 |
| | AR | .91 | .68 | .90 | .78 | ≤ 6 | 4 | .13 |
| | LN | .91 | .68 | .81 | .86 | ≤ 7 | 6 | .22 |
| PSI | | .97 | .80 | .88 | .92 | ≤ 74 | 11 | .14 |
| | SS | .96 | .80 | .94 | .86 | ≤ 6 | 7 | .07 |
| | CD | .95 | .76 | .88 | .88 | ≤ 5 | 7 | .14 |
| | CA | .90 | .62 | .81 | .80 | ≤ 5 | 4 | .23 |

For the individuals with an intellectual disability, the AUC value for FSIQ was high (.99; Table 10). The AUC of the indexes ranged from .95 (PRI) to .998 (WMI), while for the subtests, the AUC ranged from .86 (PC) to .997 (LN). FSIQ, VCI, WMI, PSI, SI, VC, IN, DS, AR, LN, and SS were the best indicators for ruling-in non-clinical individuals as an individual with an intellectual disability (LR+ = 13–∞). FSIQ, VCI, WMI, VC, DS, AR, LN, CD, and CA were the best indicators for ruling-out individuals as individuals with intellectual disability (LR– = .00 - .07). FSIQ, all indexes, and all subtests except CO and CD were good clinical indicators, with sensitivity scores over .79 and specificity scores over .86.

**Table 10: The receiver-operating characteristic (ROC) analysis for adults with an intellectual disability.**

| Adults with an intellectual disability ($n$ = 44) | | AUC | Youden index | Sensitivity | Specificity | Cut-off | LR+ | LR– |
|---|---|---|---|---|---|---|---|---|
| FSIQ | | .99 | .98 | .98 | 1.00 | ≤ 63 | ∞ | .02 |
| VCI | | .99 | .93 | .95 | .98 | ≤ 66 | 43 | .05 |
| | SI | .98 | .84 | .91 | .93 | ≤ 4 | 14 | .10 |
| | VC | .99 | .91 | .95 | .96 | ≤ 4 | 21 | .05 |
| | IN | .97 | .84 | .84 | 1.00 | ≤ 3 | ∞ | .16 |
| | CO | .93 | .66 | .84 | .82 | ≤ 2 | 5 | .20 |
| PRI | | .95 | .80 | .91 | .89 | ≤ 75 | 8 | .10 |
| | BD | .91 | .71 | .80 | .91 | ≤ 5 | 9 | .22 |
| | MR | .93 | .78 | .89 | .89 | ≤ 4 | 8 | .13 |
| | VP | .96 | .80 | .89 | .91 | ≤ 5 | 10 | .12 |
| | FW | .95 | .78 | .91 | .87 | ≤ 5 | 7 | .10 |
| | PC | .86 | .51 | .84 | .67 | ≤ 4 | 3 | .24 |
| WMI | | .998 | .96 | .98 | .98 | ≤ 60 | 44 | .02 |
| | DS | .99 | .96 | .98 | .98 | ≤ 3 | 44 | .02 |
| | AR | .99 | .91 | .93 | .98 | ≤ 3 | 42 | .07 |
| | LN | .997 | .98 | .98 | 1.00 | ≤ 3 | ∞ | .02 |
| PSI | | .96 | .86 | .91 | .96 | ≤ 65 | 20 | .10 |
| | SS | .93 | .80 | .86 | .93 | ≤ 3 | 13 | .15 |
| | CD | .94 | .78 | .93 | .84 | ≤ 4 | 6 | .08 |
| | CA | .98 | .89 | 1.00 | .89 | ≤ 4 | 9 | .00 |



## DISCUSSION

The present study aimed to examine the psychometric properties and clinical applicability of the Indonesian version of the WAIS-IV for assessing adults with Alzheimer's disease, schizophrenia, or an intellectual disability. Our results showed that the WAIS-IV-ID is reliable and has a good clinical utility. The three indexes (VCI, PRI and WMI) have excellent internal consistency for both the total clinical population and each separate clinical sample. For the WAIS-IV-ID subtests, our findings were similar to those reported in the WAIS-IV-US manual, which indicated that both PC and AR had the lowest reliability coefficients for identifying patients with Alzheimer's disease and adults with an intellectual disability, respectively (Wechsler, 2008b).

The analyses of group differences between the non-clinical and clinical populations indicated that the FSIQ, indexes, and subtest scores of the WAIS-IV-ID were sufficient to distinguish between both groups, with FSIQ, VCI, and VC being the best indicators of these differences, according to their large effect sizes. Furthermore, VCI and VC can be used to differentiate between patients with schizophrenia and their matched controls. The indexes PRI and PSI were effective for distinguishing between non-clinical individuals and those with Alzheimer's disease. WMI was adequate to identify individuals with an intellectual disability.

Further evidence of the good overall discriminative power of the WAIS-IV-ID in detecting clinical patients (individuals with Alzheimer's disease, schizophrenia, or an intellectual disability) came from the ROC analysis. The WAIS-IV-ID had the highest discriminative power for identifying the individuals with an intellectual disability, based on their test performances.

In the present study, we included a relatively small clinical sample; however, future research should attempt to recruit larger samples to facilitate a more detailed examination of the application of the WAIS-IV-ID in clinical populations. The reliability information of the SS, CD, and CA subtests of the PSI were not available for the present clinical groups as we did not perform test-retest studies. In addition, we did not compare the WAIS-IV-ID with an external criterion, such as another intelligence test or neuropsychological test. Providing stronger evidence of the validity of the WAIS-IV-ID for use with these or other clinical populations would be useful to give the clinician more confidence in utilizing the WAIS-IV-ID effectively in clinical practice.

# 7.

# External Validity of the Indonesian Wechsler Adult Intelligence Scale – Fourth Edition (WAIS–IV–ID)

**Published as**

Suwartono, C., Hidajat, L. L., Halim, M. S., Hendriks, M. P. H., and Kessels, R. P. C. (2016). External Validity of the Indonesian Wechsler Adult Intelligence Scale – Fourth Edition (WAIS-IV-ID). *Anima Indonesian Psychological Journal, 32*(1), 16–28.

## ABSTRACT

The Wechsler Adult Intelligence Scale – Fourth Edition (WAIS-IV) has been adapted for the Indonesian population (WAIS-IV-ID). The internal structure of the WAIS-IV-ID was shown to be valid and is comparable with the American version of the WAIS-IV (WAIS-IV-US; Suwartono, Hendriks, Hidajat, Halim, & Kessels, submitted); however, it is yet to be evaluated using external criteria. In this study, we investigated the external validity of the WAIS-IV-ID using other intelligence tests and educational achievements as the criteria for comparison. A total of 194 participants were recruited in total; however, not all participants were willing to complete all the tests. The results showed that the Full Scale IQ score (FSIQ) of the WAIS-IV-ID was significantly correlated with three other intelligence tests; Raven's Standard Progressive Matrices (SPM; *N* = 194), Cattell's Culture Fair Intelligence Test (CFIT; *n* = 134), and the Wechsler-Bellevue Intelligence Scale (WBIS; *n* = 44). There was also a significant positive correlation between the FSIQ and an educational achievement score, represented by a university grade point average (GPA; *n* = 51). The three of the four indexes of the WAIS-IV-ID (excluding the Verbal Comprehension Index) were significantly positively correlated with the SPM, CFIT, and WBIS scores. GPA was significantly correlated with the Processing Speed Index score, but not with the other indexes. These results demonstrate that the WAIS-IV-ID has sufficient external validity. Future research should examine the diagnostic utility of WAIS-IV-ID for groups with particular clinical conditions.



**KEYWORDS**

Academic achievement, CFIT, Indonesia, intelligence, SPM, WAIS-IV

The results of psychological assessments, particularly in the field of intelligence tests, serve as complementary information to an individual's developmental, social, educational, and occupational history, which can be used to comprehensively describe the test subject. The results help to estimate premorbid levels of cognitive functioning, formulate expectations of performance on other tests, and determine the level of discourse at which to engage an individual (Hiscock, 2007). The Indonesian version of the Wechsler Adult Intelligence Scale – Fourth Edition (WAIS-IV-ID) has been shown to have excellent structural validity, comparable with other internationally published standardized versions of the WAIS-IV (Suwartono, Halim, Hidajat, Hendriks, & Kessels, 2014); however, structural validity does not represent all aspects of a test's validity (Canivez, Konold, Collins, & Wilson, 2009), and it is important to establish that the WAIS-IV-ID has good external validity. In the present research, we validated the WAIS-IV-ID using external criteria, such as other intelligence tests (Raven's Standard Progressive Matrices (SPM; Raven, 2000), Cattell's Culture Fair Intelligence Test (CFIT; LPSP3 UI, 2009), and the Wechsler-Bellevue Intelligence Scale (WBIS; LPSP3 UI, n.d)), as well as a test of educational achievement (grade point average (GPA)).

Raven's SPM and Cattell's CFIT are among the few major intelligence tests currently available in Indonesia. Both tests have fewer verbal instructions than the Wechsler intelligence scales. Raven (2000) noted that the SPM test measures the deductive and reproductive ability of general cognitive ability (g factor). Deductive ability is the capacity to observe meaning in confusion and generate high-level schemata to facilitate the easy handling of complexity. Reproductive ability is the ability to absorb, recall, and reproduce information that has been made explicit and communicated from one person to another. The CFIT is assumed to be indifferent to cultural experience, which might differentially influence the test taker's responses to its items. This test measures fluid intelligence, a major measurable outcome of how biological factors influence the intellectual development, assumed to be unaffected by cultural factors (Schneider & McGrew, 2012). Even though individual differences might exist within a culture, there are not necessarily any differences in fluid intelligence between cultures (Nenty & Dinero, 1981). This makes intelligence measures such as the SPM and CFIT valuable in a multicultural country like Indonesia.

The predecessor of the Wechsler intelligence test, the WBIS, was developed by David Wechsler in 1939. The WBIS is still widely used for intelligence testing in Indonesia and is an individually administered measure of cognitive ability. The WBIS consists of 11 subtests; Information (I), Comprehension (C), Digit Span (D), Arithmetic (A), Similarities (S), Vocabulary (V), Picture Arrangement (PA), Picture Completion (PC), Block Design (BD), Object Assembly (OA), and Digit Symbol (DSym). The WBIS is divided into two parts, testing verbal and performance intelligence respectively (LSP3 FPUI, n.d), and provides a measurement of general intellectual functioning (Full Scale IQ (FSIQ)), a Verbal Scale (VS), and a Performance Scale (PS). The Verbal Scale comprises the scores of six subtests

(I, C, D, A, S, and V), while the Performance Scale includes the results of the other five subtests (PA, PC, BD, OA, and DSym). Internationally, the WBIS has been revised several times, generating the WAIS, the WAIS-R, the WAIS-III, and the most recently published version, the WAIS-IV (Wechsler, 2008a, 2008b). The WAIS-IV consists of 15 subtests, ten of which (Block Design (BD), Similarity (SI), Digit Span (DS), Matrix Reasoning (MR), Vocabulary (VC), Arithmetic (AR), Symbol Search (SS), Visual Puzzle (VP), Information (IN), Coding (CD)) are considered core subtests, while five (Letter Number Sequencing (LN), Figure Weights (FW), Comprehension (CO), Cancellation (CA), and Picture Completion (PC)) are supplemental (Wechsler, 2008). The WAIS-IV provides a measurement of general intellectual functioning (FSIQ), as well as four index scores that represent broad cognitive domains; Verbal Comprehension Index (VCI), Perceptual Reasoning Index (PRI), Working Memory Index (WMI), and Processing Speed Index (PSI). The index scales are composed of the core and supplemental subtests. The VCI scale comprises three core subtests (SI, VC, and IN) and one supplemental subtest (CO). The PRI scale includes three core subtests (BD, MR, and VP) and two supplemental subtests (FW and PC). The WMI scale comprises two core subtests (DS and AR) and one supplemental subtest (LN). The PSI scale consists of two core subtests (SS and CD) and one supplemental subtest (CA). We previously adapted the WAIS-IV into the Indonesian language (Suwartono et al., 2014), rearranging the item sequence in each subtest (except for the speed tests SS, CD, and CA) based on the index difficulty. The items included in the WAIS-IV-ID subtests are identical or equivalent to those of the WAIS-IV-US (Suwartono, Halim, Hidajat, Hendriks, & Kessels, 2014; Wechsler, 2008a). In this study, we investigated the correlations between the scores provided by two Wechsler intelligence scales, the WAIS-IV-ID and the WBIS.

Another external criterion by which intelligence tests can be validated is their ability to predict academic achievement (Naglieri & Bornstein, 2003; Parker & Benedict, 2002; Rohde & Thompson, 2007; Watkins, Lei, & Canivez, 2007). From a theoretical perspective, the construct of intelligence is expected to be related to academic achievement because learning itself is g-demanding (Jensen, 1998); therefore, we also conducted a correlation study between intelligence (WAIS-IV-ID) and a measure of educational achievement (GPA).

We hypothesized that there would be a significant positive correlation between the scores of the WAIS-IV-ID, the results of the other intelligence tests, and GPA. Moreover, we assumed that WAIS-IV-ID performance could contribute to the prediction of a participant's academic success and thus also serve to demonstrate the predictive validity of the WAIS-IV-ID.

## METHODS

### PARTICIPANTS

A convenience sampling method was used, based on the participant criteria listed in the WAIS-IV Technical and Interpretive Manual (Wechsler, 2008b). The present study involved 194 individuals in total, 66% of whom were women and 34% were men. Their ages ranged from 16 to 61 years (*M* = 23.53, *SD* = 7.75). The majority of the participants were university students (50.5%), although other occupations represented were high school students (25.8%), employees (10.8%), consultants (6.2%), housewives (2.1%), lecturers (1.5%), or "other" (3.1%). The participants were inhabitants of Jakarta (84.5%), Tangerang (6.2%), Bekasi (0.5%), and Denpasar (8.8%). This study was part of a larger study for the evaluation of the psychometric properties of the WAIS-IV-ID.

**Table 1: Demographic characteristics of the participants.**

| | SPM | CFIT | WBIS | GPA |
|---|---|---|---|---|
| Sample size | 194 | 134 | 44 | 51 |
| **Demographic data** | | | | |
| Men (%) | 34.02 | 32.09 | 20.45 | 11.76 |
| Women (%) | 65.98 | 67.91 | 79.55 | 88.24 |
| **Age (years old)** | | | | |
| Age range | 16 – 61 | 17 – 61 | 19 – 29 | 18 – 26 |
| *M* | 23.53 | 23.92 | 20.33 | 20.59 |
| *SD* | 7.75 | 6.89 | 2.87 | 1.72 |
| **Completed education (%)** | | | | |
| Junior high school | 26.80 | 0.75 | - | - |
| Senior high school | 42.78 | 58.21 | 86.36 | 100 |
| Undergraduate | 25.77 | 34.33 | 11.36 | |
| Master programme | 4.64 | 6.72 | 2.27 | |
| **Ethnicity (%)** | | | | |
| Balinese | 7.73 | 1.49 | 9.09 | 1.96 |
| Bataknese | 5.15 | 6.72 | 4.55 | 1.96 |
| Javanese | 17.53 | 22.39 | 15.91 | 15.69 |
| Tionghoa | 33.51 | 47.01 | 54.55 | 56.86 |
| Others* | 36.08 | 22.39 | 15.91 | 23.53 |

Note. SPM = Raven's Standard Progressive Matrices; CFIT = Cattell's Culture Fair Intelligence Test; WBIS = Wechsler-Bellevue Intelligence Scale; GPA = grade point average.
*Others = another ethnicity, including Sunda, Dayak, Minahasa, and many others. This category also contains participants who did not report their ethnicity.

The participants completed various tests of intelligence, although all participants undertook the WAIS-IV-ID. Table 1 presents the demographic characteristics of participants for whom data on both the WAIS-IV-ID and SPM (*N* = 194), the WAIS-IV-ID and CFIT (*n* = 134), the WAIS-IV-ID and WBIS (*n* = 44), and the WAIS-IV-ID and GPA (*n* = 51) were collected.

**INSTRUMENTS**

Two of Wechsler's scales were used; the WBIS (LSP3 FPUI, n.d) and the WAIS-IV-ID (Suwartono et al., 2014; Wechsler, 2008a). Both of these scales provide information about general intellectual functioning, known as the FSIQ. The WBIS is a predecessor of the WAIS-IV. The WBIS consists of 11 subtests divided into two parts, verbal (WB_VCI) and performance (WB_POI) and takes approximately 90–100 minutes to complete. The WAIS-IV-ID consists of 15 subtests divided into four factors; VCI, PRI, WMI, and PSI. Participants typically took 100–150 minutes to complete the WAIS-IV-ID. The items included in the WAIS-IV-ID subtests are identical or equivalent to those of the WAIS-IV-US. More information about the WAIS-IV can accessed at:

http://www.pearsonclinical.com/psychology/products/100000392/wechsler-adult-intelligence-scalefourth-edition-wais-iv.html#tab-details

Two other measurements of intelligence that require less verbal instruction were also used; Raven's SPM (Raven, 2000, 2008) and Cattell's CFIT (Cattell & Cattell, 1959, 1973; LSP3 FPUI, 2009). The SPM consists of 60 items presented in five sets of 12. The test has no time limit, but participants usually finish the SPM in about 25 minutes. It is relatively language-free (Raven, 2000), and the Cronbach's α is .84 (Suwartono, Amiseso, & Handoyo, 2017). CFIT form 3A was also used, which is designed to be a relatively true indicator of fluid intelligence. The CFIT a rapid test that requires about 30 minutes to administer. Its administration requires detailed verbal instructions (Colom & Abad, 2007; LSP3 FPUI, 2009), although the items of the CFIT are entirely nonverbal. The CFIT consists of four parts: Series, Analogies, Matrices, and Classification (Nenty & Dinero, 1981). Its Cronbach's α is .79 (LSP3 FPUI, 2009).

GPA was used to represent educational achievement. The GPA is a grading system employed in universities, with a score of 0 (lowest) to 4 (highest). The GPA data was obtained only for the psychology students at a private university.

**PROCEDURE**

This study is part of a more extensive study, involving the cooperation of local offices, consulting firms, non-governmental organizations, universities, and high schools. Potential participants were gathered in a meeting where the aims of the study were explained. The participants were then told about the goal of adapting the test for use in Indonesia. All participants gave their written informed consent before participating in the study. For participants who agreed to take other tests in addition to the WAIS-IV-ID, a counterbalancing method was used to eliminate the potential of test sequence bias. Those who wanted to participate again were contacted two weeks after the initial test. Each participant therefore took part in two or three test sessions of about two hours each. The WAIS-IV-ID was administered to all participants.

The WBIS and the WAIS-IV-ID are individually administered intelligence tests, but the SPM and the CFIT were administered in a group setting in a classroom or meeting room at the university. The WBIS and WAIS-IV-ID were administered according to the guidelines indicated in the administration manuals (LPSP3 FPUI, n.d.; Wechsler, 2008a). Participants were separated into groups who completed the SPM, CFIT, and WBIS. To ensure an appropriate test sequence, some of them took the WAIS-IV-ID first, while others started with another intelligence test first (SPM, CFIT, or WBIS). The administration of the WAIS-IV-ID and other intelligence tests (SPM, CFIT, and WBIS) were completed collected within three months.

**ANALYSES**

A Levene's test for the equality of variances was performed to determine whether there was an effect of test sequence between those participants who completed the WAIS-IV-ID first versus those who took the other intelligence tests (SPM and CFIT or WBIS) first. The t-values from t-tests for independent samples were used to identify any possible sequence effects for the WAIS-IV-ID test allocation. The potential sequence effects between those who took the WAIS-IV first versus those who did the SPM and CFIT first were also identified using an ABBA counterbalancing method. The descriptive statistics of the WAIS-IV-ID and all criteria for external validation study (SPM, CFIT, WBIS, and GPA) are also presented.

Pearson product-moment correlations were used for the external validation analysis, in which the FSIQ, index, and subtest scores of the WAIS-IV-ID were correlated with the total scores of the SPM, CFIT, WBIS, and GPA. The coefficient of determination ($r^2$) was then calculated to investigate the importance of the relationships and how each variable was affected by the others. The WBIS is the predecessor of the WAIS-IV-ID, so it was important to determine whether these tests yielded the same results; t-tests for dependent samples were used to compare the FSIQ scores of these two tests. Cohen's *d* was calculated following the method outlined by Lenhard and Lenhard (2016).

Pearson product-moment correlations were also used to investigate the relationship between intelligence and academic achievement. Two types of regression analyses were performed for the GPA. The first was a simple linear regression, in which GPA was estimated from the FSIQ of the WAIS-IV-ID. The second was a multiple linear regression, in which the GPA was estimated from the VCI, PRI, WMI, and PSI scores of the WAIS-IV-ID. A stepwise method was used with the following criteria: Probability-of-F-toenter ≤ .05, Probability-of-F-to-remove ≥ .10. In the regression analyses, the adjusted $R^2$ was determined. Then, developed the regression equation for transforming the standardized subtest scores into an estimate of the GPA score ($GPA_{Est}$). The adjusted $R^2$ is a modified version of $R^2$ that takes into account the predictors in the model and increases only if the new term improves the model more than would be expected by chance; therefore, the adjusted $R^2$ served as a measure of goodness of fit for this prediction model (Field, 2013).

## RESULTS

The Levene's test for equality of variances ($F$ = 3.03, $p$ =.09) between the group that took the WAIS-IV-ID before taking the other intelligence tests (SPM and CFIT) revealed an equal variance between these groups. The independent samples $t$-test was $t_{(58)} = -1.40$, $p$ = .17. The Levene's test for equality of variances ($F$ = 2.92, $p$ = .09) between the group that took the WAIS-IV-ID before completing the WBIS also revealed an equal variance between them. The independent samples t-test result was $t_{(50)} = -.02$, $p$ = .99. These statistical results indicate an equal variance between groups and suggest that the ABBA counterbalancing method was successful. The descriptive statistics of the WAIS-IV-ID and all criteria for the external validation study are presented in Table 2.

**Table 2: Descriptive statistics of the WAIS-IV-ID and the external criteria.**

| | FSIQ_WAIS-IV-ID | SPM | CFIT | FSIQ_WBIS | GPA |
|---|---|---|---|---|---|
| Sample size | 194 | 194 | 134 | 44 | 51 |
| M | 93.86 | 50.22 | 109.50 | 112.30 | 3.30 |
| SD | n/a | n/a | n/a | n/a | .32 |
| Minimum | 40 | 36 | 73 | 95 | 2.6 |
| Maximum | 133 | 65 | 140 | 129 | 3.88 |

n/a = The *SD* cannot be displayed due to the licensing regulation.

In Table 3, we present the results of the correlation analysis and coefficients of determination between the WAIS-IV-ID scores (FSIQ, four indexes, and 15 subtests) and the external validity criteria (other measures of intelligence and GPA). Weak but significant correlations were found between the SPM and the WAIS-IV-ID scores (.19 to .32), and significant weak-to-moderate correlations were found between the CFIT and the WAIS-IV-ID scores (.19 to .54). Weak-to-moderate significant correlations were found between the WBIS scores and the FSIQ, PRI, WMI, IN, BD, MR, FW, DS, and AR scores of the WAIS-IV-ID (.30 to .53). Furthermore, we found weak-to-moderate significant correlations between GPA and the FSIQ, PSI, MR, SS, and CD WAIS-IV-ID scores (.28 to .48).

**Table 3: Correlation coefficients between WAIS-IV-ID, the other intelligence tests, and academic achievement.**

| WAIS-IV-ID | SPM (N = 194) | $r^2$ | CFIT (n = 134) | $r^2$ | WBIS (n = 44) | $r^2$ | GPA (n = 51) | $r^2$ |
|---|---|---|---|---|---|---|---|---|
| **FSIQ** | .28** | .08 | .54** | .29 | .53** | .28 | .38** | .14 |
| **VCI** | .22** | .05 | .34** | .12 | .25 | .06 | .14 | .02 |
| SI | .19** | .04 | .31** | .10 | .14 | .02 | -.03 | .00 |
| VC | .20** | .04 | .27** | .07 | .04 | .00 | .20 | .04 |
| IN | .24** | .06 | .19* | .04 | .39** | .15 | .13 | .02 |
| CO | .24** | .06 | .28** | .08 | -.02 | .00 | .08 | .01 |
| **PRI** | .28** | .08 | .33** | .11 | .45** | .20 | .18 | .03 |
| BD | .29** | .08 | .26** | .07 | .30* | .09 | .14 | .02 |
| MR | .21** | .04 | .30** | .09 | .34* | .12 | .28* | .08 |
| VP | .28** | .08 | .23** | .05 | .27 | .07 | -.03 | .00 |
| FW | .32** | .10 | .43** | .18 | .41** | .17 | .16 | .03 |
| PC | .26** | .07 | .31** | .10 | .19 | .04 | -.10 | .01 |
| **WMI** | .29** | .08 | .47** | .22 | .49** | .24 | .18 | .03 |
| DS | .25** | .06 | .32** | .10 | .40** | .16 | .10 | .01 |
| AR | .30** | .09 | .48** | .23 | .49** | .24 | .20 | .04 |
| LN | .22** | .05 | .24** | .06 | .22 | .05 | .01 | .00 |
| **PSI** | .24** | .06 | .40** | .16 | .15 | .02 | .48** | .23 |
| SS | .22** | .05 | .37** | .14 | .12 | .01 | .42** | .18 |
| CD | .23** | .05 | .31** | .10 | .16 | .03 | .38** | .14 |
| CA | .27** | .07 | .36** | .13 | .17 | .03 | .15 | .02 |

** Correlation is significant at the .01 level (two-tailed).
* Correlation is significant at the .05 level (two-tailed).

The WAIS-IV-ID FSIQ score was significantly correlated with that of the SPM (r(192) =28, $r^2$ = .08, p < .01). Of the WAIS-IV-ID indexes, WMI had the highest-magnitude correlation with the SPM score ($r(192) = .29$, $r^2 = .08$, $p < .01$). The subtests with the highest-magnitude and most significant correlation for each index of the WAIS-IV-ID were IN and CO ($r_{(192)}$ =24, $r^2$= .06, p < .01), FW ($r_{(192)} = .32$, $r^2 = .10$, $p < .01$), AR ($r_{(192)} = .30$, $r^2 = .09$, $p < .01$), and CA ($r_{(192)} = .27$, $r^2 = .07$, $p < .01$).

The WAIS-IV-ID FSIQ score was significantly correlated with that of the CFIT (r(132) = .54, $r^2$ = .29, p < .01). As for the CFIT score, WMI had the highest-magnitude correlation with the CFIT score ($r(132) = .47$, $r^2 = .22$, p < .01). The subtests with the highest-magnitude and most significant correlations for each index of the WAIS-IV-ID were SI ($r_{(132)} = .31$, $r^2 = .10$, $p < .01$), FW ($r_{(132)} = .43$, $r^2 = .18$, $p < .01$), AR ($r_{(132)} = .48$, $r^2 = .23$, $p < .01$), and SS ($r_{(132)} = .37$, $r^2 = .14$, p < .01).

Chapter 7

The WAIS-IV-ID FSIQ score was significantly correlated with that of the WBIS ($r_{(42)} = .53$, $r^2 = .28$, $p < .01$). The index with the highest-magnitude correlation with the WBIS score was WMI ($r_{(42)} = .49$, $r^2 = .24$, $p < .01$). The subtests with the highest-magnitude and most significant subtest for each factor of the WAIS-IV-ID were IN ($r_{(42)} = .39$, $r^2 = .15$, $p < .01$), FW ($r_{(42)} = .41$, $r^2 = .17$, $p < .01$), AR ($r_{(42)} = .49$, $r^2 = .24$, $p < .01$). No significant results were detected between the WBIS score and those of the PSI subtests.

We present the details about the correlations between the indexes and subtests of the WAIS-IV-ID and the verbal subtests of the WBIS in Table 4.

**Table 4: Correlations between the WAIS-IV-ID and the verbal subtests of the WBIS.**

| | WB_ FSIQ | WB_ VCI | WB_I | WB_ CO | WB_D | WB_A | WB_S | WB_V |
|---|---|---|---|---|---|---|---|---|
| **FSIQ_ WAIS4** | .53** | .55** | .32 | -.002 | .53** | .60** | -.04 | .31* |
| **VCI** | .25 | .36* | .47** | -.06 | .22 | .36* | .06 | .48** |
| SI | .14 | .24 | .14 | -.003 | .27 | .28 | -.08 | .42** |
| VC | .04 | .06 | .24 | -.06 | -.05 | .03 | .08 | .33* |
| IN | .39** | .51** | .70** | -.08 | .25 | .51** | .17 | .34* |
| CO | -.02 | .06 | .21 | .17 | .03 | .01 | -.26 | .25 |
| **PRI** | .45** | .34* | .23 | -.01 | .23 | .49** | -.04 | .06 |
| BD | .30* | .16 | .17 | -.07 | .17 | .32* | -.24 | -.01 |
| MR | .34* | .35* | .30* | -.09 | .25 | .39** | .14 | .10 |
| VP | .27 | .16 | -.01 | .11 | .02 | .30 | -.01 | .04 |
| FW | .41** | .34* | .14 | .05 | .41** | .22 | .03 | .28 |
| PCm | .19 | .03 | -.21 | .09 | .27 | -.23 | .10 | .04 |
| **WMI** | .49** | .62** | .20 | .15 | .66** | .59** | -.05 | .26 |
| DS | .40** | .56** | .12 | .26 | .74** | .38** | -.10 | .24 |
| AR | .49** | .52** | .25 | -.03 | .40** | .71** | .02 | .24 |
| LN | .22 | .38* | .14 | .02 | .50** | .30 | .004 | .24 |
| **PSI** | .15 | .12 | -.01 | -.10 | .26 | .08 | -.01 | .07 |
| SS | .12 | .10 | -.03 | -.08 | .30 | .01 | -.04 | .03 |
| CD | .16 | .11 | .02 | -.09 | .16 | .11 | .03 | .08 |
| CA | .17 | .17 | .09 | .09 | .24 | -.01 | .03 | -.05 |

** Correlation is significant at the .01 level (two-tailed).
* Correlation is significant at the .05 level (two-tailed).

In Table 5, we present the details about the correlations between the indexes and subtests of the WAIS-IV-ID and the corresponding participant performances for the subtests of the WBIS.

**Table 5: Correlations between the WAIS-IV-ID and the performance subtests of the WBIS.**

| | FSIQ_WB | WB_POI | WB_PA | WB_PC | WB_B | WB_OA | WB_Dsym |
|---|---|---|---|---|---|---|---|
| **FSIQ_ WAIS4** | .53** | .20 | .05 | .18 | .34* | -.09 | .22 |
| **VCI** | .25 | -.02 | -.07 | .22 | -.10 | -.05 | .06 |
| SI | .14 | -.02 | -.09 | .15 | -.17 | .01 | .05 |
| VC | .04 | -.01 | -.04 | .20 | -.11 | .06 | .01 |
| IN | .39** | .01 | -.01 | .15 | .05 | -.19 | .09 |
| CO | -.02 | -.13 | -.26 | .19 | -.03 | -.05 | -.12 |
| **PRI** | .45** | .36* | .09 | .28 | .52** | .004 | .10 |
| BD | .30* | .33* | .01 | .06 | .62** | .16 | .04 |
| MR | .34* | .17 | .09 | .26 | .10 | -.16 | .18 |
| VP | .27 | .26 | .09 | .25 | .38* | -.01 | -.001 |
| FW | .41** | .25 | .08 | .14 | .36* | -.09 | .23 |
| PCm | .19 | .29 | -.15 | .22 | .36* | .29 | .17 |
| **WMI** | .49** | .07 | .04 | .05 | .29 | -.31* | .09 |
| DS | .40** | -.01 | -.12 | -.02 | .27 | -.23 | .06 |
| AR | .49** | .16 | .24 | .14 | .23 | -.33* | .11 |
| LN | .22 | -.07 | -.19 | -.002 | .24 | -.32* | .001 |
| **PSI** | .15 | .06 | .04 | -.09 | .07 | .16 | .34* |
| SS | .12 | .02 | -.13 | -.04 | .23 | .15 | .10 |
| CD | .16 | .09 | .20 | -.12 | -.10 | .13 | .49** |
| CA | .17 | .08 | -.18 | .01 | .06 | .25 | .25 |

** Correlation is significant at the .01 level (two-tailed).
* Correlation is significant at the .05 level (two-tailed).

We also compared each individual's scores for the WAIS-IV-ID and the WBIS using a t-test for dependent samples. Significant differences were detected between the scores obtained for the WAIS-IV-ID ($M$ = 103.68) and those of the WBIS ($M$ = 112.30); $t_{(43)} = 7.41$, $p < .01$, $d = 1.09$. The results of the WBIS were higher than for the WAIS-IV; however, these results must be carefully interpreted because of the major lack of updated information on the norms and items of the WBIS.

The WAIS-IV FSIQ score was significantly correlated with an individual's GPA ($r(49) = .38$, $r^2 = .14$, $p < .01$; Table 3). Among the four WAIS-IV-ID indexes, only PSI was significantly correlated with GPA ($r_{(49)} = .48$, $r^2$ = . 0.23, $p < .01$). The PSI subtests that were significantly correlated with the GPA were MR ($r_{(49)} = .28$, $r^2 = .08$, $p < .05$), SS ($r_{(49)} = .42$, $r^2 = .18$, $p < .01$), and CD ($r_{(49)} = .38$, $r^2 = .14$, $p < .01$). In Table 6, the relationships between the GPA and the other intelligence tests are presented, revealing no significant correlations between GPA and SPM or CFIT.

**Table 6: Correlations between the GPA and the results of the intelligence tests.**

| | SPM (n = 37) | CFIT (n = 35) | FSIQ_ WAIS-IV-ID (n = 51) | VCI (n = 51) | PRI (n = 51) | WMI (n = 51) | PSI (n = 51) |
|---|---|---|---|---|---|---|---|
| **GPA** | -.06 | .15 | .38** | .14 | .18 | .18 | .48** |

** Correlation is significant at the .01 level (two-tailed).
* Correlation is significant at the .05 level (two-tailed).

Regarding the predictive validity of the WAIS-IV-ID, our analysis showed that the FSIQ score is predictive of the GPA (adjusted R2 = .13, F(1,49) = 8.22, p < .01). The formula is GPAEst = 1.71 + 0.02 FSIQ. We also estimated the GPA using the four indexes of the WAIS-IV-ID (VCI, PRI, WMI, and PSI) as predictors in the stepwise method, obtaining an adjusted R2 of .21 (F (1,49) = 14.31, p < .01). PSI is the only index of the WAIS-IV-ID that can predict the GPA. The formula is GPAEst = 2.06 + 0.01 PSI.

## DISCUSSION

The results of the present study provide further evidence for the validity of the WAIS-IV-ID. The FSIQ, indexes, and all subtests of the WAIS-IV-ID had positive and significant correlations with the scores obtained using Raven's SPM and Cattell's CFIT. Moderate correlations were detected between the FSIQ of the WAIS-IV-ID and the other measures of intelligence. Those results were expected because the WAIS-IV-ID covers four broad cognitive areas; verbal comprehension, perceptual reasoning, working memory, and processing speed. In contrast, the SPM and CFIT only cover the nonverbal (reasoning) cognitive abilities (Cattell & Cattell, 1973; Nenty & Dinero, 1981; Raven, 2000).

We found that the VCI of the WAIS-IV-ID was moderately correlated with CFIT, but only had a weak, but still significant, relationship with SPM. Such results may stem from the fact that SPM and CFIT put less emphasis on verbal content, as both tests heavily focus on measuring fluid intelligence (Nenty & Dinero, 1981; Raven, 2000). Fluid intelligence is measured in the WAIS-IV-ID by the PRI factor; however, WAIS-IV-ID also measures VC, WM, and PS. During our experience in the administration and scoring process of the WBIS, we found that many of the items included on the test are outdated; participants got a score of 0 not only when they did not have the ability to answer the questions, but also because the content of the questions is no longer applicable to modern Indonesians. Some of the words are no longer used in everyday language.

WMI and its subtests were significantly positively correlated with SPM and CFIT. This result can be explained by the findings of Ackerman, Beier, and Boyle (2002) also Kanerva and Kalakoski (2016), who reported that working memory was related to general reasoning. Moreover, Kyllnonen and Christal (1990) and Tourva, Spanoudis, and Demetriou (2016) found that WM performance is positively and significantly related to tasks involving reasoning or fluid intelligence.

The SPM and CFIT scores were also significantly positively correlated with PSI and its subtests, which supports the findings of Jensen (1980), Tillman, Bohlin, Sorensen, and Lundervold (2009), and Vernon (1983). Jensen (1980) reported that the speed or efficiency of neural transmissions in the brain affect an individual's performance for elementary cognitive tasks as well as reasoning tasks. PSI therefore contributes to the performance of intellectual activities (Tillman et al., 2009), and is considered a bridge between working memory and general cognitive ability (Vernon, 1983).

The FSIQ score of the WAIS-IV-ID was significantly positively correlated with the WBIS. The moderate strength of this correlation (shared variance: 28%) indicated that the relationship might be explained by the reasoning aspects measured in both the WAIS-IV-ID and the WBIS. We identified significant differences in the mean FSIQ scores obtained by individuals taking the WBIS and WAIS-IV-ID assessments; the WBIS (M = 112.30) yielded higher scores than the WAIS-IV-ID (M = 103.68). This condition was also reported by Hiscock (2007), who found that the WAIS yielded higher scores than the new test, WAIS-R. Our results suggest that intelligence testing using the WBIS would result in an individual achieving a higher IQ result than if they were tested using the WAIS-IV-ID. The major concern for the use of WBIS in Indonesia is that the test is not standardized and normalized for the Indonesian population. This may likely be the reason for the higher WBIS scores observed in our study.

The WBIS is still the most used intelligence test in Indonesia. While most of the participants correctly answered all of the subtests in the WBIS' performance index in time, we suggest that the continued usage of WBIS in Indonesia is not recommended. We observed that the time limit of the WBIS subtests was too lenient, as most participants answered correctly within the time limit; for example, in the DSym subtest of the WBIS, the time limit is 90 seconds and the maximum score is 67. In our study, the time taken by the participants was 58–90 s (M = 84.68, SD = 7.12), while the range of scores achieved by the participants was 34–67 (M = 63.66, SD = 6.18). DSym is a speed test and the time limit is made so short that no one can finish all the items. Nevertheless, more than half of all participants (52.3%) got the maximum score.

Over time, certain cultural changes might have a favorable effect on intelligence; people strive for better conditions and increased standards of living, such as moving from working-class to middle-class homes. This also leads to a better environment, smaller family sizes, improved health, better nutrition, and improvements in education (Hiscock, 2007; Nisbett, Aronson, Blair, Dickens, Flynn, Halpern, & Turkheimer, 2012; Rindermann, Becker, & Coyle, 2017; Williams, 1998). In Indonesia, the improvement in education was measured as an increased rate of those who completed junior high school (from 14.51% to 19%) and a decreased illiteracy rate in adolescents (from 14.84% to 10.21%) between 1994 and 2016 (Badan Pusat Statistik Indonesia, 2016). Because of their higher education they are exposed to various opportunities, knowledge, and information in various places and media. The WBIS could, therefore, estimate a higher IQ score than the WAIS-IV-ID.

The present research represents the first time that individuals have been tested using the WAIS-IV-ID however, and further research over time is needed to confirm this hypothesis.

The Flynn effect is the theory which emphasizes that the average IQ score will increase in subsequent generations (Hiscock, 2007). We did not have access to past scores for the intelligence tests used here, so the Flynn effect cannot be investigated. A comparison of differently aged participants in the present study was also not possible, as the majority of them were of the younger generation. Arguably, individuals likely scored more highly on the old test (WBIS) because it was targeted to previous generations; modern-day participants have an increased processing speed and increased access to sources of information, so they scored better in the WBIS than previous generations.

Naglieri and Bornstein (2003) found that cognitive tests that measure basic psychological processes have considerable validity in the prediction of academic achievement. The present findings established the external validity of the WAIS-IV-ID as an intelligence test through its significant correlations with other intelligence tests and educational achievement, represented by GPA, supporting the findings of Pluck, Ruales-Chieruzzi, Paucar-Guerra, Andrade-Guimaraes, and Trueba (2016). The FSIQ, PSI, MR, SS, and CD scores of the WAIS-IV-ID were significantly positively correlated with GPA. In line with the findings of Parker and Benedict (2002), who reported that FSIQ is predictive of the IQ-achievement correlation. We found FSIQ contributed 14% of the variance in GPA. Moreover, the construct of intelligence is expected to precede and influence the development of academic achievement, since learning itself is g-demanding (Jensen, 1998). Of the WAIS-IV-ID indexes, only PSI had a positive and significant correlation with GPA. PSI reflects the mental and motor speed of an individual when solving nonverbal problems, and requires a person to plan, organize, and develop relevant strategies. This is also reflected in its subtests; for example, MR measures nonverbal abstract reasoning abilities and visual information processing, while SS and CD require the capacity to absorb, integrate, and respond to information, as well as hand-eye coordination, attention, and the capacity to work under pressure (Groth-Marnat, 2009). These abilities are vital for a university-level education, which may explain the correlations between PSI and GPA. PSI could also be useful when students need to prioritize and develop strategies to allocate their time for class, assignments, exams, and even extracurricular activities; therefore, a high PSI score would reflect a student more able to achieve higher grades at university (GPA).

Validity is not determined by a single validation study, but by a body of research that demonstrates the relationship between the test and the behavior it is intended to measure. Brown (2010) described validity as a unitary factor known as construct validity, which consists of five sources of evidence; test content, response processes, internal structure, relations to other variables, and consequences of testing. Future validation research should examine the diagnostic utility of the WAIS-IV with special (clinical) groups, which could include individuals identified as intellectually gifted, as well as those who have an intellectual disability, mild cognitive impairment, or patients with Alzheimer's disease. The ability to distinguish differences between clinical and non-clinical groups is one way of revealing whether the WAIS-IV-ID could be useful as a diagnostic test.

The present research has some limitations. First, we used convenience sampling, as the study was limited to the islands of Java and Bali. Further research is necessary to assess the generalization of these findings to people from other geographical regions in Indonesia. Another limitation was that not all of the participants wanted to participate in multiple tests, despite being asked to allocate time for three test sessions lasting approximately two hours each. The duration and sometimes the locations of the assessments were the main challenges for participants.

Taking these limitations and concerns into account, this research provides preliminary evidence that the WAIS-IV-ID is a valid instrument because its results were significantly correlated with those of other intelligence tests. Moreover, the WAIS-IV-ID could predict future achievement in university (GPA). We therefore conclude that the WAIS-IV-ID is an externally valid assessment of individual intelligence.

# 8.
# Summary and Discussion

The Wechsler intelligence scales have been frequently revised to fulfill the needs of modern assessment criteria. Of these, only the WBIS and WAIS assessments have been translated for use in Indonesia thus far. Unfortunately, the psychometric properties of the WBIS and WAIS assessments are rarely studied, and their outdated written manuals lack information about their psychometric properties. To the best of our knowledge, these properties have not been evaluated for the WBIS and WAIS assessments in Indonesia, nor for any other psychological test, as no data could be found in the literature (Suwartono & Santoso, 2016). We therefore adapted the WAIS-IV into the Indonesian language (WAIS-IV-ID) between 2012 and 2015. The following sections summarize this process, and the subsequent psychometric evaluation of the WAIS-IV-ID. The WAIS-IV-ID was administered to a total of 1,896 healthy participants in Indonesia, representing different ethnic groups. It was also administered to 195 Indonesian participants with Alzheimer's disease, schizophrenia, and intellectual disability.

## ADAPTATION

In **Chapter 2**, we described the process of adapting WAIS-IV for the Indonesian language and culture, ensuring that all ethnic groups could accomplish the subtests. A major adaptation was the sequential rearrangement of the items in most of the WAIS-IV-ID subtests. This research involved 148 healthy participants, who were representative of the overall Indonesian population with respect to gender, age, educational levels, and ethnic background. We reordered the item sequences to best assess the examinee's maximum potential.

## RELIABILITY

In **Chapter 4**, we described the reliability of the WAIS-IV-ID, including its subtests and indexes, by assessing its internal consistency, by performing a test-retest method, and by determining the inter-rater reliability for use with non-clinical participants. We also investigated the reliability of the WAIS-IV-ID when used with clinical participants in **Chapter 6**.

### INTERNAL CONSISTENCY

The reliability estimation using the internal consistency method involved determining Cronbach's α for all subtests (except SS, CD, and CA) of the WAIS-IV-ID. Then, for the four indexes and FSIQ, we used the coefficient α, calculated as the function of the reliabilities of the components and their intercorrelations based on the formula outlined by Nunnally and Bernstein (1994), as well as calculating the coefficient ω (Gignac, 2014).

The estimation of the internal consistency reliabilities of the WAIS-IV-ID using the coefficient α method proved to be reliable for the non-clinical sample. The reliabilities of the FSIQ, index, and subtest scores were all high. In contrast, the results were found to be unreliable if the reliabilities of the indexes were estimated using the coefficient ω. These lower estimations of reliability were evidence that intelligence assessed by the WAIS-IV-ID is a multidimensional construct. These results implied that psychologists must be careful when interpreting the index scores of the WAIS-IV-ID; however, we found the FSIQ to be reliable based on the coefficient ω method. The level of reliability associated with the FSIQ assessed using both coefficient α and ω methods is therefore likely sufficient for interpretation.

We applied the coefficient α method to estimate the reliability of the WAIS-IV-ID in clinical samples. The reliabilities of the indexes and subtests were found to be high when the WAIS-IV-ID was used to assess individuals with Alzheimer's disease, schizophrenia, or an intellectual disability; however, unreliable to moderate results were found for the AR, IN, and CO subtests in clinical samples of individuals with schizophrenia and intellectual disability. The results of the coefficient α method of estimating internal consistency reliabilities suggest that the indexes and subtests of the WAIS-IV-ID were sufficiently reliable for use in assessing the clinical samples studied here, except for the PSI factors and subtests, which are still an open question.

### TEST–RETEST RELIABILITY

We performed a test-retest correlation analysis to investigate the stability of the WAIS-IV-ID scores over time (Cook & Beckman, 2006; Miller, 1995). This approach was only applied in the non-clinical sample. The test-retest correlations were calculated for assessments conducted with a 10–65-day interval ($M$ = 26.66, $SD$ = 13.54), revealing a significant correlation between the scores an individual achieved on their first and second tests. This indicates that the index and subtest scores are stable across repeated administrations of the WAIS-IV-ID and leads us to conclude that the WAIS-IV-ID is a reliable instrument.

### INTER–SCORER RELIABILITY

Finally, we studied the inter-scorer reliability to determine whether the scoring guidelines for the verbal subtests are clear and could be homogenously interpreted across scorers to produce a high degree of agreement for the scoring of WAIS-IV-ID answers. The intra-class correlation coefficients for all verbal subtests were excellent, ranging from 0.94 for CO to 0.98 for IN. The results indicated that all verbal subtests could be scored reliably using the test administration guidelines provided for the WAIS-IV-ID.

## VALIDITY

A test score on a scale such as the WAIS-IV-ID is only meaningful if it reflects intelligence as a psychological construct. We confirmed the internal structure of the WAIS-IV-ID in **Chapter 3**. We also investigated the external validity of the scale in **Chapter 7**, using two other intelligence tests and educational achievement as criteria for comparison. Finally, in **Chapter 6**, we investigated the performances of the WAIS-IV-ID as a diagnostic test to differentiate between three clinical conditions; Alzheimer's disease, schizophrenia, and an intellectual disability.

**INTERNAL STRUCTURE**

The four-factor short-form models of the WAIS-IV-ID showed an adequate fit using a large sample of Indonesian participants. We found that a model containing Verbal Comprehension, Perceptual Reasoning, Working Memory, and Processing Speed factors, with a joint loading of AR on Perceptual Reasoning as well as Working Memory, was the four-factor model that best represented the internal structure of the WAIS-IV-ID. We also compared the WAIS-IV-ID with the CHC framework, which is represented by a five-factor model with AR loading on *Gsm* and *Gf*. The goodness-of-fit indexes showed that the four- and five-factor models both met the criteria for an adequate fit; however, the four-factor structure (CFI, SRMR, AGFI, and AIC model) was favorable to the other models, and. The four-factors structure was confirmed to describe the latent structure of the WAIS-IV-ID.

**CONCURRENT VALIDITY**

We studied the concurrent validity of the WAIS-IV-ID by investigating the relationships between its scores and several external criteria, including other measures of intellectual ability and educational achievement, among contrasting groups (non-clinical vs. clinical). The results show a weak but significant positive correlation between the FSIQ, indexes, and subtest scores and those of Raven's Standard Progressive Matrices (SPM), as well as significant but weak-to-moderate positive correlations between the WAIS-IV-ID scores and the Culture Fair Intelligence Test (CFIT). The weak-to-moderate strengths of these correlations were expected because WAIS-IV-ID measures more varied factors than SPM and CFIT, which both put less emphasis on verbal content, instead heavily assessing measures of fluid intelligence. Further, the WAIS-IV FSIQ score was significantly correlated with educational achievement (measured as GPA), and it could predict future achievement in university (GPA). These results demonstrate the validity of the WAIS-IV-ID by its positive and significant correlations with external criteria, such as other nonverbal intelligence tests and academic achievement.

Various validation studies determined test validity. Test-criterion relationships were determined during the assessment of special (clinical) group studies, providing further evidence of WAIS-IV-ID validity. We studied several clinical groups, including individuals with Alzheimer's disease, schizophrenia, and intelligence disability. Statistically significant differences were detected between all WAIS-IV-ID FSIQ, indexes, and subtest scores of the non-clinical and combined clinical samples, and the same results were obtained when the different clinical groups were assessed individually. These results suggest that the WAIS-IV-ID can differentiate between non-clinical and clinical individuals, including those with Alzheimer's disease, schizophrenia, and intelligence disability. This concurrent validation method using contrasted groups (non-clinical vs. clinical group) as criteria therefore further demonstrated the validity of WAIS-IV-ID.

## SHORT FORMS OF THE WAIS–IV–ID

Participants often have limited time available, and researchers and psychologists in clinical practice could often not complete all 15 subtests of the WAIS-IV-ID. This encouraged us to determine the best short forms of the WAIS-IV-ID that would still facilitate an accurate estimate of FSIQ. In **Chapter 5**, we explored 13 short-form options, developed based on previous research, an applied stepwise regression analysis, and the magnitude of the independent correlations of the subtests with the FSIQ. We found that the goodness of fit of our prediction models to estimate the FSIQ ranged from an adjusted R2 of .60 (for the short form comprising IN and PC) to an adjusted R2 of .94 (for the short form consisting of SI, IN, BD, PC, DS, AR, and CO). We checked the reliability and structural validity of the short forms, and concluded that the best short-form model of WAIS-IV-ID was SF5, which comprised the following factors: IN, MR, AR, and CO.

## STRENGTHS, LIMITATIONS, AND FUTURE DIRECTIONS

### STRENGTHS

The development and psychometric evaluation of the WAIS-IV-ID was labor-intensive and time-consuming; however, these investments were worthwhile because they have led to the first standardized intelligence test adapted for use in Indonesia. The strengths of this adaptation and psychometric evaluation are the geographical areas covered (almost all of the big islands of Indonesia, except Papua), the number of participants included (2,091 participants), and the range of the participant ages (16–90 years old).

The process of developing the WAIS-IV-ID was not simply to translate the test into the Indonesian language, unlike most of the other intelligence tests currently available in Indonesia. The psychometric adaptation of the WAIS-IV for the Indonesian population enhanced its reliability and validity, and we are convinced that the WAIS-IV-ID can capture a person's general cognitive ability and intelligence profiles for specific cognitive abilities, both for non-clinical and clinical use.

The guidelines for scoring the verbal subtests of the WAIS-IV-ID were tested by psychologists from diverse professional backgrounds; clinical, educational, and industrial psychology. The inter-rater reliability was high; therefore, the scoring guidelines are reliable and ready for use.

The WAIS-IV-ID validation process included both internal and external checks. In addition, the diagnostic validation was also performed for three clinical groups; adults with Alzheimer's disease, schizophrenia, and intellectual disability. All FSIQ, indexes, and subtest scores from the WAIS-IV-ID could be used to distinguish between the non-clinical and clinical groups.

Finally, we succeeded in developing a reliable short form of the WAIS-IV-ID as an alternative to the full test, which might be considered a strength for both research and clinical applications.

## LIMITATIONS

Our data were collected from sample populations with varying compositions of people from the major islands of Indonesia, despite designing our research to recruit a sample representative of the Indonesian population. We included population data from the six largest islands of Indonesia.

Despite reordering the item sequences in the WAIS-IV-ID, we were not able to alter the content of the items themselves. Unfortunately, Indonesian people are not familiar with some of the concepts represented in these items; for example, the PC subtest has a specific missing item, ‘snow’, a type of weather unfamiliar to most Indonesian people. The only place that snow can be found in Indonesia is Puncak Carstenz mountain on Papua. The item with an image of the type of stove typically used in the US was perceived by Indonesian participants as a washing machine, as these types of stoves are not commonly used in Indonesia; participants therefore give an incorrect response because the images were not culturally suitable.

In studying test-retest reliability, we only recruited participants from Jakarta and Surabaya, and they had a narrow range of ages. These limitations may reduce the generalizability of the findings.

## FUTURE DIRECTIONS

The fundamental assumption underlying intelligence testing is that an individual’s true intellectual ability is measurable and can be captured by the score of an intelligence assessment; however, it is possible that these tests may not adequately represent intellectual ability. Despite the vigorous and meticulous steps in the adaptation and psychometric evaluation of these assessments, intelligence tests still do not fully represent the construct of intellectual ability.

We adapted the sampling method to reflect the Indonesian population as closely as possible; however, it was not feasible to apply all relevant criteria. In future adaptations of intelligence assessments, it might be beneficial to include data from the inhabitants of Papua.

One major challenge is the complexity of standardization. Indonesia is a multicultural country, and generalized sampling procedures may not be suitable for some individuals. Attending to this complexity and developing culture-specific psychological assessment practices are important, as currently the test content is not universally suitable.

Future research could also try to reassess the test-retest reliability of the WAIS-IV-ID. The time interval could be extended to assess the long-term stability of the test, either in non-clinical or clinical populations. For clinical populations, confirming the test score stability was an essential factor in boosting clinician confidence in the WAIS-IV-ID, especially considering the time required for the psychological assessment.

## CONCLUSIONS

Intelligence assessments can provide a meaningful basis to understand the strengths and weaknesses of the potential abilities of children, adolescences, and adults; therefore, we should ensure that the tools used in intellectual assessment are psychometrically sound. The WAIS-IV-ID studies reported here have demonstrated the psychometric acceptability of the test. The item sequence was adapted to better measure the participants' abilities, and the subtests and indexes are reliable and have high inter-rater agreement. Both the internal and external structures of the assessment are valid, as demonstrated in a comparison with educational achievement tests and other intelligence tests. Furthermore, the WAIS-IV-ID is considered a clinically valid instrument to test intelligence and is capable of distinguishing between non-clinical and clinical populations. We also were able to attain a reliable estimation of the FSIQ when using a short-form test comprising the subtests IN, MR, AR, and CO. Based on the studies presented in this thesis, I conclude that the WAIS-IV-ID is a reliable and valid instrument to assess the intelligence profiles of adults in Indonesia.

## REFERENCES

Ackerman, P. L., Beier, M. E., & Boyle, M. O. (2002). Individual differences in working memory within a nomological network of cognitive and perceptual speed abilities. *Journal of Experimental Psychology: General, 131*, 567–589.

Akaike, H. (1987). Factor analysis and AIC. *Psychometrika, 52*, 317–332.

Anastasi, A., & Urbina, S. (1997). *Psychological testing (7th ed.).* New Jersey: Prentice Hall.

Atkinson, L., Cyr, J. J., Doxey, N. C. S., & Vigna, C. M. (1989). Generalizability of WAIS-R factor structure within and between populations. *Journal of Clinical Psychology, 45*, 124–129.

Ayres, L. (1911). The Binet-Simon Measuring Scale for Intelligence: Some Criticisms and Suggestions. The *Psychological Clinic,* i, 187-196.

Badan Pusat Statistik Indonesia (BPS). (2009). Data demografik Indonesia, 2008 [Indonesian demographic data 2008]. Retrieved from http://www.bps.go.id.

Badan Pusat Statistik Indonesia (BPS). (2012). Perkembangan beberapa indikator utama sosial-ekonomi Indonesia [Trends of selected socio-economic indicators of Indonesia]. Retrieved from http://www.bps.go.id/booklet/Booklet_Agustus_2012.pdf

Badan Pusat Statistik Indonesia (BPS). (2016). Indikator Pendidikan, 1994-2015 [Education Indicator 1994 - 2015]. Retrieved from https://www.bps.go.id/linkTabelStatis/ view/id/1525

Badan Pusat Statistik Indonesia (BPS). (2017a). Statistik Indonesia dalam Infografis 2017 [Statistical Yearbook of Indonesia in Infographic 2017]. Retrieved from https://www.bps.go.id/publikasi.html

Badan Pusat Statistik Indonesia (BPS). (2017b). Statistik Politik 2017 [Political Statistics 2017]. Retrieved from https://www.bps.go.id/publikasi.html

Barrett, P. (2007). Structural equation modeling: Adjudging model fit. *Personality and Individual Differences, 42*, 815–824.

Benson, E. (2003). Intelligent intelligence testing. *Monitor on Psychology, 34*(2), 48. Retrieved from http://www.apa.org/monitor/feb03/intelligent.aspx

Benson, N., Hulac, D. M., & Kranzler, J. H. (2010). Independent examination of the Wechsler adult intelligence scale-fourth edition (WAIS-IV): What does the WAIS-IV measure? *Psychological Assessment, 22*(1), 121 - 130. doi: 10.1037/a0017767.

Bentler, P. M. (1995). EQS structural equations program manual. Encino, CA: Multivariate Software.

Blanchard, J. J., & Neale, J. M. (1994). The neuropsychological signature of schizophrenia: Generalized or differential deficit? *American Journal of Psychiatry, 151*, 40–48.

Boake, C. (2002). From the Binet-Simon to the Wechsler-Bellevue: Tracing the history of intelligence testing. *Journal of Clinical and Experimental Neuropsychology, 24*, 383–405. doi:10.1076/jcen.24.3.383.981.

Bowden, S. C., Saklofske, D. H., & Weiss, L. G. (2011a). Augmenting the core battery with supplementary subtests: Wechsler Adult Intelligence Scale-IV measurement invariance across the United States and Canada. *Assessment, 18*, 133–140. doi:10.1177/1073191110381717.

Bowden, S. C., Saklofske, D. H., & Weiss, L. G. (2011b). Invariance of the measurement model underlying the Wechsler Adult Intelligence Scale-IV in the United States and Canada. *Educational and Psychological Measurement, 71*, 186–199. doi:10.1177/0013164410387382.

Brown, C. (2003). *A short history of Indonesia: The unlikely nation?* Australia: Allen & Unwin.

Brown, T. (2010). Construct validity: A unitary concept for occupational therapy assessment and measurement. *Hong Kong Journal of Occupational Theraphy, 20*(1), 30 – 42.

Canivez, G. L., & Watkins, M. W. (2010). Investigation of the factor structure of the Wechsler Adult Intelligence Scale-Fourth edition (WAIS-IV): Exploratory and higher-order factor analyses. *Psychological Assessment, 22*, 827–836. doi:10.1037/a0020429

Canivez, G. L., Konold, T. R., Collins, J. M., & Wilson, G. (2009). Construct validity of the Wechsler Abbreviated Scale of Intelligence and Wide Range Intelligence Test: Convergent and structural validity. *School Psychology Quarterly, 24*, 252–265. doi:10.1037/a0018030

Carroll, J.B. (1993). *Human cognitive abilities: A survey of factor-analytic studies.* New York: Cambridge University Press.

Carvajal, H., Schrader, M. S., & Holmes, C. B. (1996). Retest Reliability of the Wechsler Adult Intelligence Scale-Revised for 18- to 19-Year-Olds. *Psychological Reports, 78*(1), 211 - 214.

Cattell, R.B. & Cattell, A.K.S. (1973). *Measuring intelligence with the culture-fair tests: Manual for Scales 2 and 3.* Champaigne, IL: Institute for Personality and Ability Testing, Inc.

Cattell, R. B., & Cattell, A. K. S. (1959). Culture fair intelligence test (Scale 3). Champaigne, IL: Institute for Personality and Ability Testing, Inc.

Cohen, J. (1952). Factors underlying Wechsler-Bellevue performance of three neuropsychiatric groups. Journal of Abnormal and Social Psychology, 47(2), pp. 359 – 365.

Cohen, J. (1988). Statistical power analysis for the behavioral sciences (2nd ed.). Hillsdale, NJ: Lawrence Earlbaum Associates.

Colom, R., & Abad, F. (2007). Culture Fair Intelligence Test. In Neil J. Salkind, & K. Rasmussen (Eds.), *Encyclopedia of Measurement and Statistics.* (pp. 205-206). Thousand Oaks, CA: Sage Publications, Inc. doi: http://dx.doi.org/10.4135/9781412952644.n116

Cook, D. A. & Beckman, T. J. (2006). Current concepts in validity and reliability for psychometric instruments: Theory and application. *The American Journal of Medicine, 119*, 166.e7 – 166.e16.

Crawford, J. R., Allum, S., & Kinion, J. E. (2008). An index-based short form of the WAIS-III with accompanying analysis of reliability and abnormality of differences. *British Journal of Clinical Psychology, 47*, 215–237.

Crocker, L., & Algina, J. (2008). Introduction to classical and modern test theory. Ohio, USA: Cengage Learning.

Cyr, J. J., & Brooker, B. H. (1984). Use of appropriate formulas for selecting WAIS-R short forms. Journal of Consulting and Clinical Psychology, 52(5), 903 - 905.

Daniel, M. H. (1997). Intelligence testing: Status and trends. *American Psychologist, 52*(10), 1038−1045.

Deary, I. (2001). Intelligence: *A Very Short Introduction.* New York: Oxford University Press.

Deary, I., Strand, S., Smith, P., & Fernandes, C. (2007). Intelligence and educational achievement. *Intelligence, 35*, 13-21.

Deeks, J. J., & Altman, D. G. (2004). Diagnostic tests 4: likelihood ratios. *BMJ,* 329, 168-169.

Denney, D. A., Ringe, W. K., & Lacritz, L. H. (2015). Dyadic short forms of the Wechsler Adult Intelligence Scale - IV. *Archives of Clinical Neuropsychology, 30*(5), 404-412.

Donnell, A. J., Pliskin, N., Holdnack, J., Axelrod, B., & Randolph, C. (2007). Rapidly-administered short forms of the Wechsler Adult Intelligence Scale-3rd edition. *Archives of Clinical Neuropsychology, 22*, 917-924.

Doppelt, J. E. (1956). Estimating the Full Scale score on the Wechsler Adult Intelligence Scale from scores on four subtests. *Journal of Consulting Psychology, 20*(1), 63-66.

Errington, J.J. (1986). Continuity and Change in Indonesian Language Development. *The Journal of Asian Studies, 45*(2), 329-353. Retrieved from http://www.jstor.org/ stable/2055846

Farahat, F.M., Rohlman, D.S., Storzbach, D., Ammerman, T., & Anger, W.K. (2003). Measures of Short-Term Test–Retest Reliability of Computerized Neurobehavioral Tests. *Neuro Toxicology, 24*, 513 - 521.

Fawcett, T. (2006). An introduction to ROC analysis. *Pattern Recognition Letters, 27*, 861-874.

Field, A. (2013). *Discovering Statistics using IBM SPSS Statistics.* London: Sage Publication.

Fitrikasari, A., Jaeri, S., & Bintoro, A.C. (2013). Comprehensive care of epilepsy: Involvement of psychological assessment and wada test. *Basic Research Journal of Medicine and Clinical Sciences, 2*(1), 1-4.

Freund, P.A., & Holling, H. (2011). Who wants to take an intelligence test? Personality and achievement motivation in the context of ability testing. Personality and individual differences, 50, 723-728.

Garcia, L. F., Ruiz, M. A., & Abad, F. J. (2003). Factor structure of the Spanish WAIS-III. Psicothema, 15, 155–160.

Georgas, J., Weiss, L. G., van de Vijver, F. J., & Saklofske, D. H. (Eds.). (2003). *Culture and children's intelligence: Cross-cultural analysis of the WISC-III.* Cambridge, MA: Academic Press.

Gignac, G. E. (2014). On the inappropriateness of using items to calculate total scale score reliability via coefficient alpha for multidimensional scales. *European Journal of Psychological Assessment, 30*(2), 130-139.

Gignac, G. E., & Watkins, M. W. (2013). Bifactor modeling and the estimation of model-based reliability in the WAIS-IV. *Multivariate Behavioral Research, 48*, 639-662.

Gignac, G. E., Bates, T. C., & Jang, K. (2007). Implications relevant to CFA model misfit, reliability, and the Five Factor Model as measured by the NEO-FFI. *Personality and Individual Differences, 43*, 1051–1062.

Girard, T. A., Axelrod, B. N., Patel, R., & Crawford, J. R. (2015). Wechsler Adult Intelligence Scale-IV dyads for estimating global intelligence. *Assessment, 22*(4), 441-448.

Golay, P., Reverte, I., & Lecerf, T. (2011). Independent CFA of the French WAIS-IV. *IAP AP101 Brief Report #8.* Retrieved from http://www.iqscorner.com/2011/04/ap101-brief-8-guest-brief-independent.html.

Goldstein, G., & Watson, J.R. (1989). Test-retest reliability of the Halstead-Reitan battery and the WAIS in a neuropsychiatric population. *Clinical Neuropsychologist, 3*(3), 265-272.

Grieve, K.W., & Van Eeden, R. (2010). A preliminary investigation of the suitability of the WAIS-III for Afrikaans-speaking South Africans. *South African Journal of Psychology, 40*(3), 262-271.

Griffith, R. M., & Yamahiro, R. S. (1958). Reliability-stability of subtest scatter on the wechsler-bellevue intelligence scales. *Journal of Clinical Psychology, 14*(3), 317 - 318.

Groth-Marnat, G. (2009). Wechsler Intelligence Scales. In G. Groth-Marnat, *Handbook of Psychological Assessment* (5th ed., pp. 119 - 181). New Jersey: John Wiley & Sons, Inc.

Hallgren, K. A. (2012). Computing inter-rater reliability for observational data: An overview and tutorial. *Tutorials Quantitative in Methods for Psychology, 8*(1), 23–34.

Heinrichs, R. W. (1993). Schizophrenia and the brain: Conditions for a neuropsychology of madness. *American Psychologist, 48*, 221–233.

Hiscock, M. (2007). The Flynn effect and its relevance to neuropsychology. *Journal of Clinical and Experimental Neuropsychology, 29*(5), 514-29.

Horn, J. L., & Cattell, R. B. (1966). Refinement and test of the theory of fluid and crystallized intelligence. *Journal of Educational Psychology, 57*, 253–270.

Horn, J. L., & Cattell, R. B. (1967). Age differences in fluid and crystallized intelligence. *Acta Psychologica, 26*, 107–129. doi:10.1016/0001-6918(67)90011-X.

Hu, L. T., & Bentler, P. M. (1999). Cut-off criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling, 6*(1), 1–55. doi:10.1080/10705519909540118.

Hunt, E. (2011). *Human Intelligence.* Cambridge: Cambridge University Press.

International Test Commission (2010). International Test Commission Guidelines for Translating and Adapting Tests. [http://www.intestcom.org]

Jensen, A. R. (1980). *Bias in mental testing.* New York: The Free Press.

Jensen, A. R. (1993). Why is reaction time correlated with psychometric 'g'? *Current Directions in Psychological Science, 2*(2), 53-56.

Jensen, A. R. (1998). *The g factor: The science of mental ability.* Westport, CT: Praeger.

Jeyakumar, S. L., Warriner, E. M., Raval, V. V., & Ahmad, S. A. (2004). Balancing the need for reliability and time efficiency: short forms of the Wechsler adult intelligence scale-III. *Educational and Psychological Measurement, 64*(1), 71-87.

Johnson, W. & Bouchard, T. J. (2005). The structure of human intelligence: It is verbal, perceptual, and image rotation (VPR), not fluid and crystallized. *Intelligence, 33*, 393-416.

Jöreskog, K. G, & Sörbom, D. (1996). *LISREL 8 user's reference guide.* Chicago, IL: Scientific Software International.

Jöreskog, K. G., & Sörbom, D. (2006). LISREL (Version 8.80) [Computer software]. Lincolnwood, IL: Scientific Software International.

Kanerva, K., & Kalakoski, V. (2016). The Predictive Utility of a Working Memory Span Task Depends on Processing Demand and the Cognitive Task. *Applied Cognitive Psychology, 30*, 681–690.

Kaplan, R. M., & Saccuzzo, D. P. (2018). *Psychological Testing: Principles, Applications, and Issues* (9th ed.). Boston: Cengage Learning.

Kaufman, A. S., Ishikuma, T., & Kaufman-Packer, J. L. (1991). Amazingly short forms of the WAIS-R. *Journal of Psychoeducational Assessment, 9*, 4-15.

Kaufman, A. S., & Lictenberger, E. O., (2000). *Essentials of WISC-III and WPPSI-R assessment.* New York: Wiley.

Kaufman, J. C., & Kaufman, A. S. (2001). Time for the changing of the guard: A farewell to short forms of intelligence tests. *Journal of Psychoeducational Assessment, 19*(3), 245 - 267.

Keith, T. Z. & Reynolds, M. R. (2010). Cattell-Horn-Carroll abilities and cognitive tests: what we learned from 20 years of research. *Psychology in the Schools, 47*, 635-650.

Kessels R. P. C., & Hendriks, M. P. H. (2016). Neuropsychological Assessment. In Howard S. Friedman (Editor in Chief), *Encyclopedia of Mental Health* (2nd ed.), vol 3, Waltham, MA: Academic Press, pp. 197-201.

King, L. A., & King, D. W. (1982). Wechsler short forms: A brief status report. *Psychology in the Schools, 19*(4), 433–438.

Kline, R. B. (2005). Principles and practice of structural equation modeling (2nd ed.). New York: Guilford Press.

Kumar, R., & Indrayan, A. (2011). Receiver Operating Characteristic (ROC) Curve. *Indian Pediatrics, 48*, 277-287.

Kyllonen, P. C., & Christal, R. E. (1990). Reasoning ability is (little more than) working-memory capacity?! *Intelligence, 14*, 389–433.

Lakens, D. (2013). Calculating and reporting effect sizes to facilitate cumulative science: A practical primer for t-tests and ANOVAs. *Frontiers in Psychology, 4*, 1-12.

Lange, R. T., & Iverson, G. L. (2008). Concurrent validity of Wechsler Adult Intelligence Scales-Third Edition index scores short forms in the Canadian standardization sample. *Educational and Psychological Measurement, 68*(1), 139-153.

Lange, R.T. (2007). WAIS-III index score profiles in the Canadian standardization sample. Journal of Clinical and Experimental Neuropsychology, 29(1), 47-58.

Latuconsina, H., & Rafidi, D. (1996). *History for Senior High School volume 2.* Jakarta: Erlangga.

Lecerf, T., Golay, P., & Reverte, I. (2012). CHC composite scores for the WAIS-IV: French Norms. *Pratiques Psychologiques, 18*(4), 401-412.

Lembaga Pengembangan Sarana Pengukuran dan Pendidikan Psikologi Fakultas Psikologi Universitas Indonesia (LSP3 FPUI). (n.d). *Manual Wechsler-Bellevue Intelligence Scale for Adolescents and Adults Form I.* Depok: LSP3 FPUI.

Lembaga Pengembangan Sarana Pengukuran dan Pendidikan Psikologi Fakultas Psikologi Universitas Indonesia (LSP3 FPUI). (n.d). *Tabel IQ WBIS.* Depok: LSP3 FPUI.

Lembaga Pengembangan Sarana Pengukuran dan Pendidikan Psikologi Fakultas Psikologi Universitas Indonesia (LSP3 FPUI). (2009). Buku Petunjuk Penggunaan: *CFIT Culture Fair Intelligence Test Skala 3 Bentuk A dan Bentuk B.* Depok: LSP3 FPUI.

Lembaga Pengembangan Sarana Pengukuran dan Pendidikan Psikologi Fakultas Psikologi Universitas Indonesia. (n.d.). *Manual Wechsler-Bellevue Intelligence Scale for Adolescents and Adult Form I.* Depok: LSP3 FPUI.

Lembaga Pengembangan Sarana Pengukuran dan Pendidikan Psikologi (LPSP3). (2015). *Daftar Harga Alat Tes Psikologi.* Depok: LPSP3.

Lenhard, W., & Lenhard, A. (2016). Calculation of Effect Sizes. Available at http://www.psychometrica.de/effect_size.html. Bibergau (Germany): Psychometrica.

Levy, P. (1968). Short-forms tests: A methodological review. *Psychological Bulletin, 69*(6), 410-416.

Lewis, M. P., Simons, G. F., & Fennig, C. D. (eds). (2013). *Ethnologue: Languages of the World* (17th ed.). Dallas, Texas: SIL International. Online version: http://www.ethnologue.com

Lynn, R., & Dai, X. Y. (1993). Sex differences on the Chinese standardization sample of the WAIS-R. *The Journal of Genetic Psychology: Research and Theory on Human Development, 154*(4), 459-463. doi: 10.1080/00221325.1993.9914744

Mackintosh, N. J. (2011). History of theories and measurement of intelligence. In R. J. Sternberg & S. B. Kaufman (Eds.), *Cambridge Handbook of Intelligence* (pp.3-19). New York: Cambridge University Press.

Malda, M., Van de Vijver, F. J. R., Srinivasan, K., Transler, C., Sukumar, P., & Rao, K. (2008). Adapting a cognitive test for a different culture: An illustration of qualitative procedures. *Psychology Science Quarterly, 50* (4), 451 – 468. Retrieved from http://www.sjri.res.in/ mhpublicationdetail/1161

Matarazzo J. D., & Herman, D. O. (1984). Relationship of education and IQ in the WAIS-R standardization sample. *Journal of Consulting and Clinical Psychology, 52*(4), 631–634. doi: 10.1037/0022-006X.52.4.631

Maxwell, E. (1957). Validities of abbreviated WAIS scales. *Journal of Consulting Psychology, 21*(2), 121 - 126.

McDonald, R. P. (1985). *Factor analysis and related methods.* Hillsdale, NJ: Erlbaum.

McGraw, K. O., & Wong, S. P. (1996). Forming inferences about some intraclass correlation coefficients. *Psychological Methods, 1* (1), 30-46.

McGrew, K. S. (1997). Analysis of the major intelligence batteries according to a proposed comprehensive Gf-Gc framework. In D. P. Flanagan, J. L. Genshaft, & P. L. Harrison (Eds.), *Contemporary intellectual assessment: Theories, tests, and issues* (pp. 151-179). New York: Guilford.

McGrew, K. S. (2004). The Cattell-Horn-Carroll (CHC) Theory of Cognitive Abilities: Past, Present and Future. Retrieved June 22, 2017 from http://www.iapsych.com/CHCPP/ 1.Introduction.html#Topic1

McGrew, K. S. (2009). CHC theory and the human cognitive abilities project: Standing on the shoulders of the giants of psychometric intelligence research. *Intelligence, 37*, 1-10. doi: 10.1016/j.intell.2008.08.004.

McNemar, Q. (1950). On abbreviated Wechsler-Bellevue scales. *Journal of Consulting Psychology, 14*, 79 - 81.

Melendez, F. (1994). The Spanish version of the WAIS: Some ethical considerations. *Clinical Neuropsychologist, 8*(4), 388-393. doi: 10.1080/13854049408402041

Meyers, J. E., Zellinger, M. M., Kockler, T., Wagner, M., & Miller, R. M. (2013). A validated seven-subtest short form for the WAIS-IV. *Applied Neuropsychology: Adult, 20*(4), 1-8.

Miller, M. B. (1995). Coefficient alpha: A basic introduction from the perspectives of classical test theory and structural equation modeling. *Structural Equation Modeling, 2*, 255-273.

Murayama, N., Iseki, E., Tagaya, H., Ota, K., Kanasuki, K., Fujishiro, H., Arai, H., & Sato, K. (2013). Intelligence or years of education: which is better correlated with memory function in normal elderly Japanese subjects? *Psychogeriatrics, 13*, 9-16. doi: 10.1111/j.1479-8301.2012.00408.x

Naglieri, J. A., & Bornstein, B. T. (2003). Intelligence and Achievement: Just how Correlated are they? *Journal of Psychoeducational Assessment, 21*, 244-260. doi: 10.1177/ 073428290302100302

Nenty, H. J. & Dinero, T. E. (1981). A cross-cultural analysis of the fairness of the Cattell Culture Fair Intelligence Test using Rasch model. *Applied Psychological Measurement, 5* (3), 355-368. doi:10.1177/014662168100500309

Niileksela, C. R., Reynolds, M. R., & Kaufman, A. S. (2013). An alternative Cattell-Horn-Carroll (CHC) factor structure of the WAIS-IV: Age invariance of an alternative model for ages 70-90. *Psychological Assessment, 25*, 391–404. doi:10.1037/a0031175.

Nisbett, R. E., Aronson, J., Blair, C., Dickens, W., Flynn, J., Halpern, D. F., & Turkheimer, E. (2012). Intelligence: New findings and theoretical developments. *American Psychologist, 67*(2), 130-159.

Nunnally, J. C., & Bernstein, I. H. (1994). *Psychometric Theory* (3rd ed.). New York: McGraw-Hill.

Padilla, M. A., & Divers, J. (2016). A comparison of composite reliability estimators: Coefficient omega confidence intervals in the current literature. *Educational and Psychological Measurement, 76*(3), 436-453.

Parker, D. R., & Benedict, K. B. (2002). Assessment and intervention: Promoting successful transitions for college students with ADHD. *Assessment for Effective Intervention, 27*(3), 3-24.

PERSONA. (2015). *Daftar Harga Alat Tes Psikologi.* Bandung: Biro Psikologi Persona.

Peters, G.-J. Y. (2014). The alpha and the omega of scale reliability and validity: Why and how to Abandon Cronbach's Alpha. *The European Health Psychologist, 16*(2), 56-69.

Pintea, S., & Moldovan, R. (2009). The Receiver-Operating Characteristic (ROC) analysis: Fundamentals and applications in clinical psychology. *Journal of Cognitive and Behavioral Psychoterapies, 9*(1), 49-66.

Pluck, G., Ruales-Chieruzzi, C. B., Paucar-Guerra, E. J., Andrade-Guimaraes, M. V., & Trueba, A. F. (2016). Separate contributions of general intelligence and right prefrontal neurocognitive functions to academic achievement at university level. *Trends in Neuroscience and Education, 5*, 178-185.

Polhaupessy, L.F. (n.d). *Tes Wechsler-Bellevue.* Bandung: internal used publication.

Rabin, A. I. (1943). A short form of the Wechsler-Bellevue test. *Journal of Applied Psychology, 27*, 320-324.

Raven, J. (2000). The Raven's Progressive Matrices: Change and Stability over Culture and Time. *Cognitive Psychology, 41*, 1-48.

Raven, J. (2008). General introduction and overview: The Raven Progressive Matrices Tests: Their theoretical basis and measurement model. In J. Raven & J. Raven (Eds.) *Uses and Abuses* *of Intelligence: Studies Advancing Spearman and Raven's Quest for Non-Arbitrary Metrics.* Unionville, New York: Royal Fireworks Press; Edinburgh, Scotland: Competency Motivation Project; Budapest, Hungary: EDGE 2000; Cluj Napoca, Romania: Romanian Psychological Testing Services SRL. (Chapter 1, pp. 17-68).

Reckase, M. D. (1989). Adaptive testing: The evolution of a good idea. *Educational Measurement: Issues & Practice,* 8, 11-15. doi: 10.1111/j.1745-3992.1989.tb00326.x

Reise, S. P., Bonifay, W. E., & Haviland, M. G. (2012). Scoring and modeling psychological measures in the presence of multidimensionality. *Journal of Personality Assessment, 95*, 129-140.

Rindermann, H., Becker, D., & Coyle, T. R. (2017). Survey of expert opinion on intelligence: The FLynn effect and the future of intelligence. *Personality and Individual Differences, 106*, 242-247.

Ringe, W. K., Saine, K. C., Lacritz, L. H., Hynan, L. S., & Cullum, C. M. (2002). Dyadic short forms of the Wechsler Adult Intelligence Scale-III. *Assessment, 9*(3), 254-260.

Rohde, T. E., & Thompson, L. A. (2007). Predicting academic achievement with cognitive ability. *Intelligence, 35*, 83-92.

Rohmah, U. (2011). Tes inteligensi dan pemanfaatannya dalam dunia pendidikan (The intelligence test and its use in education). *Cendekia, 9*(1), 125 - 139. Retrieved from http://jurnal.stainponorogo.ac.id/index.php/cendekia/article/viewFile/869/681

Roivainen, E. (2013). Are Cross-National Differences in IQ Profiles Stable? A Comparison of Finnish and U.S. WAIS Norms. *International Journal of Testing, 13*(2), 140-151. doi: 10.1080/15305058.2012.657315

Rushton, J. P., Brainerd, C. J., & Pressley, M. (1983). Behavioral Development and Construct Validity: The Principle of Aggregation. *Psychological Bulletin, 94*(1), 18-38.

Ryan, J. J., & Ward, L. C. (1999). Validity, reliability, and standard errors of measurement for two seven subtest short forms of the Wechsler Adult Intelligence Scale-III. *Psychological Assessment, 11*(2), 207-211.

Ryan, J. J., Kreiner, D. S., Gontkovsky, S. T., & Umfleet, L. G. (2015). Classification accuracy of sequentially administered WAIS-IV short forms. Applied Neuropsychology: Adult, 22, 409-414.

Ryan, J. J., & Schnakenberg-Ott, S. D. (2003). Scoring Reliability on the Wechsler Adult Intelligence Scale–Third Edition (WAIS-III). *Assessment, 10*(2), 151-159.

Schneider, W & McGrew, Kevin. (2012). The Cattell-Horn-Carroll model of intelligence. In Dawn P. Flanagan & Patti L. Harrison (Eds.), *Contemporary intellectual assessment: Theories, tests, and issues* (pp. 99-144). New York: Guilford.

Seksi Psikodiagnostik. (1992). *Buku Pegangan Wechsler Adult Intelligence Scale.* Yogyakarta: Fakultas Psikologi Universitas Gadjah Mada.

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

Silverstein, A. B. (1982). Two- and four-subtest short forms of the Wechsler Adult Intelligence Scale-Revised. *Journal of Consulting and Clinical Psychology, 50*(3), 415-418.

Silverstein, A. B. (1990). Short forms of individual intelligence tests. *Psychological Assessment, 2*(1), 3-11.

Silverstein, A. B. (1991). Reliability of score differences on Wechsler's intelligence scales. *Journal of Clinical Psychology, 47*(2), 264-266.

Šimundić, A. M. (2008). Measures of diagnostic accuracy: basic definitions. *Med Biol Sci, 22*(4), 61-65.

Smith, G. T., McCarthy, D. M., & Anderson, K. G. (2000). On the sins of short-form development. *Psychological Assessment, 12*(1), 102-111.

Suryadinata, L., Arifin, E. N., & Ananta, A. (2003). *Indonesia's Population: Ethnicity and Religion in a Changing Political Landscape.* Singapore: Institute of Southeast Asian Studies.

Suwartono, C., Amiseso, C. P., & Handoyo, R. T. (2017). Uji Reliabilitas dan Validitas Eksternal The Raven's Standard Progressive Matrices. *Jurnal Psikologi Indonesia Humanitas, 14* (1), 1 – 15.

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. doi:10.4236/psych.2014.514171.

Suwartono, C., Hendriks, M. P. H., Hidajat, L. L., Halim, M. S., & Kessels, R. P. C. (2017). Structural Validity of the Indonesian Wechsler Adult Intelligence Scale – Fourth Edition (WAIS-IV-ID). Manuscript submitted for publication.

Suwartono, C. & Santoso, J., B. (2016). Attitudes toward Psychological Test Use in Indonesia. Anima *Indonesian Psychological Journal, 31*(4), 160-169.

Swets, J. A. (1988). Measuring the Accuracy of Diagnostic Systems. *Science, 240*(4857), 1285-1293.

Tellegen, A., & Briggs, P. F. (1967). Old wine in new skins: Grouping Wechsler subtests into new scales. *Journal of Consulting Psychology, 31*(5), 499 - 506.

Thomas, S. L., & Scroggins, W. A. (2006). Psychological testing in personnel selection: Contemporary issues in cognitive ability and personality testing. *Journal of Business Inquiry, 5*, 28–38.

Thompson, A. P. (1987). Methodological issues in the clinical evaluation of two- and four-subtest short forms of the WAIS-R. *Journal of Clinical Psychology, 43*(1), 142-144.

Tillman, C. M., Bohlin, G., Sorensen, L., & Lundervold, A. J. (2009). Intelligence and specific cognitive abilities in children. *Journal of Individual Differences, 30*(4), 209–219.

Van der Heijden, P., Van den Bos, P., Mol, B. A. W., & Kessels, R. P. C. (2013). Structural validity of the Dutch Language version of the WAIS-III in a psychiatric sample. *Applied Neuropsychology: Adult, 20*, 41-46. doi: 10.1080/09084282.2012.670166.

Van Scoyoc, S. (2017). The use and misuse of psychometrics in clinical settings. In Barry Cripps (Ed.), *Psychometric Testing* (pp.187-196). Oxford: John Wiley & Sons, Ltd.

Verhaeghen, P., & Salthouse, T. A. (1997). Meta-analyses of age-cognition relations in adulthood: Estimates of linear and nonlinear age effects and structural models. *Psychological Bulletin, 122*(3), 231-249.

Vernon, P. E. (1965). Ability factors and environmental influences. *American Psychologist, 20*, 723-733.

Ward, L. C. (1990). Prediction of verbal, performance, and Full Scale IQs from seven subtests of the WAIS-R. *Journal of Clinical Psychology, 46*, 436-440.

Ward, L. C., Bergman, M. A., & Hebert, K. R. (2012). WAIS-IV subtest covariance structure: Conceptual and statistical considerations. *Psychological Assessment, 24*, 328–340. doi:10.1037/a0025614.

Ward, L. C., Ryan, J. J., & Axelrod, B. N. (2000). Confirmatory factor analysis of the WAIS–III standardization data. *Psychological Assessment, 12*, 341–345. doi:10.1037/1040-3590.12.3.341.

Watkins, M.W. (2017). The reliability of multidimensional neuropsychological measures: from alpha to omega. *The Clinical Neuropsychologist, 31*, 1113-1126. doi: 10.1080/13854046.2017.1317364

Watkins, M.W., Lei, P.-W., & Canivez, G.L. (2007). Psychometric intelligence and achievement: A cross-lagged panel analysis. *Intelligence, 35*, 59-68.

Wechsler, D. (1939). *Wechsler-Bellevue Intelligence Scale.* New York: The Psychological Corporation.

Wechsler, D. (1944). *The Measurement of Adult Intelligence.* (3rd ed). Baltimore: Williams & Wilkins.

Wechsler, D. (1981). The psychometric tradition: Developing the Wechsler adult intelligence scale. *Contemporary Educational Psychology, 6*(2), 82 – 85.

Wechsler, D. (2008a). *Wechsler Adult Intelligence Scale–Fourth edition: Administration and scoring manual.* San Antonio, TX: Pearson.

Wechsler, D. (2008b). *Wechsler Adult Intelligence Scale–Fourth edition: Technical and interpretive manual.* San Antonio, TX: Pearson.

Wechsler, D. (2012). WAIS-IV-NL: *Wechsler Adult Intelligence Scale–Fourth Edition* (Nederlandstalige bewerking: Technische handleiding en Afname en scoringshandleiding). Amsterdam: Pearson Assessment and Information B.V.

Weiss, L G., Saklofske, D. H., Coalson, D., & Raiford, S. E. (Eds.) (2010). *WAIS-IV clinical use and interpretation.* San Diego, CA: Academic Press.

Weiss, L. G., Keith, T. Z., Zhu, J., & Chen, H. (2013). WAIS-IV and clinical validation of the four- and five-factor interpretative approaches. *Journal of Psychoeducational Assessment, 31*(2), 94 – 113. doi:10.1177/0734282913478030.

Weiss, L. G., Saklofske, D. H., Coalson, D., & Raiford, S. E. (Eds.). (2010). WAIS-IV: Clinical use and interpretation. London, UK: Academic Press.

Westin, L. K. (2001). Receiver operating characteristic (ROC) analysis: Evaluating discriminance effects among descision support systems. Sweden: Umeå universitet-Department of Computing Science. Retrieved June 16, 2016, from http://nutkin.cs.umu.se/research/reports/2001/018/part1.pdf

Williams, W. M. (1998). Are we raising smarter children today? School- and home- related influences on IQ. In U. Nesser (Ed.), *The rising curve: Long-term gains in IQ and related measures* (pp. 125 - 154). Washington, DC: American Psychological Association.

Youden, W. J. (1950). *Index for rating Diagnostic Tests. Cancer, 3*(1), 32-35.

Zinbarg, R. E., Revelle, W., Yovel, I., & Li, W. (2005). Cronbach's α, Revelle's β, and McDonald's ωh: Their relations with each other and two alternative conceptualizations of reliability. *Psychometrika, 70*, 123–133.

Zou, K. H., O'Malley, J., & Mauri, L. (2007). Receiver-operating characteristic analysis for evaluating diagnostic tests and predictive models. *Circulation, 115*, 654-657.

## SUMMARY IN ENGLISH

Psychological assessment is a series of many activities undertaken by psychologists and psychological services for various purposes. Intelligence testing is always conducted during a comprehensive psychological assessment. Some of the most well-known and widely accepted tests are Wechsler's intelligence scales for individual intelligence testing. The Wechsler-Bellevue Intelligence Scale (WBIS; 1939) and the Wechsler Adult Intelligence Scale – Revised (WAIS-R; 1981) tests have previously been adapted for use in Indonesia; however, the latest revised version of Wechsler's intelligence scale, the Wechsler Adult Intelligence Scale – Fourth Edition (WAIS-IV; 2008), has not been adapted for the Indonesian population, despite its popularity worldwide.

The Wechsler intelligence scales available in Indonesia were adapted decades ago, and the minimum documentation, adaptation process, and the psychometric properties of these tests are not available. The present study aimed to provide a valid, reliable, and accurate measurement of individual cognitive function for Indonesian people. This thesis therefore covers the adaptation of the WAIS-IV, as well as reporting its validation in a series of academic papers, standing as a breakthrough for intelligence measurement through empirical studies in Indonesia. In addition, the development of an accurate and reliable WAIS-IV for use in psychological testing will benefit both healthy adults and those with special needs or clinical problems.

First, we selected translators with a psychological background, who translated the text into Indonesian and back to English. Several experts, including psychologists, Indonesian language experts, and psychometricians, evaluated the translation of the test. We then rearranged the sequence of items based on their index difficulty. The scoring guidelines for the verbal subtests were also tested for inter-rater agreement. The next step was to test the structural validity of the Indonesian version of the WAIS-IV (WAIS-IV-ID). The WAIS-IV-ID has a four-factor structure identical to the US version of the WAIS-IV.

The results of the WAIS-IV-ID studies were psychometrically acceptable. The reliabilities of the indexes and subtests are good, in terms of their internal consistency and inter-rater agreement for the scoring of the verbal subtests. Moreover, a test-retest approach indicated that the results were stable over time.

For some applications, the test user only requires an estimation of the Full Scale IQ (FSIQ), leading us to explore the possibility of developing a shorter test battery, consisting of four subtests, which can be used to reduce the administration time while maintaining the validity and reliability of the FSIQ estimation. The short form is useful for providing an indicative overview of intelligence and facilitating the rapid classification of research participants, as well as indicating whether further (neuro)psychological assessment is needed.

It was also important to investigate the psychometric properties of the WAIS-IV-ID in clinical groups. The WAIS-IV-ID was tested in three clinical groups: adults with

Alzheimer's disease, schizophrenia, and intellectual disability. The WAIS-IV-ID showed good internal consistency and clinical utility, reliably and validly distinguishing between the non-clinical and clinical groups.

The WAIS-IV-ID was externally validated through a comparison with an educational achievement score, represented by a university grade point average (GPA), and the results of other intelligence tests, such as Raven's Standard Progressive Matrices (SPM) and Cattell's Culture Fair Intelligence Test (CFIT). The WAIS-IV-ID was proved to be externally valid.

The WAIS-IV-ID was developed for use in modern Indonesia. Here, the need for the investigation of the psychometric properties of the WAIS-IV-ID was addressed for both non-clinical and clinical groups. In the future, the WAIS-IV-ID could contribute to intelligence assessments performed in clinical, education, social, and organizational contexts. Such assessments are increasingly important in Indonesia, and the WAIS-IV-ID will likely serve as the leading test used by Indonesian researchers and practitioners in (neuro)psychology.

## SUMMARY IN DUTCH

Een psychologisch onderzoek is één van vele activiteiten, die door psychologen en psychologische dienstverleners voor verschillende doeleinden worden uitgevoerd. In een uitgebreid psychologisch onderzoek, wordt vrijwel altijd een intelligentietest opgenomen. De meeste van de beschikbare tests in Indonesië zijn echter reeds decennia geleden in gebruik genomen, met een geringe documentatie over het proces van aanpassing en de psychometrische eigenschapen van de testen.

Een van de bekendste en wereldwijd meest gebruikte intelligentietests is de intelligentietest van Wechsler. In Indonesië wordt op dit moment nog steeds gebruik gemaakt van vertalingen van één van de eerste versies: de Wechsler-Bellevue Intelligence Scale (WBIS) uit 1939, de Wechsler Adult Intelligence Scale (WAIS) uit 1955 en de herziene versie (WAIS-R; 1981). Internationaal zijn deze sterk verouderde versies en niet aangepast aan de Indonesische taal en demografische kenmerken. Het gebrek aan goed gestandaardiseerde psychologische tests, heeft meetfouten en foute beslissingen van clinici en beleidsmakers over de mogelijkheden van individuen, tot gevolg. In Indonesië is dan ook reeds een groot aantal jaren behoefte aan een betrouwbaar en valide meetinstrument van de intelligentie.

De meest recente en herziene versie van de Wechsler intelligentieschaal in de Verenigde Staten, is de Wechsler Adult Intelligence Scale – Fourth Edition (WAIS-IV). In dit proefschrift wordt de ontwikkeling en bewerking van de WAIS-IV voor Indonesië beschreven: de WAIS-IV-ID. De adaptatie van de WAIS-IV-ID aan demografische kenmerken van de Indonesische bevolking wordt beschreven en een aantal empirische studies naar de psychometrische kenmerken van de test. Het onderzoek is gericht op het verbeteren van een meer valide, betrouwbare en nauwkeurige meting van de individuele intelligentie van de Indonesische bevolking. Tevens draagt een valide en betrouwbare psychologische test als WAIS-IV-ID, bij aan de diagnostiek en hulpverlening aan volwassenen met speciale behoeften en klinische problemen.

Voor de bewerking is de handleiding van de Amerikaanse WAIS-IV eerst vertaald door vertalers met psychologische kennis en de beheersing van het Indonesisch en Engels. Daarna is de volgorde van de items bepaald op basis van hun moeilijkheid. Vervolgens zijn de scoringsregels voor de verbale subtests onderzocht met behulp van de inter-beoo rdeelbaarheidsbetrouwbaarheid.

De resultaten van het onderzoek naar de psychometrische kenmerken van de WAIS-IV-ID stellen tevreden. De betrouwbaarheid van de indexen en subtesten is goed, ook wat betreft de onderlinge overeenstemming met de score van de mondelinge subtesten. Factoranalyses tonen de meest evidentie voor een vier-factorenstructuur, overeenkomstig de factorstructuur van de WAIS-IV-US. De betrouwbaarheid van het model is vergeleken met de Index Prestasi Kumulatip (IPK) en met zogenoemde neventests als het Grade Point Average (GPA), de Standard Progressive Matrices (SPM) en Cattel's Culture

Fair Intelligence Test (CFIT). De WAIS-IV-ID bleek verder in staat om een betrouwbaar en valide onderscheid te maken tussen niet-klinische en klinische groepen, zoals patiënten met de ziekte van Alzheimer, schizofrenie, of intellectuele beperkingen. In dit onderzoek is aangetoond dat de WAIS-IV-ID over voldoende externe validiteit beschikt.

Ondanks de populariteit van de Wechsler intelligentieschalen in de klinische praktijk en het wetenschappelijk onderzoek, wordt de afnameduur van het totaal aantal subtests vaak als tijdrovend ervaren. Wij onderzochten mogelijke verkorte versies van de WAIS-IV-ID. De meest accurate, verkorte versie betreft een combinatie van de subtests Matrix Redeneren, Informatie, Rekenen en Symbool Substitutie Coderen.

De studies in dit proefschrift over de WAIS-IV-ID laten zien dat deze intelligentietest betrouwbaar en valide is voor gebruik in de praktijk. De test kan derhalve aanbevolen worden voor beoordelingen van de intelligentie in klinische settings, onderwijsgebied en binnen organisaties.

## SUMMARY IN INDONESIAN

Penilaian psikologis merupakan serangkaian kegiatan yang dilakukan oleh psikolog dan biro layanan psikologis untuk berbagai keperluan. Tes kecerdasan selalu digunakan dalam sebuah penilaian psikologis yang komprehensif. Salah satu tes yang dikenal dan diterima secara luas sebagai tes kecerdasan individu adalah skala kecerdasan Wechsler. Wechsler-Bellevue Intelligence Scale (WBIS; 1939) dan Wechsler Adult Intelligence Scale – Revised (WAIS-R; 1981) telah diadaptasi untuk digunakan di Indonesia; namun, versi terkini dari skala kecerdasan Wechsler, Wechsler Adult Intelligence Scale – Fourth Edition (WAIS-IV; 2008), belum disesuaikan untuk penduduk Indonesia, meskipun popularitasnya sudah mendunia.

Skala inteligensi Wechsler yang tersedia di Indonesia telah diadaptasi beberapa dekade yang lalu dengan pencatatan yang minim bahkan hampir tidak ada, baik dari segi proses adaptasi maupun properti psikometri dari tes tersebut. Disertasi ini ditujukan untuk mendorong pengukuran yang sahih, terpercaya, dan akurat dari pengukuran fungsi kognitif individu di Indonesia. Lebih jauh daripada sekedar mengadaptasi instrumen psikologis yang lebih komprehensif, disertasi ini dirancang untuk menghasilkan artikel-artikel akademis dan menjadi mercusuar untuk pengukuran kecerdasan masyarakat Indonesia melalui studi-studi empiris. Penelitian ini bertujuan untuk meningkatkan pengukuran fungsi kognitif individu Indonesia yang lebih valid, andal dan akurat. Selain itut, tes psikologis yang akurat dan dapat diandalkan seperti WAIS-IV juga dapat berkontribusi baik terhadap kelompok orang dewasa pada umumnya maupun dengan kebutuhan khusus atau pun dengan masalah klinis.

Pada awalnya, kami memilih penerjemah dengan latar belakang psikologi (psikolog) untuk melakukan penterjemahan ke dalam Bahasa Indonesia dan kembali ke Bahasa Inggris. Beberapa ahli, termasuk psikolog, pakar bahasa Indonesia, dan psikometri mengevaluasi hasil kerja penerjemah tes. Kemudian, urutan item disusun ulang berdasarkan tingkat kesulitannya. Pedoman penilaian untuk subtes verbal juga diuji untuk kesepakatan antar penilai. Langkah selanjutnya adalah menguji validitas struktural dari Wechsler Adult Intelligence Scale – Fourth Edition versi Bahasa Indonesia (WAIS-IV-ID). WAIS-IV-ID memiliki empat faktor struktur model, struktur yang sama dengan WAIS-IV-US.

Hasil-hasil pengujian psikometrik dari WAIS-IV-ID dapat diterima secara psikometri. Keandalan dari semua subtes dan indeks tergolong bagus, dalam hal konsistensi internal dan kesepakatan antar-penilai untuk penyekoran subtes verbal. Kemudian, pengujian dengan pendekatan metode reliabilitas test-retest, hasil pengukurannya stabil dari waktu ke waktu.

Untuk beberapa aplikasi, pengguna tes hanya membutuhkan estimasi dari fungsi kecerdasan umum (FSIQ). Kebutuhan ini membawa kita untuk mengeksplorasi kemungkinan komposisi baterai tes yang lebih singkat yang terdiri dari empat subtests.

Hal ini tentunya dapat mengurangi waktu administrasi, tentunya dengan tetap menjaga validitas dan reliabilitas estimasi dari FSIQ. Bentuk singkat ini berguna untuk memberikan ikhtisar indikatif dari kecerdasan dan memfasilitasi klasifikasi secara cepat dalam suatu penelitian, serta menunjukkan apakah diperlukan penilaian (neuro)psikologis yang lebih lanjut.

Penyelidikan mengenai properti psikometrik dari WAIS-IV-ID juga penting dilakukan pada kelompok-kelompok klinis. Dalam hal ini, WAIS-IV-ID diuji pada tiga kelompok klinis: orang dewasa dengan Alzheimer, skizofrenia, dan disabilitas intelektual. WAIS-IV-ID menunjukkan konsistensi internal dan utilitas klinis yang baik serta dapat membedakan antara kelompok non-klinis dan klinis dengan handal dan valid.

WAIS-IV-ID juga telah divalidasi secara eksternal melalui perbandingan dengan skor prestasi pendidikan yang diwakili oleh Indeks Prestasi Kumulatif (IPK) di perguruan tinggi dan tes-tes kecerdasan lainnya, seperti Raven’s Standard Progressive Matrices (SPM) dan Cattell's Culture Fair Intelligence Test (CFIT). WAIS-IV-ID terbukti valid secara eksternal.

WAIS-IV-ID dikembangkan untuk digunakan di Indonesia modern. Di sini, kebutuhan untuk penyelidikan properti psikometrik dari WAIS-IV-ID ditujukan untuk kedua kelompok non-klinis dan klinis. Di masa depan, WAIS-IV-ID dapat berkontribusi pada penilaian inteligensi yang dilakukan dalam konteks klinis, pendidikan, sosial, dan organisasi. Penilaian semacam ini semakin penting di Indonesia, dan WAIS-IV-ID kemungkinan akan berfungsi sebagai tes terkemuka yang digunakan oleh peneliti dan praktisi Indonesia dalam bidang (neuro)psikologi.

**LIST OF ABBREVIATIONS**

| | | |
|---|---|---|
| SPM | = *The Raven's Standard Progressive Matrices* | |
| CFIT | = *The Cattell's Culture Fair Intelligence Test* | |
| GPA | = Grade Point Average | |
| IPK | = Indeks Prestasi Kumulatif (GPA in Indonesian) | |
| | | |
| WBIS | = *The Wechsler-Bellevue Intelligence Scale* | |
| I | = *Information* | = Informasi |
| C | = *Comprehension* | = Pemahaman |
| D | = *Digit Span* | = Rentang Digit |
| A | = *Arithmetic* | = Aritmatik |
| S | = *Similarities* | = Persamaan |
| V | = *Vocabulary* | = Kosa kata |
| PA | = *Picture Arrangement* | = Penyusunan Gambar |
| PC | = *Picture Completion* | = Penyelesaian Gambar |
| BD | = *Block Design* | = Desain Kubus |
| OA | = *Object Assembly* | = Perakitan Objek |
| DSym | = *Digit Symbol* | = Digit Simbol |
| POI | = *The Perceptual Organization Index* | = Indeks Skala Unjuk Kerja |
| WB | = *The Wechsler-Bellevue Intelligence Scale* | |
| | | |
| WAIS-IV | = *The Wechsler Adult Intelligence Scale-Fourth Edition* | |
| | = Skala Inteligensi Wechsler edisi Keempat | |
| WAIS-IV-ID | = *The Indonesian version of Wechsler Adult Intelligence Scale-Fourth Edition* | |
| | = Skala Inteligensi Wechsler edisi Keempat versi Bahasa Indonesia | |
| WAIS-IV-US | = *The American of version Wechsler Adult Intelligence Scale-Fourth Edition* | |
| | = Skala Inteligensi Wechsler edisi Keempat versi Amerika | |
| BD | = *Block Design* | = Desain Kubus |
| SI | = *Similarity* | = Persamaan |
| DS | = *Digit Span* | = Rentang Digit |
| MR | = *Matrix Reasoning* | = Penalaran Matriks |
| VC | = *Vocabulary* | = Kosa Kata |
| AR | = *Arithmetic* | = Aritmatik |
| SS | = *Symbol Search* | = Pencarian Simbol |
| VP | = *Visual Puzzle* | = Teka-teki Visual |
| IN | = *Information* | = Informasi |
| CD | = *Coding* | = Kode |
| LN | = *Letter Number Sequencing* | = Urutan Huruf-Angka |

| | | |
|---|---|---|
| FW | = *Figure Weights* | = Timbangan Bagun |
| CO | = *Comprehension* | = Pemahaman |
| CA | = *Cancellation* | = Pembatalan |
| PC | = *Picture Completion* | = Kelengkapan Gambar |
| VCI | = *Verbal Comprehension Index* | = Indeks Pemahaman Verbal |
| PRI | = *Perceptual Reasoning Index* | = Indeks Penalaran Persepsi |
| WMI | = *Working Memory Index* | = Indeks Working Memory |
| PSI | = *Processing Speed Index* | = Indeks Kecepatan Pemrosesan |
| FSIQ | = *general intellectual functioning, Full Scale Intelligence Quotient* | = fungsi kecerdasan umum |

## ACKNOWLEDGMENTS

I would like to thank my promotor, prof. dr. Roy P.C. Kessels and my daily supervisor, dr. Marc P.H. Hendriks from the Donders Institute of Brain, Cognition, and Behaviour from the Radboud University, Nijmegen. Thank you for your trust, advice (regarding either the articles or places to visit during my yearly three months stays in Nijmegen), freedom, and support (connecting me with others with the same interest). It was incredible to work with you and having you as mentors. I hope that we will continue to collaborate.

Thank you to my supervisor, Dr. Magdalena S. Halim, Psi at Unika Atma Jaya. Thank you for your initiative to bring this project to my Ph.D. research and to work together with Dr. Lidia L. Hidajat, Psi to get the grant from the Indonesian Ministry of Education and Culture, mostly used for the data collection all around Indonesia.

I kindly thank Sarah Jose, Rosa Delima, Lufiana Harnani Utami, Colleen Monaghan, Antonius Wiriadjaja, Danny Yatim, Novita Sutantoputri, and Agatha Hirst for their willingness to review this manuscript's language and style.

Next, I wish to thank my wonderful landlady and her husband, Jeanny, and Iman Slis. I have enjoyed our small talks, and every time I arrived, you always checked my day. You made me feel at home. Also, thank you for my friends at Atma Jaya: Weny Sembiring, Darma Juwono, Justinus, Venie, Yapina, Yenny, Adi, Aya, Ratri, Atiek, Stefani, Hana, Gigis, Danny, and others, and even at Radboud: Saskia, Paul, Agnes, Kevin, Lily, Casper, Evelyn, Johanna, Zita, Dirk, Fenny, Daniel, Valerio, Magda, and others for our conversations and traveling during my stay in Nijmegen. I would also thank Charlotte and Liselot for our yearly reunion in various places in the Netherlands since we met at Atma Jaya, Jakarta.

Thank you for my mentors in life from the Netherlands: Lanny Suhendra, Peter Merkx, Kitty Merkx, Jolanda Hendriks, Johannes Veldkamp, Joseph Geelen Crl, and Irenia Kristiani. You taught me a lot from your humor, stories, and books that you gave me. Thank you for showing me beautiful nature and historical buildings during my visits to the Netherlands. I enjoyed our conversations and hoped we will keep in touch.

I also would like to thank all those who have supported and who have shared the experiences with me during the last four years, especially during the data collection. In general, thank you for all mentors, participants, and research assistants (Achie, Adeline, Adi Ganjar, Alvernas Aniskripsa, Amelia Yulistiani Ratawinata, Andika Octavianto, Ariesti P., Ariny, Astrid Wulandari, B. Dian Suryandari, Bernadette Duisters, Bunga Andriani, Bunga Rassie A., Christina Maria A., Cindy Komara, Cucus W, Dede Fitriana A, Devita, Dima Wuenta Caesaria, Edo, Egie, Ellen Sumual, Eni, Erimson, Esther Agustin W., Eunike

Frederika, Eunike Nathania W., Eva Marina, Evelyn, Farida Dwi Rahma, Feblin Follen P.S, Febrika Kaswang, Firda Tamtelahitu, Firsta Andrina, Fitry, Gabriela Erika, Giovany Heny H., Glorry, Grace, Hadyan Dhiozandi, Hanifah, Hernilen, Intan, Ira, Irene Raflesia, Isabella Chan, Jessica Emily, Joanna Natalie, Justinus Budi Santoso, Karin Lucia Tanojo, Kartika Amelia, Kartika Indradjaja, Khristina Kencana, Kornelia Gracia, Kumala, Laura Ajawaila, Liliani S, Listra Senobaan, Lysa Laurensia, Made Syanesti Adishesa, Maria Helena Suprapto, Maria Natalia, Medwin Wisnu P, Mega Dewi, Michael Dirk, Morde, Nadya C, Nanda Rossalia, Nelly D. Pardede, Nesia Natalie, Nessia, Olivia Belinda, Prischa Nova, Priska, Putri, Putri Makmur, Rebeka Pinaima, Rendy, Reni Widowati, Retno Pratiwi, Riama Theresia. S, Rizky Haris Ilyas, Rosa Delima, Sad Yuli Prihartati, Sariri, Siska Natalia, Sjania Malik, Stefanus Cahya CW, Stefi, Stella, Stepani Kartika Bintang, Stephanie, Swastika Ayu, Tania, Triapermatasari, Tyas, Vania, Vilta, Widhi Adhiatma, Yenny Astria, Yenny Sahra, Yulianti Sanjaya). I remembered a friendly, collaborative, and supportive environment between us. As well, I remember many great moments, enjoyable routines, and the adventure time when we traveled to many places in Indonesia to collect the data. We became very creative, with limited resources (transportation and accommodation); sometimes they were enjoyable and sometimes they were also stressful ones. What a tremendous effort we did! Great job! Hopefully, we shall meet again soon. It was a joy and pleasure for me to know and work together with all of you. Some of you even now become my colleagues at the Faculty of Psychology, Unika Atma Jaya. Welcome to the academic life! Some of you have chosen another line of work. God's abundant blessings accompany you on your path of life.

Special thanks to Thomas Aquino who was ready to help with a spreadsheet program for data processing and the following institutions for their help and cooperation for the data collection: **Java Island** (**Bandung**: Jayamas; **Jakarta**: Universitas Katolik Indonesia Atma Jaya, Atjong Tex, Recruitment & Assessment Center Division - PT Astra International Tbk., Yayasan Spiritia, PWK Hana, Bright Consulting, Inti Data Consulting, PDA Consulting, Sekolah Menengah Atas Luar Biasa Tri Asih, Yayasan Tri Asih, Rumah Sakit Jiwa Soeharto Heerdjan, Woka Finance; Malang: Rumah Sakit Jiwa Dr. Radjiman Wediodiningrat; Surabaya: Universitas Pelita Harapan, Universitas Widya Mandala, Universitas Ciputra; Tangerang: Universitas Pelita Harapan, National Center for Gifted and Talented - Surya University, Universitas Matana, Siloam Hospital Lippo Village; **Yogyakarta**: Biro Psikologi Inner Power, Perpustakaan ICBC, RS Bethesda); **Sumatra Island** (**Medan**: Bunga Rassie (personal), **Palembang**: Universitas Bina Darma, Akademi Kebidanan Rizki Patya, Biara San Damiano, Biara Charitas Laverna, Biro Psikologi Magna Penta; **Riau**: Kapolresta Pekanbaru, RS Bhayangkara); **Bali Island** (**Badung**: SMAN 1 Kuta Utara; **Denpasar**: SMA Negeri 1 Denpasar); **Kalimantan** (**Borneo**) **Island** (**Banjarmasin**: Klinik Millenia; **Pontianak**: Andika Octavianto (personal), RS St. Antonius, Paroki St. Agustinus, SMA. St. Paulus, SMAK Immanuel, dan SMA St. Ignatius Singkawang); **Sulawesi** (**Celebes**) **Island**

(**Makassar**: Universitas Atma Jaya Makassar, Universitas Negeri Makassar; **Manado**: Max Tene (personal) and Skolastikat MSC Hati Kudus); and **Nusa Tenggara Islands** (**Kupang**: Alfani Alithea (personal), Universitas Widya Mandira, and Universitas Nusa Cendana).

Last but not least, I also thank my mother and my brother. They are my real supporters during my Ph.D. journey. Without the inspiration, motivation, and support you gave me, I might not be the person I am today.

In this work, I unfold myself then I get to know myself and others. These things are not only a cause of joy; often they are the cause of sadness. It is a blessing that I am allowed to experience them all, and everything equips me in my future tasks in my life.

Finally, in the journey, I have also experienced how diverse my country is. We have a different approach to solving a problem, and it made me curious for Indonesians. Even though I also found some challenges, I think this will be the exciting part of the Indonesian people. This dissertation is the first step to open more and more topics to be explored.

Warm regards,
Christiany Suwartono
October 8, 2018.

## CURRICULUM VITAE

Christiany Suwartono was born on October 29, 1982, in Purwokerto, Central Java, Indonesia. She obtained her bachelor's in psychology at the University of Indonesia in 2005. Then a master degree in Psychometric at the University of Indonesia in 2007. Since her master's degree, she has a growing interest in test construction and evaluations. After experienced worked in educational (University of Indonesia) and commercial (Ascendsys), she took a career as a full-time lecturer in psychological measurement department at the Atma Jaya Catholic University of Indonesia in 2007. To express and sharpen her skills, she actively collaborates with other academics and professionals in the area of testing and psychometric evaluations. In 2012, she began to work with Dr. Magdalena Halim to start the process of adapting The Wechsler Adult Intelligence Scale-Fourth Edition into the Indonesian language. The data collection for the pilot studies began in 2013 until 2015. In 2014, she started her Ph.D. under the supervision of Dr. Magdalena S. Halim, Psi, Dr. Marc P.H. Hendriks, and Prof. dr. Roy P. C. Kessels at The Donders Institute for Brain, Cognition and Behaviour at Radboud University, the Netherlands. Therefore, in the last five years, she had started traveling actively - mainly due to the data collection for her dissertation. She discovered much new knowledge and began many collaborations with various people and institutions.

## PUBLICATIONS

### ARTICLES

**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. DOI: http://dx.doi.org/10.4236/psych.2014.514171

**Suwartono, C.**, Hidajat, L. L., Halim, M. S., Hendriks, M. P. H., & Kessels, R. P. C. (2016). External Validity of the Indonesian Wechsler Adult Intelligence Scale – Fourth Edition (WAIS-IV-ID). *Anima Indonesian Psychological Journal, 32*(1), 16 - 28.

Oktaviano, A., Priadi, M. A. G., Halim, M. S., & **Suwartono, C.** (2016). The study of intelligence profiles between islands: A preliminary study towards norm development. *Anima Indonesian Psychological Journal, 32*(1), 55 - 64.

Van Dijk, M.J.A.M., Claassen, T., **Suwartono, C.**, van der Veld, W.M., van der Heijden, P.T., & Hendriks, M.P.H. (2017). Evaluating WAIS-IV structure through a different psychometric lens: structural causal model discovery as an alternative to confirmatory factor analysis. *The Clinical Neuropsychologist, 31*(6-7), 1141-1154. doi: 10.1080/13854046.2017.1352029.

### CONFERENCES

2018 The International Neuropsychological Society 2018 Mid-Year Meeting, Prague, Czech Republic.

2016 The 124th Annual Convention of the American Psychological Association, Denver, Colorado, USA.

2015 National Seminar at Faculty of Psychology, Universitas Sumatera Utara, Medan, North Sumatra, Indonesia.

2015 The 5th Conference of the European Societies of Neuropsychology and the 12th Nordic Meeting in Neuropsychology, Tampere Hall, Finland.

2015 The 13th European Conference on Psychological Assessment, Zurich, Switzerland.

2015 The 5th Pacific Rim Conference of the International Neuropsychology Society (INS) and the 38th Annual Brain Impairment Conference of the Australasian Society for the Study of Brain Impairment (ASSBI).

2014 The Scientific National Meeting / The 12th Congress of Indonesian Psychological Association (Temu Ilmiah Nasional/Kongres XII HIMPSI), Manado, Sulawesi Utara, Indonesia.

2014 The mid-year meeting of the International Neuropsychology Society (INS), Jerusalem, Israel

2014 The 9th Conference of the International Test Commission, San Sebastian, Spain.

## DONDERS GRADUATE SCHOOL FOR COGNITIVE NEUROSCIENCE

For a successful research Institute, it is vital to train the next generation of young scientists. To achieve this goal, the Donders Institute for Brain, Cognition and Behaviour established the Donders Graduate School for Cognitive Neuroscience (DGCN), which was officially recognised as a national graduate school in 2009. The Graduate School covers training at both Master's and PhD level and provides an excellent educational context fully aligned with the research programme of the Donders Institute.

The school successfully attracts highly talented national and international students in biology, physics, psycholinguistics, psychology, behavioral science, medicine and related disciplines. Selective admission and assessment centers guarantee the enrolment of the best and most motivated students.

The DGCN tracks the career of PhD graduates carefully. More than 50% of PhD alumni show a continuation in academia with postdoc positions at top institutes worldwide, e.g. Stanford University, University of Oxford, University of Cambridge, UCL London, MPI Leipzig, Hanyang University in South Korea, NTNU Norway, University of Illinois, North Western University, Northeastern University in Boston, ETH Zürich, University of Vienna etc.. Positions outside academia spread among the following sectors: specialists in a medical environment, mainly in genetics, geriatrics, psychiatry and neurology. Specialists in a psychological environment, e.g. as specialist in neuropsychology, psychological diagnostics or therapy. Positions in higher education as coordinators or lecturers. A smaller percentage enters business as research consultants, analysts or head of research and development. Fewer graduates stay in a research environment as lab coordinators, technical support or policy advisors. Upcoming possibilities are positions in the IT sector and management position in pharmaceutical industry. In general, the PhDs graduates almost invariably continue with high-quality positions that play an important role in our knowledge economy.

For more information on the DGCN as well as past and upcoming defenses please visit: http://www.ru.nl/donders/graduate-school/phd/